Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Επιθέσεις και Άμυνες σε Βαθιά Νευρωνικά Δίκτυα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΛΕΥΘΕΡΙΑ ΑΝΝΑ ΒΑΛΗ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Γεώργιος Αλεξανδρίδης
Εργαστηριακό και Διδακτικό Προσωπικό Ε.Μ.Π.

Αθήνα, Οκτώβριος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Επιθέσεις και Άμυνες σε Βαθιά Νευρωνικά Δίκτυα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΛΕΥΘΕΡΙΑ ΑΝΝΑ ΒΑΛΗ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Γεώργιος Αλεξανδρίδης
Εργαστηριακό και Διδακτικό Προσωπικό Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12η Οκτωβρίου 2021.

Ανδρέας Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2021

Ελευθερία Άννα Βαλή

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η πληθώρα δεδομένων που παρατηρείται στις μέρες μας καθώς και η εξέλιξη που έχει σημειωθεί τα τελευταία χρόνια στις κάρτες γραφικών δίνουν τη δυνατότητα ανάπτυξης ισχυρών δικτύων βαθιάς μηχανικής μάθησης. Τα βαθιά νευρωνικά δίκτυα, ως δίκτυα πολλαπλών κρυφών επιπέδων, αποτελούν υλοποιήσεις αλγορίθμων βαθιάς μηχανικής μάθησης που εφαρμόζονται σε διάφορους τομείς με κυριότερους την αναγνώριση εικόνας, κειμένου και ήχου. Συχνά, βαθιά νευρωνικά δίκτυα χρησιμοποιούνται και σε εφαρμογές υψηλού κινδύνου σε περίπτωση σφάλματος όπως στα συστήματα ανίχνευσης οικονομικών απατών ή στα αυτόνομα οχήματα.

Παρά την αποτελεσματικότητα και την ευρεία χρήση αυτών των δικτύων, έχει αποδειχθεί ότι τα βαθιά νευρωνικά δίκτυα είναι ευάλωτα σε κακόβουλες επιθέσεις. Τα κίνητρα καθώς και οι γνώσεις γύρω από το μοντέλο που μπορεί να διαθέτει ένας εισβολέας έχουν απασχολήσει σε βάθος την επιστημονική κοινότητα. Συγχρόνως έχουν αναπτυχθεί διάφορες στρατηγικές άμυνας χωρίς όμως να υπάρχει κάποια καθολική λύση για την αντιμετώπιση των επιθέσεων. Οι έρευνες εστιάζουν στην ενδελεχή μελέτη πιθανών επιθέσεων καθώς και στις τεχνικές αντιμετώπισης τους με απώτερο σκοπό τη δημιουργία εύρωστων πολυεπίπεδων νευρωνικών δικτύων που θα είναι σε θέση να αναγνωρίζουν και να διαχειρίζονται τυχόν επιθέσεις.

Στην παρούσα διπλωματική εργασία - εστιάζοντας στον τομέα της αναγνώρισης εικόνας - σε πρώτη φάση μελετήσαμε την one-pixel επίθεση σε συνδυασμό με την τεχνική ανίχνευσης αντιφατικών παραδειγμάτων PCA-whitening. Πιο συγκεκριμένα, εφαρμόζοντας την one-pixel επίθεση στα σύνολα δεδομένων MNIST και CIFAR10, δημιουργήσαμε αντιφατικά παραδείγματα με στόχο την παραπλάνηση του προεκπαιδευμένου VGG16 δικτύου. Στη συνέχεια, εφαρμόσαμε τον PCA-whitening ανιχνευτή στο σύνολο των καλοηθών και αντιφατικών παραδειγμάτων των δύο συνόλων δεδομένων, διεξάγοντας συμπεράσματα τόσο ως προς την ισχύ της one-pixel επίθεσης όσο και ως προς την αποτελεσματικότητα του δεδομένου ανιχνευτή.

Σε δεύτερη φάση, επεκτείνοντας την παραπάνω έρευνα, εφαρμόσαμε την FGSM επίθεση συνδυαστικά με την one-pixel δημιουργώντας νέα αντιφατικά παραδείγματα. Εισάγοντας στη συνέχεια τα κακόβουλα αυτά παραδείγματα στον PCA-whitening ανιχνευτή, μελετήσαμε τη δική του απόδοση στη συνδυαστική αυτή επίθεση. Τέλος, πειραματιστήκαμε με μια δεύτερη τεχνική ανίχνευσης, την squeezing color bits, και με τον συνδυασμό της με την PCA-whitening, εφαρμόζοντάς τες στις παραπάνω επιθέσεις.

Λέξεις κλειδιά

Βαθιά νευρωνικά δίκτυα, Επιθέσεις και Άμυνες, One-pixel επίθεση, PCA-whitening ανιχνευτής, Τεχνική μείωσης του βάθους του χρώματος των εικόνων

Abstract

Nowadays, powerful networks based on deep learning techniques are increasingly developed due to the abundance of data and the evolution of graphics cards. Deep neural networks, which contain multiple hidden layers, can be identified as deep learning algorithms with many applications in different scientific fields, including image, text and voice recognition. Deep neural networks are often applied to tasks critical to security, such as autonomous vehicles or financial fraud detection systems.

Despite the widespread use of these networks and their proven effectiveness, deep neural networks have been shown to be vulnerable to malicious attacks. Scientists are trying to figure out what motivates an attack and what kind of information is available to the attacker. Although different defense strategies have been developed in recent years, there has not yet been a universal solution to the problem of attacks. Therefore, the scientific research focuses on the thorough study of possible attacks as well as on their defense techniques, with the ultimate goal of creating robust multilayer neural networks that will be able to recognize and manage any attacks.

In this diploma dissertation we initially focused on the field of image recognition and studied the one-pixel attack in combination with the PCA-whitening detector of adversarial examples. In particular, by applying the one-pixel attack to the MNIST and CIFAR10 datasets, we created adversarial examples aimed at misleading the pre-trained VGG16 network. We then applied the PCA-whitening detector to all the benign and adversarial examples of the two datasets and we drew conclusions about the power of the one-pixel attack and the effectiveness of the given detector.

In addition to the above research, we applied the FGSM attack in combination with the one-pixel and created new adversarial examples. Moreover, we used these new examples as input to the PCA-whitening algorithm in order to study the performance of the detector in the double attack. Finally, we experimented with a second detection technique, the squeezing color bits, combining it with PCA-whitening and applying this double-defense to the aforementioned attacks.

Key words

Deep neural networks, Adversarial Attacks and Defenses, One-pixel attack, PCA-whitening technique, Squeezing color bits

Ευχαριστίες

Η παρούσα διπλωματική εργασία σηματοδοτεί την ολοκλήρωση των προπτυχιακών σπουδών μου. Συνεπώς, θα ήθελα να ευχαριστήσω θερμά όλους όσους με στήριξαν στα χρόνια των σπουδών μου και ιδιαίτερα κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα αυτής της διπλωματικής εργασίας, κ. Ανδρέα-Γεώργιο Σταφυλοπάτη, Καθηγητή Ε.Μ.Π., για την ευκαιρία που μου προσέφερε να εκπονήσω τη συγκεκριμένη εργασία καθώς και για την εμπιστοσύνη που μου έδειξε. Παράλληλα, θα ήθελα να ευχαριστήσω τον συνεπιβλέποντα, κ. Γεώργιο Αλεξανδρίδη, Εργαστηριακό και Διδακτικό Προσωπικό Ε.Μ.Π., για τη συνεχή υποστήριξη και καθοδήγησή του. Η άριστη συνεργασία μας, τα εποικοδομητικά του σχόλια και οι πολύτιμες συμβουλές του με βοήθησαν να οργανώσω την έρευνα μου και να ξεπεράσω σημαντικά εμπόδια που συνάντησα κατά την εκπόνηση της εργασίας. Επιπροσθέτως, θα ήθελα να ευχαριστήσω τους κ. κ. Στέφανο Κόλλια και Γεώργιο Στάμου, Καθηγητές Ε.Μ.Π., που με τίμησαν με τη συμμετοχή τους στην τριμελή επιτροπή εξέτασης.

Σε προσωπικό επίπεδο θα ήθελα να ευχαριστήσω από καρδιάς την οικογένειά μου και όλους μου τους φίλους για τη συνεχή ενθάρρυνση και υποστήριξή τους όλα αυτά τα χρόνια που συνέβαλαν ενεργά στην ολοκλήρωση αυτού του κύκλου σπουδών μου.

Ελευθερία Άννα Βαλή,

Αθήνα, 12η Οκτωβρίου 2021

Περιεχόμενα

Κατάλογος πινάκων

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

Στη σύγχρονη εποχή η πρόοδος που έχει σημειώσει ο άνθρωπος στον τομέα της βαθιάς μηχανικής μάθησης έχει οδηγήσει στην επίλυση πολύπλοκων προβλημάτων μέσω πλήρως αυτοματοποιημένων συστημάτων χωρίς καμία παρέμβαση ανθρώπινου παράγοντα. Χαρακτηριστικό παράδειγμα της εξέλιξης αυτής αποτελούν τα *βαθιά νευρωνικά δίκτυα* (ΒΝΔ), τα οποία έχουν σημειώσει πρωτοφανή επιτυχία σε πολλά προβλήματα μηχανικής μάθησης σε διάφορους τομείς.

Ωστόσο, η ύπαρξη αντιφατικών παραδειγμάτων που έχουν προκύψει ως αποτέλεσμα διαφόρων επιθέσεων έχει προκαλέσει ανησυχίες σχετικά με την εφαρμογή της βαθιάς μάθησης σε εφαρμογές κρίσιμες για την ασφάλεια της ζωής ενός ανθρώπου ή της επιβίωσης μιας επιχείρησης. Ως εκ τούτου, παρατηρείται μια συνεχής αύξηση του ενδιαφέροντος της επιστημονικής κοινότητας γύρω από τη μελέτη μεθόδων επίθεσης και άμυνας για μοντέλα ΒΝΔ σε διαφορετικούς τύπους δεδομένων, όπως εικόνες, γραφήματα και κείμενο.

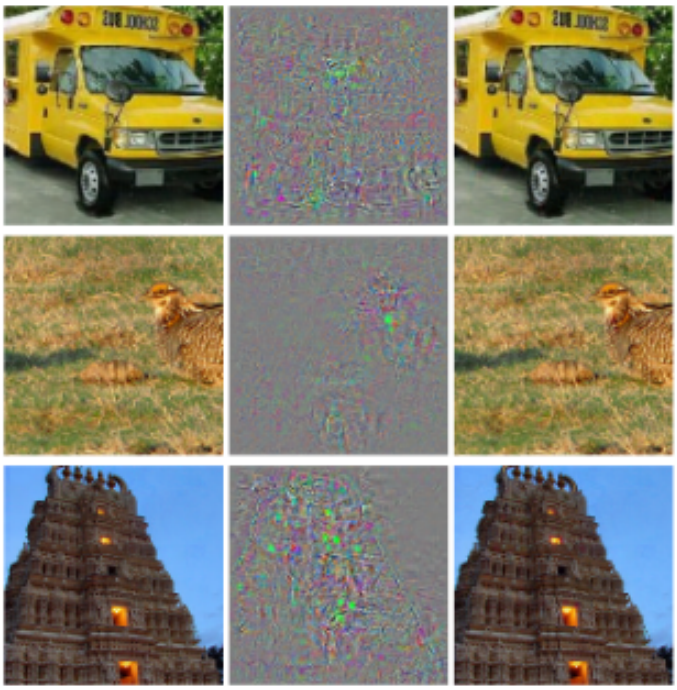
1.1 Κίνητρο

Ένα από τα σπουδαιότερα επιτεύγματα που έχουν σημειωθεί στον επιστημονικό χώρο της τεχνητής νοημοσύνης αποτελεί η ανάπτυξη δικτύων μηχανικής μάθησης. Τα βαθιά νευρωνικά δίκτυα, που αποτελούν χαρακτηριστικό παράδειγμα μηχανικής μάθησης, σταδιακά γίνονται όλο και πιο δημοφιλή καθώς αναπτύσσονται επιτυχώς όλο και περισσότερο σε διαφορετικά προβλήματα αναγνώρισης που απαντώνται κυρίως στους τομείς εικόνας, γραφημάτων, κειμένου και ομιλίας. Μάλιστα, στον τομέα της αναγνώρισης εικόνας τα ΒΝΔ είναι πλέον σε θέση να αναγνωρίζουν αντικείμενα με ακρίβεια σχεδόν ανθρώπινου επιπέδου [Kriz12b, He16].

Λόγω αυτών των επιτευγμάτων τους, τεχνικές που βασίζονται στα ΒΝΔ εφαρμόζονται συχνά σε εργασίες όπου η ασφάλεια του δικτύου κρίνεται ιδιαίτερα σημαντική, καθώς τα αποτελέσματα του καθορίζουν κρίσιμες αποφάσεις. Για παράδειγμα, στα πλαίσια ανάπτυξης αυτόνομων οχημάτων, η διαδικασία αναγνώρισης των οδικών πινακίδων [Cire12] πραγματοποιείται μέσω ορισμένων βαθιών *συνελικτικών νευρωνικών δικτύων* (convolutional neural networks - CNN) κατάλληλα εκπαιδευμένων στον τομέα της οδικής σήμανσης. Η τεχνική μηχανικής μάθησης που εφαρμόζεται σε μια τέτοια περίπτωση απαιτείται να είναι εξαιρετικά ακριβής, σταθερή και αξιόπιστη. Σε αυτό το σημείο ένα από τα πρώτα ερωτήματα που μας δημιουργείται είναι τι θα συμβεί στην περίπτωση που το μοντέλο του CNN δεν αναγνωρίσει την πινακίδα «STOP» στην άκρη του δρόμου και το όχημα συνεχίσει κανονικά την πορεία του χωρίς περαιτέρω έλεγχο. Ένα άλλο παράδειγμα εφαρμογής ΒΝΔ σε χώρους ύψιστης ασφάλειας αποτελεί το συστήματα ανίχνευσης της οικονομικής απάτης. Συχνά οι ενδιαφερόμενες εταιρείες χρησιμοποιούν *συνελικτικά δίκτυα γράφων* (graph convolutional networks - GCN) [Kipf17] προκειμένου να αποφασίσουν εάν οι πελάτες τους είναι αξιόπιστοι ή όχι. Εάν υπάρχουν απατεώνες που συγκαλύπτουν τα προσωπικά στοιχεία της ταυτότητας τους προκειμένου να αποφύγουν τον εντοπισμό τους από το σχετικό σύστημα ανίχνευσης, θα προκαλέσουν τεράστια απώλεια στην εταιρεία. Συνεπώς, η επικινδυνότητα τέτοιων καταστάσεων εγείρει πολλαπλά ερωτήματα γύρω από ζητήματα ασφαλείας των ΒΝΔ και ως εκ τούτου η δημιουργία ασφαλών ΒΝΔ αποτελεί μείζον μέλημα της επιστημονικής κοινότητας.

Την τελευταία δεκαετία, πολλές εργασίες [Szeg14, Good15, He16] έχουν δείξει ότι τα μοντέλα

ΒΝΔ είναι ευάλωτα σε αντιφατικά παραδείγματα, τα οποία μπορούν να οριστούν ως τροποποιημένα δείγματα εισόδου σε μοντέλα μηχανικής μάθησης, όπου ο εισβολέας σχεδίασε σκόπιμα προκειμένου να οδηγήσει το μοντέλο σε λανθασμένα αποτελέσματα. Ιδιαίτερα στον ευρύτατο τομέα της ταξινόμησης εικόνων, τα αντιφατικά παραδείγματα προκύπτουν από την εσκεμμένη σύνθεση τροποποιημένων εικόνων που μοιάζουν σχεδόν ολόιδιες με τις αρχικές εικόνες, όπως φαίνεται στο Σχήμα 1.1, αλλά δύνανται να παραπλανήσουν τον ταξινομητή οδηγώντας τον στο να παράγει λανθασμένα αποτελέσματα πρόβλεψης. Στο σύνολο δεδομένων MNIST [LeCu], σχεδόν όλα τα ψηφία-δείγματα μπορούν να δεχθούν επίθεση από μια ανεπαίσθητη διαταραχή που προστίθεται στην αρχική εικόνα και να παραπλανήσουν ένα καλά εκπαιδευμένο ΒΝΔ πάνω στο συγκεκριμένο σύνολο δεδομένων. Συγχρόνως, παρόμοια σχήματα επιθέσεων για τη σύγχυση των μοντέλων βαθιάς μάθησης υπάρχουν και σε άλλους τομείς εφαρμογών συμπεριλαμβανομένων γραφημάτων, κειμένου και ήχου. Η ευρεία αυτή εφαρμογή των αντιφατικών παραδειγμάτων σε όλους τους τομείς εφαρμογών αποτρέπει τους ερευνητές από το να χρησιμοποιήσουν άμεσα ΒΝΔ σε εφαρμογές με ζητήματα κρίσιμης ασφάλειας. Για την αντιμετώπιση της απειλής των αντιφατικών παραδειγμάτων, έχουν δημοσιευθεί διάφορες μελέτες με στόχο την εξεύρεση αντιμέτρων για την προστασία των βαθιών νευρωνικών δικτύων, οι οποίες αφορούν σε πρώτη φάση τεχνικές ανίχνευσης τέτοιων κακόβουλων δειγμάτων.



Σχήμα 1.1: Δείγματα αντιφατικών παραδειγμάτων (Πηγή:[Szeg14]).

Η μελέτη των επιθέσεων ενάντια σε ΒΝΔ και των αντιμέτρων τους αποτελεί βασική προϋπόθεση για την δημιουργία ασφαλών και αξιόπιστων μοντέλων βαθιάς μάθησης. Επιπλέον, πρόκειται για μια διαδικασία ιδιαίτερα ωφέλιμη για τη βαθύτερη κατανόηση της φύσης των ΒΝΔ και κατά συνέπεια της βελτίωσής τους. Για παράδειγμα, το γεγονός ότι στον τομέα της αναγνώρισης εικόνας - που συναντάται κατεξοχήν σε πολλές εφαρμογές- υπάρχουν διαταραχές οι οποίες είναι ανεπαίσθητες στα ανθρώπινα μάτια αλλά μπορούν να μπερδέψουν ένα ΒΝΔ, υποδηλώνει ότι η προσέγγιση που ακολουθεί ένα ΒΝΔ για να καταλήξει σε μια πρόβλεψη δεν συμβαδίζει απόλυτα με τον ανθρώπινο συλλογισμό.

Όλοι οι παραπάνω προβληματισμοί σε συνδυασμό με την κρισιμότητα του ζητήματος δημιουργίας εύρωστων βαθιών νευρωνικών δικτύων αποτέλεσαν το κίνητρο για την ερευνητική μας ενασχόληση με τη μελέτη επιθέσεων εναντίων των ΒΝΔ και τεχνικών αντιμετώπισης τους στα πλαίσια της αναγνώρισης εικόνας· τομέας ο οποίος απαντάται συχνότερα στις σύγχρονες εφαρμογές.

1.2 Δομή εργασίας

Η παρούσα διπλωματική εργασία δομείται σε 6 Κεφάλαια. Η Εισαγωγή (Κεφάλαιο 1) στοχεύει να δώσει στον αναγνώστη να κατανοήσει τη σπουδαιότητα του προβλήματος και τα αίτια που μας οδήγησαν σε περαιτέρω διερεύνησή του. Στο Κεφάλαιο 2 καλύπτεται το θεωρητικό υπόβαθρο γύρω από τα νευρωνικά δίκτυα προκειμένου να είναι δυνατή η κατανόηση της αρχιτεκτονικής των συστημάτων που θα παρουσιαστούν σε επόμενες ενότητες και γίνεται σαφής ο διαχωρισμός μεταξύ νευρωνικών δικτύων και βαθιάς μηχανικής μάθησης με περαιτέρω ανάλυση της δεύτερης. Στο Κεφάλαιο 3, παρουσιάζονται αναλυτικά οι διάφορες κατηγοριοποιήσεις των γνωστών μέχρι σήμερα επιθέσεων σε ΒΝΔ και συγχρόνως δίνονται εν συντομία περιγραφές των αλγορίθμων τους καθώς και σχολιασμός των αποτελεσμάτων τους. Στο Κεφάλαιο 4, παρουσιάζονται οι βασικές τεχνικές άμυνας απέναντι στις επιθέσεις του προηγούμενου Κεφαλαίου. Στο Κεφάλαιο 5, αναλύονται σε βάθος τα πειράματα που πραγματοποιήθηκαν, τα σύνολα δεδομένων και οι αρχιτεκτονικές που χρησιμοποιήθηκαν, οι λεπτομέρειες υλοποίησης του συστήματος καθώς και τα αποτελέσματα που προέκυψαν. Τέλος, στο Κεφάλαιο 6 ολοκληρώνεται η διπλωματική εργασία, παρουσιάζοντας συγκεντρωτικά τα συμπεράσματα που διεξήχθησαν από το σύνολο της διαδικασίας και γίνεται αναφορά σε σκέψεις για μελλοντικές ερευνητικές κατευθύνσεις.

Στο Παράρτημα Α παρουσιάζονται ορισμένα συμπληρωματικά αποτελέσματα της πειραματικής διαδικασίας που κρίνονται απαραίτητα για την ολοκληρωμένη παρουσίαση της παρούσας εργασίας, συμβάλλοντας στη βαθύτερη κατανόηση ορισμένων πρακτικών αποφάσεων αλλά και συμπερασμάτων.

Κεφάλαιο 2

Νευρωνικά Δίκτυα

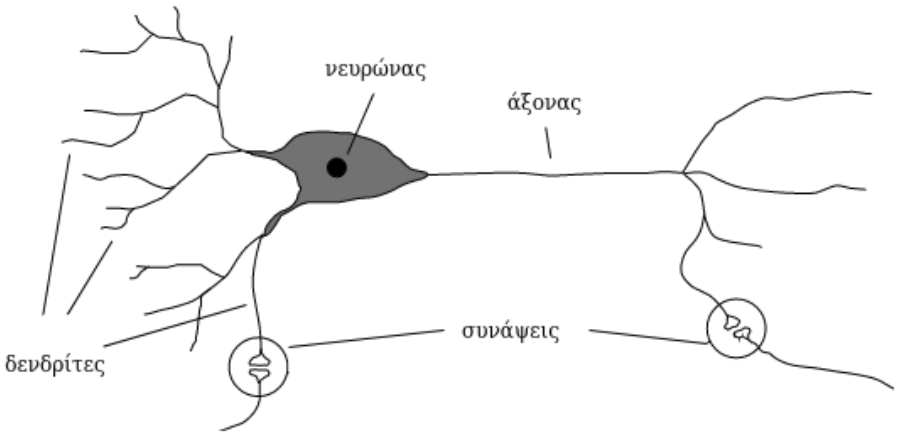
Η δομή και η λειτουργία των νευρωνικών δικτύων των ζωντανών οργανισμών αποτέλεσαν πηγή έμπνευσης και κυρίως τη βάση για την κατασκευή υπολογιστικών νευρωνικών δικτύων που μπορούν να επιτελούν ένα πλήθος από διεργασίες και να λύνουν ικανοποιητικά πολλά προβλήματα.

Το 1943, οι McCulloch και Pitts προσπάθησαν να καταλάβουν πώς ο ανθρώπινος εγκέφαλος θα μπορούσε να παράγει πολύπλοκα μοτίβα μέσω συνδεδεμένων εγκεφαλικών κυττάρων ή νευρώνων [MCCU43]. Μία από τις κύριες ιδέες που προέκυψαν από την έρευνα τους ήταν η σύγκριση των νευρώνων με την Boolean λογική. Την επόμενη δεκαετία ο Rosenblatt επεκτείνοντας το έργο των προαναφερθέντων, δημοσιεύει την έρευνα του πάνω στο perceptron [ROSE58]. Πρόκειται για ένα πιθανοτικό μοντέλο με βάση το οποίο ένας υπολογιστής δύναται να μάθει να διακρίνει τις κάρτες που έχουν σημειωθεί στην αριστερή πλευρά έναντι των καρτών που έχουν σημειωθεί στη δεξιά πλευρά. Το 1974, ο Werbos υπήρξε από τους πρώτους που εφάρμοσε την τεχνική της προς τα πίσω διάδοσης του σφάλματος (backpropagation) σε νευρωνικά δίκτυα [Werb74]. Μερικά χρόνια αργότερα, το 1989, ο LeCun ανέλυσε τον τρόπο με τον οποίο η χρήση περιορισμών στην προς τα πίσω διάδοση του σφάλματος και η ενσωμάτωσή της τεχνικής αυτής στην αρχιτεκτονική του νευρωνικού δικτύου μπορούν να χρησιμοποιηθούν για την εκπαίδευση αλγορίθμων [YLeC89]. Μέσω αυτής της έρευνας υλοποιήθηκε επιτυχώς ένα νευρωνικό δίκτυο αναγνώρισης χειρόγραφων ψηφίων ταχυδρομικού κώδικα.

2.1 Βιολογικά νευρωνικά δίκτυα

Ο ανθρώπινος εγκέφαλος αποτελείται από δισεκατομμύρια διασυνδεδεμένους νευρώνες. Όπως χαρακτηριστικά αναφέρεται στην έρευνά του [Bail01], πρόκειται για κύτταρα που έχουν εξειδικευμένες μεμβράνες που επιτρέπουν τη μετάδοση σημάτων σε γειτονικούς νευρώνες. Ένας μεμονωμένος άξονας εκτείνεται από το βασικό κυτταρικό σώμα του νευρώνα, ο οποίος συνήθως διακλαδίζεται έντονα σε πολλαπλές απολήξεις, πριν σχηματίσει μια σύναψη με δενδρίτη άλλου νευρώνα. Οι ηλεκτρικοί παλμοί διαδίδονται κατά μήκος του άξονα καταλήγοντας στις συνάψεις, όπως φαίνεται στο Σχήμα 2.1. Η επικοινωνία των νευρώνων επιτυγχάνεται με χημικό τρόπο μέσω της ταχύτατης έκκρισης μορίων νευροδιαβιβαστών. Κατά την άφιξη ενός ηλεκτρικού παλμού στην απόληξη του νευράξονα, διεγείρεται η απελευθέρωση χημικών νευροδιαβιβαστών στο συναπτικό κενό. Μέσω αυτής της διαδικασίας, το προ-συναπτικό νευρωνικό κύτταρο - το οποίο απελευθερώνει το νευροδιαβιβαστή - μπορεί να επάγει στο μετα-συναπτικό κύτταρο - το οποίο αποτελεί μέρος λήψης του επόμενου νευρώνα - μια ηλεκτρική διέγερση που θα διαβιβαστεί μέσω του δενδρίτη στο κύριο μέρος του σώματος του επόμενου νευρώνα. Στη συνέχεια, οι είσοδοι από τους διαφορετικούς δενδρίτες συνδυάζονται για να παράγουν ένα σήμα εξόδου το οποίο περνά κατά μήκος του άξονα και έτσι η διαδικασία επαναλαμβάνεται. Ωστόσο, ένα σήμα παράγεται στον άξονα μόνο εάν υπάρχουν αρκετές είσοδοι επαρκούς αντοχής ώστε να ξεπεραστεί η τιμή κατωφλίου και η έξοδος είναι μη γραμμική συνάρτηση των ερεθισμάτων εισόδου.

Ο ανθρώπινος εγκέφαλος αποτελείται από περίπου 10^{11} νευρώνες οι οποίοι συνδέονται μεταξύ τους 10.000 συνάψεις κατά μέσο όρο. Ο ισχυρός δεσμός της συναπτικής σύνδεσης μεταξύ των νευρώνων μπορεί να αλλάξει χημικά από τον εγκέφαλο σε απάντηση ευνοϊκών και δυσμενών ερεθισμάτων, με τέτοιο τρόπο ώστε να προσαρμόζει τον οργανισμό να λειτουργεί βέλτιστα στο περιβάλλον του. Για



Σχήμα 2.1: Δομή ενός εγκεφαλικού νευρώνα (**Πηγή**: [Bail01])

τον λόγο αυτό άλλωστε είναι κοινώς παραδεκτό από διάφορους επιστημονικούς χώρους ότι οι συνάψεις αποτελούν το κλειδί για τη μάθηση σε βιολογικά συστήματα. Ένα τεχνητό νευρωνικό δίκτυο αποτελεί μίμηση του φυσικού νευρικού δικτύου, όπου οι τεχνητοί νευρώνες συνδέονται με παρόμοιο τρόπο όπως το εγκεφαλικό δίκτυο [Shir16].

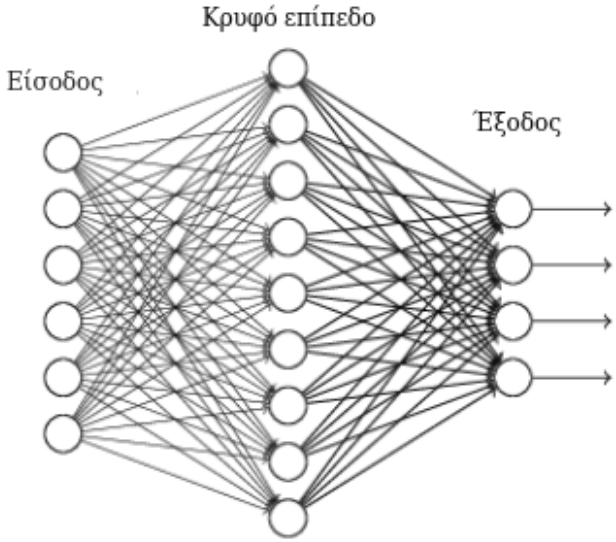
2.2 Τεχνητά Νευρωνικά Δίκτυα

2.2.1 Γενική περιγραφή

Τα *τεχνητά νευρωνικά δίκτυα* (artificial neural networks), γνωστά και ως απλώς νευρωνικά δίκτυα, αποτελούν υποσύνολο της μηχανικής μάθησης και βασικό παράγοντα στην δημιουργία αλγορίθμων βαθιάς μάθησης. Ο ανθρώπινος εγκέφαλος αποτέλεσε προϊόν έμπνευσης τόσο για την ονομασία όσο και για τη δομή των νευρωνικών δικτύων. Σε εννοιολογικό επίπεδο, το μοντέλο ενός νευρωνικού δικτύου αποτελείται από νευρώνες, που ονομάζονται επίσης μονάδες ή κόμβοι. Οι τελευταίοι μπορούν να συλλέξουν πληροφορίες είτε από κάποιον εξωτερικό παράγοντα είτε από άλλους νευρώνες και στη συνέχεια μπορούν να τις μεταδώσουν κατ' αντιστοιχία είτε σε άλλους νευρώνες είτε να τις βγάλουν ως τελικό αποτέλεσμα.

Μια βασική διάκριση γίνεται μεταξύ των νευρώνων εισόδου, των κρυφών νευρώνων και των νευρώνων εξόδου. Η πρώτη κατηγορία νευρώνων λαμβάνει πληροφορίες με τη μορφή μοτίβων ή σημάτων από το εξωτερικό περιβάλλον του κλειστού νευρωνικού δικτύου. Στη δεύτερη κατηγορία βρίσκονται οι κρυμμένοι νευρώνες, οι οποίοι τοποθετούνται αρχιτεκτονικά μεταξύ των νευρώνων εισόδου και εξόδου, όπως χαρακτηριστικά μπορούμε να διακρίνουμε στο Σχήμα 2.2. Οι νευρώνες της συγκεκριμένης κλάσης αποτελούν τον πυρήνα της λειτουργίας του νευρωνικού δικτύου, χαρτογραφώντας εσωτερικά μοτίβα πληροφοριών. Το ερώτημα σχετικά με το ποιο είναι το βέλτιστο πλήθος κρυφών νευρώνων ενός δικτύου παραμένει ανοιχτό. Παρόλα αυτά, ερευνητές έχουν αποδείξει ότι ένα ΤΝΔ με ένα απλό κρυφό επίπεδο μπορεί να προσεγγίσει οποιαδήποτε συνάρτηση, αρκεί να έχει αρκετούς νευρώνες. Τέλος, οι νευρώνες εξόδου αποτελούν το δίαυλο επικοινωνίας με τον έξω κόσμο γνωστοποιώντας και μεταδίδοντας το αποτέλεσμα υπό τη μορφή πληροφοριών και σημάτων.

Οι διάφοροι νευρώνες συνδέονται μεταξύ τους μέσω των λεγόμενων ακμών. Συνεπώς, η έξοδος ενός νευρώνα μπορεί να γίνει η είσοδος του επόμενου νευρώνα. Σε κάθε ακμή αντιστοιχίζεται ένα συγκεκριμένο βάρος το οποίο αντιπροσωπεύει το πόσο ισχυρός είναι ο δεσμός μεταξύ των νευρώνων και καθορίζεται από το ρόλο του δεσμού αυτού στην επίλυση του ευρύτερου στόχου του δικτύου. Όσο μεγαλύτερο είναι το βάρος της ακμής τόσο μεγαλύτερη είναι η επίδραση που μπορεί να ασκήσει ένας νευρώνας στη σύνδεση με τον άλλο νευρώνα. Τα βάρη δύναται να πάρουν θετικές, αρνητικές ή μηδενικές τιμές επιδρώντας αντίστοιχα ενθαρρυντικά, ανασταλτικά ή ουδέτερα στο νευρωνικό δίκτυο. Ένα μηδενικό βάρος συνεπάγεται ότι ο ένας νευρώνας δεν ασκεί καμία επίδραση στη σύνδεση



Σχήμα 2.2: Βασική δομή Νευρωνικού δικτύου (**Πηγή**: [Niel15])

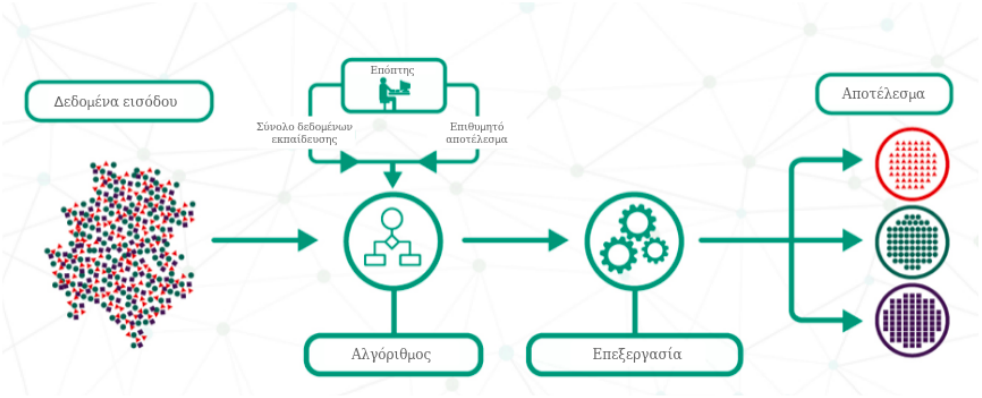
με τον άλλο νευρώνα. Η γνώση και κατ' επέκταση η τεχνητή νοημοσύνη ενός νευρωνικού δικτύου αποθηκεύονται τελικά στις συνδέσεις μεταξύ των νευρώνων και στα βάρη των ακμών τους. Ανάλογα με τον τρόπο με τον οποίο συνδέονται οι νευρώνες μεταξύ τους ορίζουν διαφορετικά νευρωνικά επίπεδα, έτσι ώστε οι νευρώνες που βρίσκονται στο ίδιο επίπεδο να μην συνδέονται μεταξύ τους, αλλά να συνδέονται με όλους τους νευρώνες του επόμενου επιπέδου.

2.2.2 Μάθηση

Προτού χρησιμοποιηθεί ένα νευρωνικό δίκτυο για την επίλυση του επικείμενου προβλήματος θα πρέπει πρώτα να εκπαιδευτεί. Δεδομένων δειγμάτων εισόδου και δοσμένων κανόνων εκμάθησης, το νευρωνικό δίκτυο προσπαθεί να αναπτύξει συγκεκριμένη «νοημοσύνη». Οι κανόνες μάθησης υπαγορεύουν τον τρόπο με τον οποίο το εκπαιδευτικό υλικό, που εισάγεται ως είσοδος στο μοντέλο, μεταβάλλει το νευρωνικό δίκτυο.

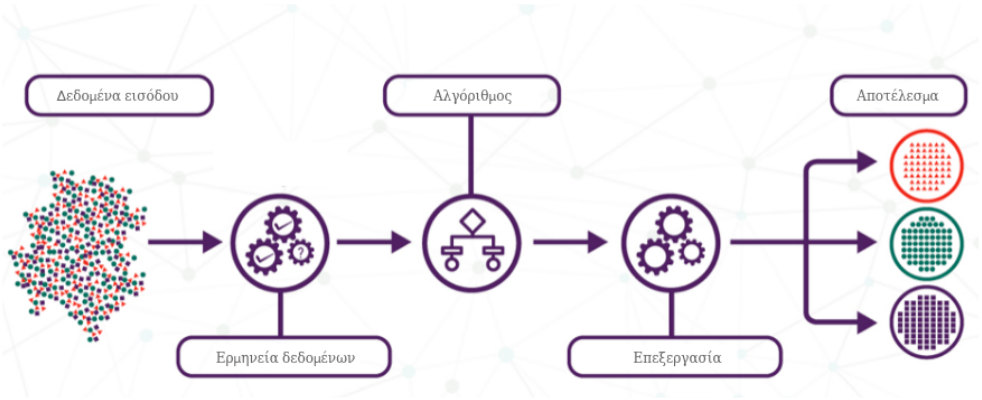
Καταρχήν, η μηχανική μάθηση κατηγοριοποιείται στην *επιβλεπόμενη* και στη *μη-επιβλεπόμενη* μάθηση. Στην πρώτη περίπτωση, όπως φαίνεται και στο Σχήμα 2.3, ζεύγη δειγμάτων και ετικετών τους δίνονται ως είσοδοι στο νευρωνικό δίκτυο. Με βάση τη συνεχή σύγκριση μεταξύ των ετικετών και του αποτελέσματος που υπολογίζει το νευρωνικό, το δίκτυο μαθαίνει να ρυθμίζει κατάλληλα τα βάρη που συνδέουν τους νευρώνες. Συνεπώς, στην επιβλεπόμενη μάθηση θα πρέπει κατά την εκπαίδευση να υπάρχει πάντα ένας εξωτερικός εκπαιδευτής. Συνηθέστερη μέθοδος είναι η μάθηση με διόρθωση σφάλματος.

Αντίθετα, στη μη-επιβλεπόμενη μάθηση, η μαθησιακή διαδικασία βασίζεται αποκλειστικά στις πληροφορίες των πολλών διαφορετικών προτύπων που έχουν εισαχθεί στο δίκτυο. Πιο συγκεκριμένα, παρατηρώντας το Σχήμα 2.4 βλέπουμε ότι το νευρωνικό δίκτυο πραγματοποιεί τις αλλαγές μόνο με βάση τα μοτίβα εισόδου, καθώς πλέον δεν δίνονται ως είσοδοι ζεύγη δειγμάτων και ετικετών. Το σύστημα προσπαθεί να ανακαλύψει στατιστικά εμφανή χαρακτηριστικά του πληθυσμού εισόδου. Σε αντίθεση με την επιβλεπόμενη μάθηση, δεν υπάρχει εκ των προτέρων κάποιο σύνολο κατηγοριών στις οποίες ταξινομούνται τα πρότυπα. Συνεπώς, το σύστημα πρέπει να αναπτύξει τη δική του αναπαράσταση των ερεθισμάτων εισόδου. Για τον σκοπό αυτό υπάρχουν διάφοροι κανόνες μάθησης, όπως η θεωρία προσαρμοστικού συντονισμού ή ο κανόνας μάθησης του Hebb. Πλέον, η διαδικασία ανανέωσης των βαρών δεν απαιτεί τον υπολογισμό του σφάλματος. Στην εκπαίδευση με διόρθωση



Σχήμα 2.3: Επιβλεπόμενη μάθηση (**Πηγή**: [onlia]).

σφάλματος τα βάρη ανανεώνονται με τρόπο ώστε η ολική έξοδος του δικτύου να πλησιάζει όσο το δυνατόν περισσότερο την επιθυμητή, ενώ αντίθετα στην μάθηση Hebb η ενεργοποίηση ενός νευρώνα συνεπάγεται την ενεργοποίηση ενός άλλου και κατ' επέκταση η σύναψη μεταξύ των δυο νευρώνων ενισχύεται. Με άλλα λόγια το βάρος μεταξύ δύο νευρώνων αυξάνεται αναλογικά με το πόσο ισχυρή είναι η συσχέτιση μεταξύ τους.



Σχήμα 2.4: Μη-επιβλεπόμενη μάθηση (**Πηγή**: [Soft19]).

Ο αριθμός των νευρώνων και των νευρωνικών επιπέδων καθώς και η διασύνδεση των νευρώνων μεταξύ των διαφορετικών επιπέδων καθορίζει την πολυπλοκότητα ή αλλιώς το βάθος του νευρωνικού δικτύου και την ικανότητά του να επιλύει προβλήματα. Κατά τη διάρκεια της εκπαίδευσης του νευρωνικού δικτύου τα βάρη μεταξύ των συνδέσεων αλλάζουν, ανάλογα με τους κανόνες μάθησης που έχουν τεθεί και τα ληφθέντα αποτελέσματα. Σε θεωρητικό επίπεδο, ο αριθμός των νευρώνων σε ένα τεχνητό νευρωνικό δίκτυο δύναται να είναι απεριόριστος. Στην πράξη ωστόσο, ο αριθμός των νευρώνων, των νευρωνικών επιπέδων και των συνδέσεων αυξάνουν την απαιτούμενη υπολογιστική ισχύ για την εκπαίδευση και λειτουργία του δικτύου.

2.2.3 Συναρτήσεις ενεργοποίησης

Τα τεχνητά νευρωνικά δίκτυα έχουν τη δυνατότητα γενίκευσης και λήψης αποφάσεων από μεγάλα και κάπως ασαφή δεδομένα εισόδου. Συγχρόνως, μπορούν να μοντελοποιούν μηχανισμούς κατάλληλα εξειδικευμένους στην επίλυση μη γραμμικών προβλημάτων.

Βηματική συνάρτηση (Step function)

Η απλούστερη συνάρτηση που χρησιμοποιεί ένα νευρωνικό προκειμένου να προσομοιάσει τη λειτουργία του εγκεφάλου κατά την οποία οι νευρώνες πυροδοτούνται μόνο πάνω από ένα ορισμένο κατώφλι, είναι η βηματική όπως ορίζεται στην Εξίσωση 2.1.

$$\varphi(\vec{w} \cdot \vec{x} + b) = \begin{cases} 1 & \text{, if } \vec{w} \cdot \vec{x} + b \geq 0 \\ 0 & \text{, if } \vec{w} \cdot \vec{x} + b < 0 \end{cases} \tag{2.1}$$

όπου $\vec{w}$ το διάνυσμα βαρών των δεσμών μεταξύ του δεδομένου νευρώνα και όσων συνδέονται ως είσοδοι μαζί του, $\vec{x}$ οι είσοδοι του νευρώνα και b η τιμή κατωφλίου.

Η δεδομένη συνάρτηση ενεργοποίησης στην ουσία αποτελεί έναν γραμμικό ταξινομητή, ο οποίος στις δύο διαστάσεις μπορεί να ερμηνευθεί ως μια ευθεία με κλίση $\vec{w}$ και μετατόπιση b ως προς τον $y'y$ άξονα. Ρυθμίζοντας κατάλληλα τις τιμές των $\vec{w}$ και b, η βηματική συνάρτηση μπορεί να προσαρμόσει το γραμμικό της όριο και να μάθει να χωρίζει τις εισόδους της σε κλάσεις 0 και 1. Ως επακόλουθο, διαφορετικές τιμές των $\vec{w}$ και b για πολλαπλές μονάδες βηματικής συνάρτησης θα παράγουν πολλαπλούς διαφορετικούς γραμμικούς ταξινομητές. Ένα από τα σημαντικότερα πλεονεκτήματα των ΤΝΔ είναι η ικανότητά τους να προσαρμόζουν τα $\vec{w}$ και b για πολλές μονάδες ταυτόχρονα, μαθαίνοντας έτσι αποτελεσματικά πολλούς γραμμικούς ταξινομητές.

Ωστόσο, η διαδικασία σταδιακής εκμάθησης του δικτύου προϋποθέτει τη σταδιακή αλλαγή των βαρών ή/και των πολώσεων. Το πρόβλημα έγκειται στο γεγονός ότι όταν το δίκτυο χρησιμοποιεί τη βηματική συνάρτηση ενεργοποίησης μια μικρή μεταβολή στο διάνυσμα των βαρών ή στην πόλωση μπορεί να προκαλέσει πλήρη αναστροφή της εξόδου. Συνεπώς, είναι δύσκολο να προβλέψουμε πώς να τροποποιούμε σταδιακά τα βάρη και τις πολώσεις, έτσι ώστε το δίκτυο να πλησιάζει περισσότερο στην επιθυμητή συμπεριφορά.

Σιγμοειδής ή λογιστική συνάρτηση ενεργοποίησης (Sigmoid function)

Μια συνεχής συνάρτηση η οποία είναι παρόμοια με την βηματική αλλά μπορεί να προσεγγίσει καθολικά στο σύνολο τους γραμμικά και μη γραμμικά προβλήματα είναι η σιγμοειδής. Η έξοδος αυτής της συνάρτησης κυμαίνεται σε όλες τις τιμές μεταξύ 0 και 1 και ορίζεται σύμφωνα με την Εξίσωση 2.2. Πρόκειται επί της ουσίας για μία εξομαλυμένη μορφή της βηματικής συνάρτησης. Επομένως, σε μια υπολογιστική μονάδα που χρησιμοποιεί τη σιγμοειδή συνάρτηση, αντί να ενεργοποιεί το 0 ή το 1 όπως στην περίπτωση της βηματικής, η έξοδος της θα είναι μεταξύ 0 και 1 χωρίς να λαμβάνει ποτέ τις τιμές 0,1 αυτές καθ' αυτές. Το γεγονός αυτό αλλάζει ελαφρώς την ερμηνεία του μοντέλου του νευρώνα, καθώς αποκλίνει πλέον από τη συμπεριφορά «όλα-ή-τίποτα» του βιολογικού νευρώνα που εφαρμόζεται μέσω μιας τιμής κατωφλίου. Ωστόσο, η σιγμοειδής συνάρτηση είναι πολύ κοντά στο 0 για $\vec{x} < 0$ και πολύ κοντά στο 1 για $\vec{x} > 0$, οπότε μπορεί να ερμηνευθεί όπως η βηματική συνάρτηση καθώς στην πράξη εμφανίζει την επιθυμητή συμπεριφορά σε όλες τις μη μηδενικές εισόδους. Στην ουσία, η σιγμοειδής ή αλλιώς λογιστική συνάρτηση μπορεί να προσομοιαστεί μέσω ενός γραμμικού ταξινομητή με όριο στο $\vec{w} \cdot \vec{x} + b = 0$. Αυτό σημαίνει ότι μικρές αλλαγές στα βάρη και στη πόλωση θα μας δώσουν μικρές αλλαγές και στην τιμή εξόδου του δικτύου. Η τιμή της συνάρτησης στο γραμμικό όριο είναι $\sigma(0) = 0.5$. Οι είσοδοι $\vec{x}$ που βρίσκονται κοντά στο όριο πλησιάζουν την τιμή 0.5, ενώ οι είσοδοι που είναι πιο απομακρυσμένοι λαμβάνουν τιμές πολύ κοντά στο 0 ή στο 1.

$$\sigma(\vec{w} \cdot \vec{x} + b) = \frac{1}{1 + \exp(-(\vec{w} \cdot \vec{x} + b))} \tag{2.2}$$

όπου $\vec{w}$ το διάνυσμα βαρών των δεσμών μεταξύ του δεδομένου νευρώνα και όσων συνδέονται ως είσοδοι μαζί του, $\vec{x}$ οι είσοδοι του νευρώνα και b η τιμή κατωφλίου.

Στην πράξη, τα περισσότερα σύγχρονα ΤΝΔ χρησιμοποιούν μία από τις τρεις συναρτήσεις ενεργοποίησης που παρατίθενται εν συντομία στη συνέχεια.

Συνάρτηση ενεργοποίησης υπερβολικής εφαπτομένης

Μια συνήθης συνάρτηση ενεργοποίησης, εναλλακτική της σιγμοειδούς, είναι η υπερβολική εφαπτομένη. Παρόλο που σχηματικά είναι παρόμοια με την εκθετική σιγμοειδή συνάρτηση, όπως φαίνεται στο Σχήμα 2.5, το μεγαλύτερο εύρος του πεδίου τιμών της καθώς και το γεγονός ότι επιστρέφει και αρνητικές τιμές αποτελεί σημαντικό πλεονέκτημα. Από την Εξίσωση 2.3 είναι φανερό ότι η υπερβολική εφαπτομένη επιστρέφει ως έξοδο έναν πραγματικό αριθμό στο $[-1, 1]$.Το γεγονός ότι οι τιμές εξόδου είναι κεντραρισμένες στο μηδέν, την κάνει συχνά προτιμότερη της σιγμοειδούς συνάρτησης.

$$\varphi_{tanh}(\vec{w} \cdot \vec{x} + b) = \frac{\sin(\vec{w} \cdot \vec{x} + b)}{\cos(\vec{w} \cdot \vec{x} + b)} = \frac{\exp(\vec{w} \cdot \vec{x} + b) - \exp(-(\vec{w} \cdot \vec{x} + b))}{\exp(\vec{w} \cdot \vec{x} + b) + \exp(-(\vec{w} \cdot \vec{x} + b))} \tag{2.3}$$

όπου $\vec{w}$ το διάνυσμα βαρών των δεσμών μεταξύ του δεδομένου νευρώνα και όσων συνδέονται ως είσοδοι μαζί του, $\vec{x}$ οι είσοδοι του νευρώνα και b η τιμή κατωφλίου.

Ανορθωμένη γραμμική συνάρτηση ράμπας

Η *ανορθωμένη γραμμική συνάρτηση ράμπας* (rectified linear unit - ReLU) ορίζεται στην Εξίσωση 2.4 και χρησιμοποιείται κατά κόρων στα κρυφά στρώματα των νευρωνικών δικτύων. Έχει την δυνατότητα να εκπαιδεύει ένα δίκτυο αρκετά γρήγορα, δίνοντας συγχρόνως ακριβή αποτελέσματα. Η αυξημένη απόδοση της οφείλεται στο γεγονός ότι η ReLU είναι μια μη-κορεσμένη γραμμική συνάρτηση. Πιο συγκεκριμένα, αντιστοιχίζει την τιμή εισόδου με έναν θετικό πραγματικό αριθμό ή μηδέν. Για αρνητικές τιμές εισόδου η έξοδος είναι 0 ενώ για θετικές εισόδους η έξοδος είναι ίδια με την είσοδο.

$$\varphi(\vec{w} \cdot \vec{x} + b) = \begin{cases} \vec{w} \cdot \vec{x} + b & , \text{αν } \vec{w} \cdot \vec{x} + b > 0 \\ 0 & , \text{αν } \vec{w} \cdot \vec{x} + b \leq 0 \end{cases} \tag{2.4}$$

όπου $\vec{w}$ το διάνυσμα βαρών των δεσμών μεταξύ του δεδομένου νευρώνα και όσων συνδέονται ως είσοδοι μαζί του, $\vec{x}$ οι είσοδοι του νευρώνα και b η τιμή κατωφλίου.

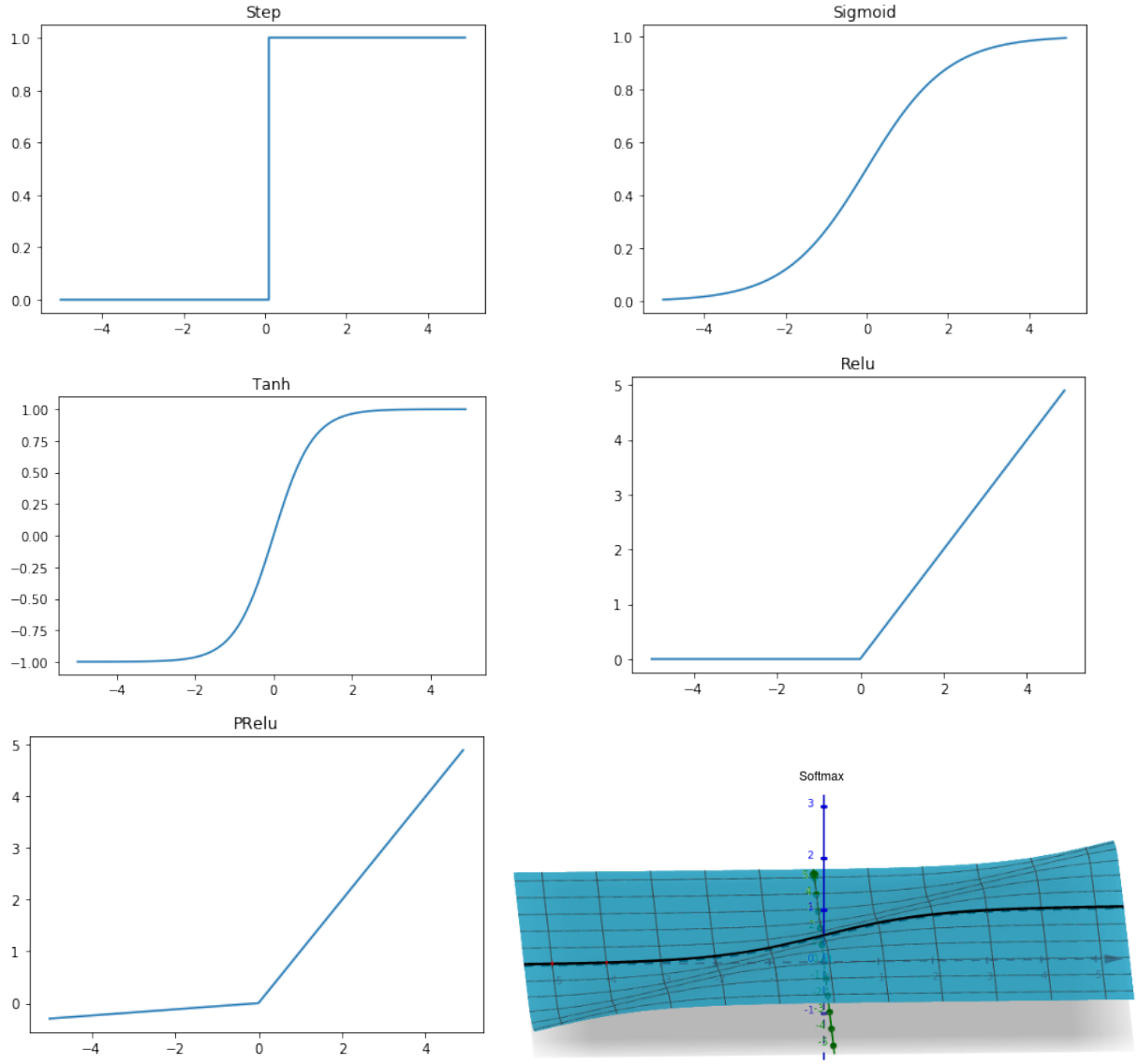
Το γεγονός ότι η ReLU δεν διαθέτει κάποιο σημείο κορεσμού για τις θετικές εισόδους - σε αντίθεση με τις τρεις προηγούμενες συναρτήσεις ενεργοποίησης - αποτελεί σημαντικό πλεονέκτημα και έναν από τους βασικότερους λόγους που αιτιολογεί την εκτεταμένη χρήση της. Ωστόσο, ο κίνδυνος ότι ορισμένοι νευρώνες του δικτύου μπορεί να «νεκρώσουν» κατά τη διαδικασία της εκπαίδευσης είναι υπαρκτός, καθώς κάποιες φορές η συνάρτηση ράμπας μπορεί να οδηγήσει ορισμένους νευρώνες σε τιμές βαρών, οι οποίες τους αποτρέπουν από το να ενεργοποιηθούν. Η παραμετροποιημένη συνάρτηση ράμπας είναι μια τροποποιημένη εκδοχή της ReLU, η οποία δίνει τη λύση στον παραπάνω προβληματισμό. Η συνάρτηση αυτή, στην περίπτωση που η τιμή εισόδου είναι αρνητική, πολλαπλασιάζει την τιμή εξόδου με μία μικρή θετική σταθερά η οποία διαφέρει από δίκτυο σε δίκτυο.

Συνάρτηση softmax

Σε προβλήματα κατηγοριοποίησης χρησιμοποιείται συχνά μια γενίκευση της σιγμοειδούς συνάρτησης, η softmax. Η συνάρτηση αυτή, όπως ορίζεται στην Εξίσωση 2.5, συναντάται συνήθως στο τελευταίο επίπεδο του νευρωνικού δικτύου και αντιστοιχίζει κάθε τιμή εισόδου με μία τιμή εξόδου στο $[0, 1]$. Η διαφορά με τη σιγμοειδή συνάρτηση έγκειται στο γεγονός ότι η softmax εξαναγκάζει το άθροισμα των τιμών εξόδου να ισούται πάντα με τη μονάδα, έτσι ώστε κάθε τιμή εξόδου να περιέχει την πιθανότητα η είσοδος που δόθηκε στο δίκτυο να ανήκει στην αντίστοιχη κλάση. Χωρίς τη χρήση της softmax στο επίπεδο εξόδου του δικτύου, οι έξοδοι των νευρώνων είναι απλά αριθμητικές τιμές, με την υψηλότερη ένδειξη να αντιστοιχεί στην επικρατούσα κλάση.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_j e^{z_j}}, \tag{2.5}$$

όπου $i = 1, ...,$ και $\vec{z}$ διάνυσμα με $\vec{z} = (z_{1,...,z_K}) \in R^K$.



Σχήμα 2.5: Συναρτήσεις ενεργοποίησης (από αριστερά προς τα δεξιά και από πάνω προς τα κάτω): (i) βηματική συνάρτηση, (ii) σιγμοειδής συνάρτηση, (iii) συνάρτηση υπερβολικής εφαπτομένης, (iv) ανορθωμένη γραμμική συνάρτηση ράμπας και (v) συνάρτηση softmax (**Πηγή**: [Refa]).

2.2.4 Βασικά σημεία εκπαίδευσης δικτύων

Προς τα εμπρός διάδοση

Με τον όρο *προς τα εμπρός διάδοση* (forward propagation) καθορίζουμε ότι η ροή διάδοσης της πληροφορίας εντός του νευρωνικού δικτύου είναι προς μια κατεύθυνση και συγκεκριμένα από το επίπεδο εισόδου προς το επίπεδο εξόδου. Τα δεδομένα εισόδου εισέρχονται στο νευρωνικό δίκτυο μέσω του επιπέδου εισόδου, επεξεργάζονται και μέσω των κρυφών επιπέδων καταλήγουν στο επίπεδο εξόδου, ακολουθώντας αποκλειστικά αυτή τη μονόδρομη πορεία. Όπως έχει ήδη αναφερθεί σε προηγούμενη ενότητα, σε κάθε νευρώνα, λαμβάνοντας υπόψη τα δεδομένα εισόδου του, τις τιμές των βαρών του, τη συνολική του πόλωση και κάνοντας χρήση μιας συνάρτησης ενεργοποίησης μπορούμε να παράγουμε την έξοδο του δεδομένου νευρώνα που θα αποτελέσει είσοδο για νευρώνες επόμενου

επιπέδου. Η παραπάνω διαδικασία, ξεκινώντας από το πρώτο επίπεδο με είσοδο την αρχική πληροφορία, επαναλαμβάνεται για κάθε νευρώνα του δικτύου μέχρις ότου παραχθεί στο τελευταίο επίπεδο το τελικό αποτέλεσμα.

Συνάρτηση κόστους

Ένα νευρωνικό δίκτυο βαθιάς μάθησης εκπαιδεύεται να αντιστοιχεί ένα σύνολο δεδομένων εισόδων σε ένα σύνολο δεδομένων εξόδων (ετικέτες) κάνοντας χρήση του συνόλου δεδομένων εκπαίδευσης. Η εύρεση των κατάλληλων βαρών και της αντίστοιχης πόλωσης για κάθε νευρώνα του δικτύου θα μας επιστρέψει ένα μοντέλο που θα εκτελεί ορθά τη διαδικασία της επιθυμητής αντιστοίχισης. Ωστόσο στην πράξη, ο υπολογισμός των ιδανικών βαρών ενός δικτύου είναι αδύνατος καθώς υπάρχουν πολλοί άγνωστοι παράμετροι.

Αντ' αυτού, το πρόβλημα της μάθησης μετατρέπεται σε πρόβλημα βελτιστοποίησης και για την επίλυσή του απαιτείται ένας αλγόριθμος πλοήγησης στο χώρο των πιθανών συνόλων βαρών που μπορεί να χρησιμοποιήσει το μοντέλο για να κάνει καλές ή αρκετά καλές προβλέψεις. Στο πλαίσιο του αλγορίθμου βελτιστοποίησης, η συνάρτηση που χρησιμοποιείται για την αξιολόγηση μιας υποψήφιας λύσης - δηλαδή ενός συνόλου βαρών - αναφέρεται ως η *αντικειμενική συνάρτηση* (objective function). Μπορεί να επιδιώξουμε να μεγιστοποιήσουμε ή να ελαχιστοποιήσουμε την αντικειμενική συνάρτηση, που σημαίνει ότι αναζητούμε μια υποψήφια λύση που έχει την υψηλότερη ή τη χαμηλότερη τιμή, αντίστοιχα. Συνήθως, στα νευρωνικά δίκτυα, επιδιώκουμε να ελαχιστοποιήσουμε το σφάλμα. Ως εκ τούτου, η αντικειμενική συνάρτηση αναφέρεται συχνά ως *συνάρτηση κόστους* (cost function) ή *συνάρτηση σφάλματος* (loss function) και η τιμή που υπολογίζεται από αυτή τη συνάρτηση αποτελεί το σφάλμα του δικτύου.

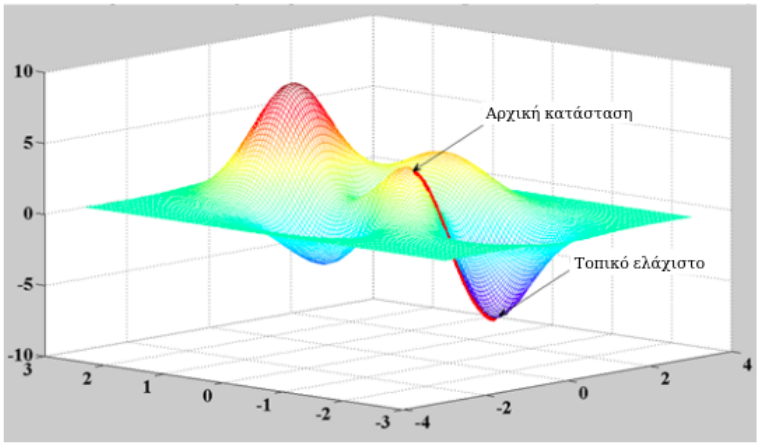
Στην ουσία, μέσω της συνάρτησης κόστους προσπαθούμε να προσομοιώσουμε την επιβολή κάποιας «διόρθωσης» στο δίκτυο, στην περίπτωση πιθανών λαθών, έχοντας πάντα υπόψιν ότι στη γενική περίπτωση η αρχικοποίηση των παραμέτρων βαρών και πόλωσης ενός δικτύου γίνεται με τυχαίο τρόπο. Κύριος στόχος μας είναι η αύξηση της ακρίβειας πρόβλεψης και η μείωση του σφάλματος, ελαχιστοποιώντας τη συνάρτηση κόστους. Η επίτευξη αυτού του στόχου πραγματοποιείται αναπροσαρμόζοντας τις παραμέτρους του δικτύου με τέτοιο τρόπο ώστε οι προβλέψεις του να προσεγγίζουν όσο το δυνατόν περισσότερο τις τιμές των δεδομένων. Χαρακτηριστικά παραδείγματα συναρτήσεων κόστους αποτελούν το *μέσο τετραγωνικό σφάλμα* (mean squared error), το *μέσο απόλυτο σφάλμα* (mean absolute error) και η *διασταυρούμενη εντροπία* (cross entropy).

Αλγόριθμος κατάβασης κλίσης

Ένας συνήθης επαναληπτικός αλγόριθμος βελτιστοποίησης που συναντάται συχνά στα ΒΝΔ είναι ο αλγόριθμος *κατάβασης κλίσης* (gradient descent). Ο αλγόριθμος αυτός χρησιμοποιείται για την εύρεση του τοπικού ελαχίστου μια συνάρτησης, μετακινώντας το αρχικό σημείο επαναληπτικά προς την αντίθετη κατεύθυνση της κλίσης. Στη μηχανική μάθηση, χρησιμοποιούμε τον αλγόριθμο κατάβασης κλίσης για να ενημερώσουμε τις παραμέτρους του μοντέλου μας, δηλαδή τα βάρη και τις πολώσεις όταν πρόκειται για νευρωνικό δίκτυο. Δεδομένης μιας συνάρτησης μιας μεταβλητής, θα αναλύσουμε στη συνέχεια τα βήματα του αλγορίθμου για την εύρεση ενός τοπικού ελαχίστου της.

Ξεκινώντας από ένα τυχαίο σημείο της συνάρτησης, υπολογίζουμε πρώτα την κλίση της συνάρτησης στο σημείο αυτό μέσα από την εύρεση της παραγώγου της. Ανάλογα με το πρόσημο της κλίσης και χρησιμοποιώντας το κατάλληλο βήμα επιλέγουμε ένα σημείο αριστερά του αρχικού, αν η κλίση είναι θετική, ή ένα σημείο δεξιά του αρχικού, αν η κλίση είναι αρνητική αντίστοιχα. Εάν η κλίση είναι μηδενική τότε η διαδικασία λήγει με επιτυχία καθώς το συγκεκριμένο σημείο αποτελεί τοπικό ελάχιστο της συνάρτησης μας. Σε αντίθετη περίπτωση, επιλέγοντας ένα κατάλληλο βήμα - δηλαδή μια κατάλληλη απόσταση μεταξύ των επιλεγμένων διαδοχικών σημείων - θα καταλήξουμε μετά από διαδοχικές επαναλήψεις της παραπάνω διαδικασίας σε τοπικό ελάχιστο. Εάν το πρόσημο της κλίσης σε ένα σημείο είναι αντίθετο από αυτό που είχε η κλίση στο προηγούμενο σημείο, αυτό σημαίνει ότι το βήμα που χρησιμοποιούμε είναι μεγάλο και για αυτό ξεπεράσαμε το επιθυμητό σημείο. Οπότε σε

αυτή την περίπτωση θα πρέπει να αναπροσαρμόσουμε το βήμα μικραίνοντάς το και να επιλέξουμε ένα σημείο προς τα πίσω. Βελτιώνοντας την παραπάνω σκέψη μας, μπορούμε να μικραίνουμε την τιμή του βήματος αναλογικά με την τιμή της κλίσης ώστε να μην ξεπερνάμε κατά πολύ το σημείο τοπικού ελαχίστου.



Σχήμα 2.6: Αναπαράσταση της πορείας του αλγορίθμου κατάβασης κλίσης για συνάρτηση δύο μεταβλητών (**Πηγή**: [onlic]).

Όπως φαίνεται στο Σχήμα 2.6, η διαδικασία αυτή μπορεί εύκολα να γενικευτεί και για συναρτήσεις περισσοτέρων μεταβλητών, με τη διαφορά ότι ο υπολογισμός της κλίσης περιλαμβάνει πλέον τον υπολογισμό του διανύσματος μερικών παραγώγων και η κατεύθυνση κατά την οποία πρέπει να εκτελέσουμε το επόμενο βήμα μας καθορίζεται από το αρνητικό διάνυσμα μερικών παραγώγων.

Προς τα πίσω διάδοση

Η μέθοδος της προς τα πίσω διάδοσης στοχεύει στην εκμάθηση των παραμέτρων του δικτύου (βάρη και πολώσεις) μέσω της διαδικασίας της βελτιστοποίησης της συνάρτησης κόστους. Η μέθοδος αυτή χρησιμοποιεί τον αλγόριθμο κατάβασης κλίσης για τον υπολογισμό της ελαχιστοποίησης της συνάρτησης κόστους πάνω στο σύνολο δεδομένων εκπαίδευσης, από τον οποίο θα προκύψουν τελικά οι κατάλληλες τιμές των παραμέτρων του δικτύου. Η τροποποίηση των βαρών και των πολώσεων ενός νευρωνικού δικτύου ξεκινά από το τελευταίο επίπεδο του δικτύου και ολοκληρώνεται στο επίπεδο εισόδου, διατρέχοντας όλα τα ενδιάμεσα κρυφά επίπεδα και επιλέγοντας σε κάθε επανάληψη ένα νέο γειτονικό ως προς το προηγούμενο επίπεδο. Η μεταβολή που θα υποστούν οι παράμετροι κάθε νευρώνα καθορίζεται από την αντίστοιχη τιμή του διανύσματος κλίσεων. Πιο συγκεκριμένα, η μεταβολή των βαρών ενός δικτύου δίνεται από την Εξίσωση 2.6:

$$W_i(t+1) = W_i(t) + \Delta W_i(t) \tag{2.6}$$

όπου $w_i(t+1)$ είναι η νέα τιμή του βάρους i, $w_i(t)$ η παλιά τιμή του βάρους i και ο όρος $w_i(t)$ εκφράζει το πόσο πρέπει να μεταβληθεί το βάρος ώστε να ελαχιστοποιηθεί το σφάλμα της συνάρτησης κόστους.

2.2.5 Βαθιά μηχανική μάθηση

Η *βαθιά μάθηση* (deep learning) και τα νευρωνικά δίκτυα τείνουν να χρησιμοποιούνται συχνά ως συνώνυμοι όροι, πράγμα το οποίο μπορεί να προκαλέσει σύγχυση στη βαθύτερη κατανόηση των δύο εννοιών. Σε αυτό το σημείο αξίζει να σημειωθεί ότι ο όρος «βαθιά» στη βαθιά μηχανική μάθηση αναφέρεται στο πλήθος των κρυφών επιπέδων ενός νευρωνικού δικτύου [onlie]. Ένα νευρωνικό δίκτυο που έχει δύο ή περισσότερα κρυφά επίπεδα μπορεί να θεωρηθεί βαθύ. Η ύπαρξη των κρυφών επιπέδων

συμβάλλει καθοριστικά στην αύξηση της αποδοτικότητας και της αποτελεσματικότητάς τους. Πρακτικά, η διαφορά μεταξύ των νευρωνικών δικτύων και της βαθιάς μηχανικής μάθησης έγκειται στο γεγονός ότι στην περίπτωση της βαθιάς μηχανικής μάθησης η διαδικασία εξαγωγής χαρακτηριστικών από τα δεδομένα εισόδου - σύμφωνα με τα οποία θα σχηματιστεί στη συνέχεια το μοντέλο - γίνεται αυτοματοποιημένα μέσα στο δίκτυο, ενώ στα απλά νευρωνικά δίκτυα η παρέμβαση του ανθρώπινου παράγοντα κρίνεται απαραίτητη.

Τα τελευταία χρόνια χρησιμοποιούνται όλο και περισσότερο βαθιά δίκτυα μηχανικής μάθησης ενώ τα «ρηχά» δίκτυα περιορίζονται στην επίλυση πιο απλών προβλημάτων. Η ραγδαία πρόοδος που έχει σημειωθεί στις κάρτες γραφικών τους δίνει τη δυνατότητα να επεξεργάζονται παράλληλα και σε πολύ υψηλές ταχύτητες μεγάλο όγκο δεδομένων. Έτσι, με τη χρήση καρτών γραφικών τελευταίας γενιάς η μεγάλη υπολογιστική πολυπλοκότητα που συνεπάγονται τα δίκτυα με χιλιάδες νευρώνες είναι πλέον διαχειρίσιμη. Επιπλέον, η πληθώρα δεδομένων που παρατηρείται στις μέρες μας καθώς και η εύκολη πρόσβαση σε μεγάλες βάσεις δεδομένων με συχνά κατηγοριοποιημένα σύνολα αποτελούν καθοριστικό παράγοντα στην ανάπτυξη και την ισχυροποίηση των δικτύων βαθιάς μηχανικής μάθησης.

Μια συνήθης πρακτική που ακολουθείται συχνά στο χώρο της μηχανικής μάθησης και στοχεύει κυρίως στη σημαντική μείωση του χρόνου εκπαίδευσης των μοντέλων είναι η χρήση προεκπαιδευμένων μοντέλων. Όπως έχει αποδειχθεί και στην πράξη, ένα μοντέλο που έχει εκπαιδευτεί επιτυχώς σε μεγάλα σύνολα δεδομένων και συνεπώς διαθέτει αυτή τη γνώση θα συγκλίνει πολύ πιο γρήγορα κατά τη διαδικασία της εκπαίδευσής του σε ένα εξειδικευμένο σύνολο δεδομένων από ότι ένα μοντέλο που έχει αρχικοποιηθεί με τυχαίες τιμές βαρών. Η τεχνική αυτή ανήκει στην ευρύτερη κατηγορία της μεταφοράς γνώσης μεταξύ των δικτύων, μέθοδος η οποία χρησιμοποιείται όλο και περισσότερο στη μηχανική μάθηση.

Συνελικτικά δίκτυα

Τα *συνελικτικά δίκτυα* (convolutional networks) αποτελούν μια υποκατηγορία βαθιών νευρωνικών δικτύων και χρησιμοποιούνται ευρέως στην αναγνώριση εικόνων και βίντεο. Σε αφαιρετικό επίπεδο, προσπαθούν να προσομοιάσουν τη βιολογική λειτουργία κατά την οποία μεμονωμένοι φλοιώδεις νευρώνες ανταποκρίνονται σε ερεθίσματα μόνο σε μια περιορισμένη περιοχή του οπτικού πεδίου που είναι γνωστή ως το δεκτικό πεδίο. Τα δεκτικά πεδία διαφορετικών νευρώνων επικαλύπτονται εν μέρει έτσι ώστε να καλύπτουν ολόκληρο το οπτικό πεδίο. Έτσι τα δίκτυα αυτά επιχειρούν να διαχωρίσουν τα χαρακτηριστικά της εικόνας ώστε η επιμέρους επεξεργασία τους από τους νευρώνες να οδηγήσει στην αναγνώριση της εικόνας εισόδου. Από την Εξίσωση 2.7 γίνεται φανερό ότι, στην πράξη, τα συνελικτικά επίπεδα εφαρμόζουν στην είσοδο την πράξη της συνέλιξης και προωθούν το αποτέλεσμα στο επόμενο επίπεδο. Σε αντίθεση με τα καθιερωμένα πλήρως συνδεδεμένα στρώματα, κάθε νευρώνας ενός συνελικτικού επιπέδου δεν συνδέεται με όλους τους νευρώνες του προηγούμενου επιπέδου, πράγμα το οποίο θα απαιτούσε την εκπαίδευση πολλαπλών παραμέτρων δικτύου, αφού το μέγεθος των δεδομένων εισόδου είναι εκτεταμένο εφόσον πρόκειται για εικόνες. Στα συνελικτικά επίπεδα, κάθε νευρώνας συνδέεται μόνο με ένα μέρος του συνόλου των νευρώνων του προηγούμενου επιπέδου και όλοι οι νευρώνες του ίδιου επιπέδου μοιράζονται τα ίδια βάρη. Επιπλέον, σημειώνεται ότι κάθε συνελικτικό επίπεδο είναι οργανωμένο σε τρεις διαστάσεις, το ύψος, το πλάτος και το βάθος· η τελευταία διάσταση καθορίζεται από των αριθμό των συνελικτικών φίλτρων που θα εφαρμοστούν. Όπως και στην παραδοσιακή επεξεργασία εικόνας με χρήση φίλτρων, έτσι και στα συνελικτικά δίκτυα τα φίλτρα μπορούν να αποθηκεύουν πληροφορία για κάποια χρήσιμη έννοια, ανάλογα με το εκάστοτε πρόβλημα.

$$\vec{y} = \vec{w} \cdot \vec{x} + \vec{b} \tag{2.7}$$

Κεφάλαιο 3

Επιθέσεις σε βαθιά νευρωνικά δίκτυα

3.1 Βασικοί παράγοντες επιθέσεων

Όπως αναφέρθηκε σε προηγούμενο κεφάλαιο, τα βαθιά νευρωνικά δίκτυα, παρ' ότι αποτελούν ισχυρό εργαλείο επίλυσης πολλών προβλημάτων και μοντέλα τους εφαρμόζονται σε πλήθος καθημερινών συσκευών που χρησιμοποιεί ο σύγχρονος άνθρωπος, είναι ευάλωτα σε επιθέσεις. Η αδυναμία αυτή των ΒΝΔ αποτελεί ένα από τα σημαντικότερα προβλήματα τα οποία η επιστημονική κοινότητα καλείται να αντιμετωπίσει. Η δημιουργία εύρωστων πολυεπίπεδων νευρωνικών δικτύων που θα είναι σε θέση αρχικά να αναγνωρίζουν πιθανές επιθέσεις και σε δεύτερη φάση να τις διαχειρίζονται αποτελεί φλέγον ζήτημα στον τομέα των ΒΝΔ. Μια τέτοια υλοποίηση που θα ξεπερνάει τον κίνδυνο των επιθέσεων μπορεί να επιφέρει ραγδαία πρόοδο ακόμα και σε εν εξελίξει προγράμματα που αντιμετωπίζουν το αδιέξοδο αυτό, όπως η κατασκευή αυτόνομων οχημάτων για ευρεία χρήση. Η σπουδαιότητα και η κρισιμότητα του δεδομένου προβλήματος είναι αδιαμφισβήτητη και συνεπώς η βαθύτερη κατανόηση και μελέτη των υπαρχόντων επιθέσεων κρίνεται απαραίτητη.

Μια επίθεση που δέχεται ένα νευρωνικό δίκτυο είναι πολυπαραγοντική και οφείλει να μελετηθεί υπό αυτό το πρίσμα. Ερωτήματα όπως που αποσκοπεί η επίθεση, ποιες πληροφορίες είναι διαθέσιμες στον εισβολέα σχετικά με τη δομή και την αρχιτεκτονική του μοντέλου και τι είδους μοντέλα βαθιάς μηχανικής μάθησης συνήθως γίνονται στόχοι επιθέσεων, αποδεικνύουν την πολύπλευρη φύση των επιθέσεων και θα απαντηθούν στη συνέχεια.

3.1.1 Στόχος αντιπάλου

Πολλές φορές, κύριος στόχος του εισβολέα είναι να παραπλανήσει τον ταξινομητή για κάποιο συγκεκριμένο δείγμα ή μια ορισμένη κατηγορία δειγμάτων ώστε ο ταξινομητής να αποφανθεί εσφαλμένα, ενώ άλλες φορές ο εισβολέας προσπαθεί να επηρεάσει συνολικά τη γενική απόδοση του ταξινομητή. Ανάλογα λοιπόν με τις προθέσεις και τον απώτερο σκοπό του επιτιθέμενου, οι κακόβουλες επιθέσεις μπορούν να διακριθούν σε δύο βασικές κατηγορίες: (i) στις *επιθέσεις δηλητηρίασης* (poisoning attacks) και (ii) στις *επιθέσεις διαφυγής* (evasion attacks).

Επιθέσεις δηλητηρίασης

Οι επιθέσεις δηλητηρίασης αναφέρονται στους αλγόριθμους επίθεσης που επιτρέπουν σε έναν εισβολέα να εισάγει πλαστά δείγματα στο σύνολο εκπαίδευσης ενός αλγορίθμου ΒΝΔ. Τα τεχνητά αυτά δείγματα μπορούν να προκαλέσουν αστοχίες στον εκπαιδευμένο ταξινομητή και να οδηγηθεί σε εσφαλμένες αποφάσεις. Συγκεκριμένα, μπορούν να οδηγήσουν σε χαμηλά ποσοστά ακρίβειας του μοντέλου [Bigg13b], ή σε λανθασμένη πρόβλεψη ορισμένων δεδομένων ελέγχου [Züg18]. Συχνά, σε τέτοιου είδους επιθέσεις υπάρχει μεγάλη πιθανότητα ο εισβολέας να έχει πρόσβαση στη διαδικασία εκπαίδευσης του μοντέλου ή στη βάση με τα δεδομένα που προορίζονται για την εκπαίδευση του δικτύου. Για παράδειγμα, τα αποθετήρια που υπάρχουν στο διαδίκτυο συχνά συλλέγουν δείγματα κακόβουλου λογισμικού για τη διαδικασία της εκπαίδευσης, γεγονός που παρέχει την ευκαιρία στους εν δυνάμει εισβολείς να δηλητηριάσουν τα δεδομένα. Στην πραγματικότητα, η διαδικασία εκμάθησης ενός μοντέλου με δεδομένα που περικλείουν θόρυβο δεν αποτελεί καινούργιο πρόβλημα για την

επιστημονική κοινότητα, καθώς η πρώτη τοποθέτηση γύρω από τον συγκεκριμένο προβληματισμό πραγματοποιήθηκε το 1993 [Kear93]. Ωστόσο, οι περιπτώσεις που είχαν μελετηθεί αφορούσαν μια μικρή ποσότητα θορύβου που είχε προστεθεί στα δεδομένα χωρίς απαραίτητα την επέμβαση κάποιου κακόβουλου αντιπάλου, σε αντίθεση με την επίθεση δηλητηρίασης που αφορά τη διαδικασία όπου κάποιος εισβολέας προσπαθεί σκόπιμα να εκμεταλλευτεί το μοντέλο.

Η ιστορία των επιθέσεων δηλητηρίασης στη μηχανική μάθηση ξεκινά το 2008 με τη δημοσίευση ενός επιστημονικού άρθρου που ανέπτυσσε ένα παράδειγμα επίθεσης σε φίλτρα ανεπιθύμητων μηνυμάτων [Nels08]. Οι συγγραφείς του [Bigg13b] ήταν οι πρώτοι που πρότειναν μια επίθεση δηλητηρίασης εναντίον *μηχανών διανυσμάτων υποστήριξης* (support vector machines - SVM) εισάγοντας κακόβουλα δεδομένα στη βάση των δεδομένων εκπαίδευσης του μοντέλου, προκειμένου να πετύχουν σημαντική μείωση της ακρίβειας του. Η μέθοδος αυτή στοχεύει στον υπολογισμό της κατάβασης κλίσης με βάση τα χαρακτηριστικά του SVM για τη δημιουργία ορισμένων σημείων που μπορούν να απορριφθούν, μεγιστοποιώντας την ακρίβεια του SVM. Το 2017, στην έρευνα [Yang17] οι συγγραφείς προτείνουν μια μέθοδο επίθεσης δηλητηρίασης κατά των νευρωνικών δικτύων. Η μέθοδος αυτή χρησιμοποιεί για τη δημιουργία δεδομένων ενός *αυτοκωδικοποιητή* (autoencoder) βασισμένο στα *παραγωγικά δίκτυα μάθησης με αντιπαλότητα* (Generative Adversarial Networks - GAN) [Good14]. Παρόμοιες συστηματικές επιθέσεις δηλητηρίασης έχουν προταθεί και στον τομέα της υγειονομικής περίθαλψης. Χαρακτηριστικά, στην έρευνα [Moza15] αναλύεται η επίθεση δηλητηρίασης σε ένα σύνολο δεδομένων υγειονομικής περίθαλψης, γεγονός το οποίο υποδηλώνει τη γενικότερη επέκταση των συγκεκριμένων επιθέσεων και στον κλάδο της ιατρικής.

Επιθέσεις διαφυγής

Οι επιθέσεις διαφυγής αποτελούν τις συνηθέστερες επιθέσεις στον τομέα της μηχανικής μάθησης, που πραγματοποιούνται εναντίον του μοντέλου κατά τη διαδικασία ελέγχου της αποτελεσματικότητας του. Στις επιθέσεις διαφυγής, οι ταξινομητές είναι σταθεροί και συνήθως έχουν καλή απόδοση σε *«καλοήθη»* (benign) δείγματα ελέγχου. Οι εισβολείς δεν έχουν τη δυνατότητα να αλλάξουν τον ταξινομητή ή τις παραμέτρους του, αλλά μπορούν να κατασκευάσουν ψεύτικα δείγματα τα οποία ο ταξινομητής δεν μπορεί να αναγνωρίσει. Πιο συγκεκριμένα, οι αντίπαλοι δημιουργούν μερικά πλαστά δείγματα για να αποφύγουν τον εντοπισμό από τον ταξινομητή. Μια επίθεση πραγματοποιείται όταν ο επιτιθέμενος προσθέτει θόρυβο σε μια κατά τ' άλλα κανονική είσοδο, έτσι ώστε να οδηγήσει τον ταξινομητή στην πρόβλεψη λανθασμένου αποτελέσματος για τη συγκεκριμένη είσοδο. Η προσθήκη θορύβου πραγματοποιείται από τον εισβολέα με τέτοιο τρόπο έτσι ώστε να μην είναι αντιληπτή από τον άνθρωπο και η τροποποιημένη αυτή είσοδος ονομάζεται *αντιφατικό παράδειγμα* (adversarial example). Η επίθεση διαφυγής πραγματοποιείται λοιπόν όταν στο μοντέλο εισάγονται αντιφατικά παραδείγματα. Για παράδειγμα, στην περίπτωση των αυτόνομων οχημάτων οδήγησης, το να κολλήσει κάποιος μερικά κομμάτια ταινιών στις πινακίδες σημάτων οδικής κυκλοφορίας μπορεί να προκαλέσει σύγχυση στο μηχανισμό αναγνώρισης της οδικής σήμανσης που διαθέτει το όχημα [Eykh18].

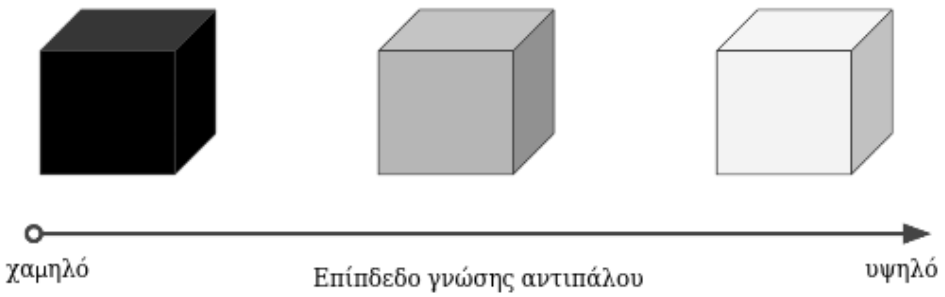
Υπάρχουν τέσσερις διαφορετικοί τρόποι με τους οποίους ο αντίπαλος μπορεί να παραποιήσει τα δείγματα εισόδου προκειμένου να αποφύγει την ανίχνευση τους από τον ταξινομητή, οι οποίοι αναλύονται στη συνέχεια [Ayub20]. Ένας τρόπος με τον οποίο ο εισβολέας μπορεί να οδηγήσει το μοντέλο σε εσφαλμένη ταξινόμηση είναι μειώνοντας τον *βαθμό εμπιστοσύνης* (confidence reduction) που επιδεικνύει το δίκτυο στην πρόβλεψη του σωστού αποτελέσματος για μια δεδομένη είσοδο. Ένας δεύτερος τρόπος με τον οποίο ο αντίπαλος επιδιώκει να αποπροσανατολίσει τον ταξινομητή είναι προσπαθώντας να *αλλάξει το αποτέλεσμα της κατηγοριοποίησης* (misclassification), θεωρώντας ότι η είσοδος ανήκει σε κλάση διαφορετική από την αντικειμενική της. Εναλλακτικά, ο αντίπαλος προσπαθεί να παράξει ένα δείγμα που ξεγελά το μοντέλο για να το ταξινομήσει σε μια συγκεκριμένη επιθυμητή κλάση, που φυσικά διαφέρει από την πραγματική. Τέλος, ο εισβολέας μπορεί να επιδιώκει κάθε είσοδος μιας κλάσης να κατηγοριοποιείται εσφαλμένα από τον ταξινομητή σε *κάποια άλλη* στοχευμένη κλάση (source/target misclassification).

Στοχευμένες και μη επιθέσεις

Συγχρόνως, με βάση τους στόχους και τις προθέσεις του επιτιθέμενου, μπορεί να υπάρξει και μια δεύτερη, άλλου τύπου διάκριση ανάμεσα στις κακόβουλες επιθέσεις, η οποία περιλαμβάνει τις στοχευμένες και τις μη στοχευμένες επιθέσεις. Στην στοχευμένη επίθεση, για κάθε δείγμα (x, y) - όπου το x είναι διάνυσμα χαρακτηριστικών και το $y \in Y$ είναι η πραγματική ετικέτα-κλάση του x - το οποίο αποτελεί εν δυνάμει στόχο του εισβολέα, ο αντίπαλος επιδιώκει να επηρεάσει τον ταξινομητή δίνοντας μια συγκεκριμένη ετικέτα $t \in Y$ στο τροποποιημένο δείγμα x'. Για παράδειγμα, ένας απατεώνας είναι πιθανό να επιτεθεί στο μοντέλο αξιολόγησης της πιστοληπτικής ικανότητας μιας χρηματοοικονομικής εταιρείας προκειμένου να αξιολογηθεί από το μοντέλο ως ένας εξαιρετικά αξιόπιστος και φερέγγυος πελάτης αυτής της εταιρείας. Εάν δεν υπάρχει καθορισμένη ετικέτα στόχος t για το δείγμα x το οποίο πρόκειται να τροποποιήσει ο εισβολέας, η επίθεση ονομάζεται μη στοχευμένη. Ο αντίπαλος στοχεύει απλά στο να καταφέρει να κάνει τον ταξινομητή να προβλέψει εσφαλμένα το δείγμα x', χωρίς να τον ενδιαφέρει να προκαθορίσει την ετικέτα t.

3.1.2 Γνώση αντιπάλου

Η αυστηρά περιορισμένη ή η πλήρης γνώση που μπορεί να διαθέτει ο εισβολέας σχετικά με την αρχιτεκτονική του μοντέλου, τις τιμές των παραμέτρων και τη γενικότερη δομή του δικτύου επηρεάζουν καθοριστικά την επιλογή και την εκτέλεση μιας επίθεσης. Αν, για παράδειγμα, ο αντίπαλος γνωρίζει τη δομή του ταξινομητή, τις παραμέτρους του ή το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση του, τότε έχει μεγαλύτερη ευελιξία κινήσεων και μπορεί να εκτελέσει με σχετική ευκολία μια αποτελεσματική επίθεση, σε αντίθεση με την περίπτωση όπου καμία από αυτές τις πληροφορίες δεν είναι διαθέσιμες. Σε συνθήκες όπου ο αντίπαλος δεν γνωρίζει καμία πληροφορία σχετικά με τη δομή του μοντέλου, εξακολουθεί να είναι εφικτή μια επιτυχημένη επίθεση από πλευράς του, αλλά συνήθως απαιτούνται περισσότερος χρόνος και υπολογιστικοί πόροι για την υλοποίηση της επίθεσης. Συνεπώς, όσο περισσότεροι φραγμοί τεθούν από την πλευρά του αμυνόμενου, τόσο περισσότερο θα αυξηθεί η πολυπλοκότητα στη διαδικασία πραγματοποίησης μιας επίθεσης, χωρίς ωστόσο να αίρεται ο κίνδυνος μιας κακόβουλης επίθεσης με ό,τι συνέπειες αυτό συνεπάγεται. Με βάση λοιπόν τις γνώσεις που διαθέτει ή όχι ο αντίπαλος, οι επιθέσεις χωρίζονται σε τρεις βασικές κατηγορίες, όπως φαίνεται και στο Σχήμα 3.1: στις (i) white box, στις (ii) black box και στις (iii) gray box επιθέσεις.



Σχήμα 3.1: Αναπαράσταση του επιπέδου των γνώσεων που διαθέτει ένας αντίπαλος σε κάθε μία από τις κατηγορίες επιθέσεων από αριστερά προς τα δεξιά: (i) white box, στις (ii) black box και στις (iii) gray box επιθέσεις. (**Πηγή**: [onli18])

Επίθεση white box

Σε μια επίθεση white box, ο αντίπαλος έχει πρόσβαση σε όλες τις πληροφορίες του νευρωνικού δικτύου-στόχου, συμπεριλαμβανομένης της αρχιτεκτονικής του, των παραμέτρων και των βαρών. Ο αντίπαλος μπορεί να κάνει πλήρη χρήση των πληροφοριών του δικτύου για να επεξεργαστεί και να κατασκευάσει προσεκτικά αντιφατικά δείγματα. Οι επιθέσεις αυτής της κατηγορίας έχουν μελετηθεί

εκτενώς, παρέχοντας μας πλούσια βιβλιογραφία, αφού η παράθεση της αρχιτεκτονικής του μοντέλου και των παραμέτρων του συνέβαλε καθοριστικά στο να γίνει πλήρως αντιληπτή και κατανοητή από την επιστημονική κοινότητα η αδυναμία των μοντέλων ΒΝΔ, καθώς και να μπορεί να αναλυθεί με μαθηματικά μοντέλα και εργαλεία. Η ασφάλεια των δικτύων εναντίον των επιθέσεων white box είναι η ιδανική που επιθυμούμε να έχουν τα όλα τα μοντέλα μηχανικής μάθησης [Tram20], καθώς αποτελεί τη μεγαλύτερη πρόκληση σε θέματα προστασίας των βαθιών νευρωνικών δικτύων.

Επίθεση black box

Σε μια επίθεση black box, η εσωτερική δομή και η διαμόρφωση του μοντέλου ενός ΒΝΔ υποψήφιου στόχου δεν είναι διαθέσιμη σε κανέναν εξωτερικό παράγοντα. Οι αντίπαλοι δεν διαθέτουν καμία γνώση σχετική με το μοντέλο-στόχο τους· μπορούν να τροφοδοτήσουν μόνο τα δεδομένα εισόδου και να υποβάλουν ερώτημα για τις εξόδους του δικτύου. Σε αυτή την κατηγορία επιθέσεων, οι εισβολείς συνήθως προτού επιτεθούν στα μοντέλα χρειάζεται να εκτελέσουν μια προεργασία προκειμένου να ανακαλύψουν την αδυναμία του εκάστοτε μοντέλου. Η διαδικασία ξεκινά με την εισαγωγή επεξεργασμένων δειγμάτων στο μοντέλο και συνεχίζεται μέσω παρατήρησης της εξόδου. Οι εισβολείς προσπαθούν να εκμεταλλευτούν τη σχέση εισόδου-εξόδου του μοντέλου εισάγοντας διαρκώς αντιφατικά δείγματα μέχρις ότου ανακαλύψουν ένα μοτίβο σύμφωνα με το οποίο μπορούν να παραπλανήσουν τον ταξινομητή του μοντέλου. Σε σύγκριση με τις επιθέσεις white box, οι επιθέσεις black box είναι πιο πρακτικές και για αυτό απαντώνται και περισσότερο σε εφαρμογές, καθώς οι σχεδιαστές των μοντέλων συνήθως δεν δημοσιοποιούν τις παραμέτρους και άλλες λεπτομέρειες του μοντέλου τους, κυρίως για λόγους πνευματικής ιδιοκτησίας.

Επίθεση gray (semi-white) box

Η επίθεση gray box, ή αλλιώς semi-white box, αποτελεί μια υβριδική μέθοδο επίθεσης, καθώς συνδυάζει τεχνικές που εφαρμόζονται και στις δύο προηγούμενες κατηγορίες. Οι αντίπαλοι γνωρίζουν την αρχιτεκτονική του ταξινομητή του ΒΝΔ, συχνά και άλλες λεπτομέρειες, όπως τον τύπο και τον αριθμό των επιπέδων του δικτύου ή το σύνολο δεδομένων που χρησιμοποιήθηκαν κατά τη διαδικασία της εκπαίδευσης, αλλά δεν γνωρίζουν τις παραμέτρους του. Οι επιτιθέμενοι εκπαιδεύουν ένα μοντέλο για την κατασκευή αντιφατικών δειγμάτων, χρησιμοποιώντας την αρχιτεκτονική του αρχικού μοντέλου, όπως θα έκαναν αντίστοιχα και σε μια επίθεση white box. Μόλις εκπαιδευτεί το νέο μοντέλο, ο εισβολέας δεν χρειάζεται πια το αρχικό μοντέλο, αφού είναι σε θέση πλέον να δημιουργεί αντιφατικά δείγματα και να τα διαχειρίζεται όπως σε μια επίθεση black box.

3.1.3 Μοντέλο θύμα

Πέρα από τα κίνητρα του εισβολέα και τις γνώσεις που διαθέτει γύρω από το μοντέλο-στόχο είναι σημαντικό να διερευνήσουμε και το ρόλο των ίδιων των μοντέλων που πέφτουν θύματα επιθέσεων, έτσι ώστε να αποκτήσουμε μια ολοκληρωμένη εικόνα για τις επιθέσεις στα ΒΝΔ. Σε αυτή την ενότητα θα συνοψίσουμε εν συντομία τα μοντέλα μηχανικής μάθησης που παρουσιάζουν μια σχετική ευαισθησία σε αντιφατικά δείγματα και είναι επιρρεπή στο να δέχονται επιθέσεις, μελετώντας μερικές δημοφιλείς αρχιτεκτονικές βαθιάς μάθησης που χρησιμοποιούνται σε τομείς δεδομένων εικόνας, γράφων και κειμένου.

Συμβατικά μοντέλα μηχανικής μάθησης

Για τα συμβατικά εργαλεία της μηχανικής μάθησης, υπάρχει μακρά ιστορία μελέτης ζητημάτων ασφάλειας και προστασίας. Πιο συγκεκριμένα, στην εργασία [Bigg13a] οι συγγραφείς μελετούν την επίθεση σε SVM καθώς και σε πλήρως συνδεδεμένα νευρωνικά δίκτυα, με μικρό βάθος χρησιμοποιώντας το σύνολο δεδομένων MNIST [LeCu]. Παράλληλα, σε μια άλλη έρευνα [Barr10] εξετάζεται η ασφάλεια του SpamBayes, ενός λογισμικού ανίχνευσης ανεπιθύμητων μηνυμάτων βασισμένο στη

μπεϋζιανή συμπερασματολογία. Επιπλέον, στην εργασία [Dalv04] ελέγχεται η ασφάλεια των αφελών μπεϋζιανών ταξινομητών. Πολλές από αυτές τις ιδέες και τις στρατηγικές έχουν υιοθετηθεί στη μελέτη επιθέσεων στα βαθιά νευρωνικά δίκτυα.

Μοντέλα βαθιάς μηχανικής μάθησης

Σε αντίθεση με τις παραδοσιακές τεχνικές μηχανικής μάθησης οι οποίες απαιτούν την ανθρώπινη παρέμβαση για την εξαγωγή των χαρακτηριστικών καθώς και γνώσεις σε βάθος πάνω στο συγκεκριμένο τομέα, τα ΒΝΔ είναι αλγόριθμοι μάθησης *«από άκρο σε άκρο»* (end-to-end). Τα ΒΝΔ χρησιμοποιούν τα δεδομένα απευθείας ως είσοδο στο μοντέλο, χωρίς να έχουν εκτελέσει κάποιου είδους προεργασία πριν, και μαθαίνουν τις δομές και τα χαρακτηριστικά των αντικειμένων που επεξεργάζονται αυτόματα κατά τη διαδικασία της εκπαίδευσης. Ωστόσο, το σημαντικό αυτό πλεονέκτημα που αποκτούν τα ΒΝΔ μέσω της «από άκρο σε άκρο» αρχιτεκτονικής τους αποτελεί σοβαρό μειονέκτημα σε ότι αφορά την ασφάλεια τους. Η «από άκρο σε άκρο» αρχιτεκτονική των ΒΝΔ διευκολύνει τους αντιπάλους να εκμεταλλευτούν τις πιθανές αδυναμίες τους και να δημιουργήσουν αντιφατικά δείγματα. Επιπλέον, λόγω των απεριόριστων δυνατοτήτων που διαθέτουν τα ΒΝΔ, ορισμένες από τις ιδιότητές τους δεν έχουν γίνει ακόμα πλήρως κατανοητές από την επιστημονική κοινότητα και ως λογικό επακόλουθο δεν υπάρχουν καλώς ορισμένες ερμηνείες τους. Επομένως, κρίνεται απαραίτητη η μελέτη των θεμάτων ασφαλείας των μοντέλων των ΒΝΔ χωριστά από τα υπόλοιπα μοντέλα μηχανικής μάθησης. Στη συνέχεια, θα παρουσιάσουμε εν συντομία μερικά δημοφιλή μοντέλα-θύματα βαθιάς μηχανικής μάθησης τα οποία χρησιμοποιούνται ως μοντέλα αναφοράς σε μελέτες που αφορούν επιθέσεις και άμυνες σε ΒΝΔ.

Πλήρως συνδεδεμένα νευρωνικά δίκτυα

Τα *πλήρως συνδεδεμένα νευρωνικά δίκτυα* (fully-connected neural networks) αποτελούνται από επίπεδα τεχνητών νευρώνων. Σε κάθε επίπεδο, οι νευρώνες λαμβάνουν την είσοδο από προηγούμενα επίπεδα, την επεξεργάζονται χρησιμοποιώντας μία συνάρτηση ενεργοποίησης και την στέλνουν στο επόμενο επίπεδο. Η είσοδος του πρώτου επιπέδου είναι το δείγμα x και η έξοδος του τελευταίου επιπέδου είναι το αποτέλεσμα της $F(x)$. Ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο m-επιπέδων μπορεί να σχηματιστεί με βάση την Εξίσωση 3.1.

$$z^{(0)} = x; \;\; z^{(l+1)} = \sigma\left(W^l z^l \; + \; b^l\right) \tag{3.1}$$

όπου σ η συνάρτηση ενεργοποίησης, W το διάνυσμα βαρών και b η πόλωση.

Συνελικτικά νευρωνικά δίκτυα

Στον τομέα της όρασης υπολογιστών, τα συνελικτικά νευρωνικά δίκτυα είναι από τα πιο ευρέως χρησιμοποιούμενα μοντέλα [Kriz12b]. Τα CNN συγκεντρώνουν τα τοπικά χαρακτηριστικά από την εκάστοτε εικόνα για να μάθουν τις αναπαραστάσεις των αντικειμένων της. Τα μοντέλα αυτά μπορούν να θεωρηθούν ως μια λιγότερο «πυκνή» έκδοση των πλήρως συνδεδεμένων νευρωνικών δικτύων καθώς τα περισσότερα βάρη μεταξύ των επιπέδων είναι μηδενικά. Επίσης, ο αλγόριθμος εκπαίδευσης των CNN και ο υπολογισμός της κλίσης τους συνήθως είναι κοινός με τον αντίστοιχο των πλήρως συνδεδεμένων νευρωνικών δικτύων.

Συνελικτικά δίκτυα γράφων

Η εργασία [Kipf17] εισάγει τα συνελικτικά δίκτυα γράφων, τα οποία αργότερα έγιναν δημοφιλή μοντέλα ταξινόμησης κόμβων για δεδομένα γράφων. Η ιδέα αυτών των δικτύων είναι παρόμοια με αυτή των CNN· το μοντέλο συγκεντρώνει τις πληροφορίες από τους γειτονικούς κόμβους για να μάθει τις αναπαραστάσεις για κάθε κόμβο v και εξάγει το αποτέλεσμα της $F(v, X)$ ως πρόβλεψη. Η μαθηματική σχέση της παραπάνω διατύπωσης δίνεται από την Εξίσωση 3.2.

$$H^{(0)} = X; \quad H^{(l+1)} = \sigma\left(\hat{A}H^{(l)}W^{l}\right) \tag{3.2}$$

όπου X ο πίνακας χαρακτηριστικών του γράφου εισόδου, W το διάνυσμα βαρών, σ η συνάρτηση ενεργοποίησης και ο $\hat{A}$ εξαρτάται από τον βαθμό του πίνακα του γράφου και τον πίνακα γειτνίασης.

Επαναληπτικά νευρωνικά δίκτυα

Τα *επαναληπτικά νευρωνικά δίκτυα* (recurrent neural networks - RNN) είναι πολύ χρήσιμα για την αντιμετώπιση διαδοχικών δεδομένων εισόδου και για αυτόν το λόγο χρησιμοποιούνται ευρέως στην επεξεργασία της φυσικής γλώσσας. Τα μοντέλα των RNN, ειδικότερα τα μοντέλα *μακράς βραχυπρόθεσμης μνήμης* [Hoch97], είναι σε θέση να αποθηκεύουν τις χρονικά προγενέστερες πληροφορίες στη μνήμη και να εκμεταλλεύονται χρήσιμες πληροφορίες από προηγούμενη ακολουθία δεδομένων για την πρόβλεψη του επόμενου βήματος.

3.2 Δημιουργία αντιφατικών παραδειγμάτων

Σε αυτήν την ενότητα, θα παρουσιάσουμε τις κύριες μεθόδους που χρησιμοποιούνται στον τομέα ταξινόμησης εικόνων για τη δημιουργία αντιφατικών παραδειγμάτων. Η μελέτη τέτοιων δειγμάτων στον τομέα της εικόνας θεωρείται απαραίτητη διότι: (i) η ομοιότητα μεταξύ ψεύτικων και κανονικών εικόνων γίνεται αντιληπτή από τους παρατηρητές διαισθητικά, μέσω της ανθρώπινης αίσθησης της όρασης και (ii) τα δεδομένα και οι ταξινομητές των εικόνων έχουν απλούστερη δομή από τα αντίστοιχα σε άλλους τομείς, όπως για παράδειγμα στον ήχο. Έτσι, πολλές μελέτες επικεντρώνονται εξ' ολοκλήρου στις επιθέσεις εναντίον ταξινομητών στον τομέα της εικόνας ως τυπική περίπτωση επιθέσεων. Σε όλες τις περιπτώσεις που θα αναλυθούν στην πορεία, υποθέτουμε ότι οι ταξινομητές εικόνας αναφέρονται σε πλήρως συνδεδεμένα νευρωνικά δίκτυα και σε συνελικτικά νευρωνικά δίκτυα, καθώς αυτά αποτελούν και το αντικείμενο μελέτης της διατριβής. Τα πιο συνηθισμένα σύνολα δεδομένων που χρησιμοποιούνται στις παρακάτω μελέτες περιλαμβάνουν το σύνολο χειρόγραφων εικόνων MNIST [LeCu] και τα σύνολα εικόνων CIFAR10 [onlib] και ImageNet [onlid]. Στη συνέχεια, θα εξετάσουμε ορισμένες βασικές μεθόδους που χρησιμοποιούνται για τη δημιουργία αντιφατικών παραδειγμάτων εικόνων, καλύπτοντας τις επιθέσεις δηλητηρίασης, τις τρεις κατηγορίες επιθέσεων διαφυγής που αναλύσαμε σε προηγούμενη ενότητα (white box, black box και gray box) καθώς και μια τέταρτη κατηγορία, τις επιθέσεις στο φυσικό κόσμο, που ανήκει και αυτή στον ευρύτερο τομέα των επιθέσεων διαφυγής.

3.2.1 Επιθέσεις white box

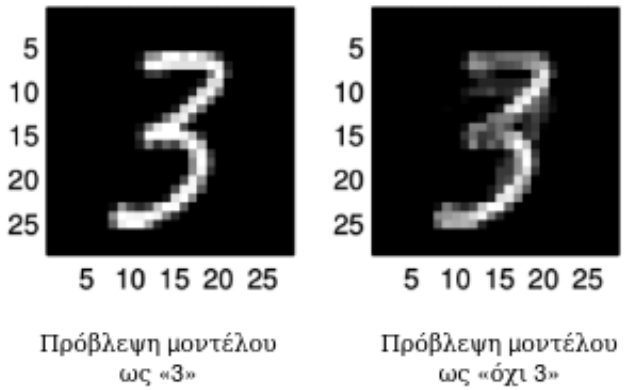
Γενικά, σε μια επίθεση white box, όταν ο ταξινομητής C ενός μοντέλου-στόχου F και το δείγμα προς τροποποίηση (x, y) δοθούν στον εισβολέα, ο στόχος του είναι να συνθέσει μια ψεύτικη εικόνα x' παρόμοια - υπό το πρίσμα της ανθρώπινης αντίληψης - με την αρχική εικόνα x που να μπορεί όμως να παραπλανήσει τον ταξινομητή C για να δώσει λάθος αποτελέσματα πρόβλεψης. Η παραπάνω σκέψη μπορεί να διατυπωθεί μαθηματικά μέσω της Εξίσωσης 3.3.

$$\begin{aligned} &\exists x' \ \Rightarrow ||x' - x|| \ \leq \ \varepsilon \\ &s.t. \ C(x') = t' \neq y \end{aligned} \tag{3.3}$$

όπου το $||\cdot||$ μετρά την ανομοιότητα μεταξύ x' και x και συνήθως είναι κάποια l_p νόρμα. Στη συνέχεια, θα αναλύσουμε μερικές από τις κύριες μεθόδους αυτής της κατηγορίας για να συνειδητοποιήσουμε καλύτερα την παραπάνω μαθηματική διατύπωση.

Επίθεση Biggio

Στην εργασία [Bigg13a], οι συγγραφείς δημιουργούν αντιφατικά δείγματα εισόδου χρησιμοποιώντας το σύνολο δεδομένων MNIST [LeCu] με στόχο συμβατικούς ταξινομητές μηχανικής μάθησης, όπως SVM και πλήρως συνδεδεμένα νευρωνικά δίκτυα 3-επιπέδων . Η επίθεση, όπως περιγράφεται, εστιάζει στη βελτιστοποίηση της συνάρτησης διαχωρισμού των δειγμάτων στην εκάστοτε κατηγορία. Για την καλύτερη κατανόηση του όρου «συνάρτηση διαχωρισμού» αλλά και της γενικότερης λειτουργίας του μοντέλου-θύματος της συγκεκριμένης επίθεσης θα δώσουμε ένα πρακτικό παράδειγμα για την περίπτωση ενός γραμμικού ταξινομητή SVM. Η συνάρτηση διαχωρισμού σε μια τέτοια περίπτωση θα είναι της μορφής $g(x) = < w, x > + b$. Έστω ότι θέλουμε να εξετάσουμε αν ένα δείγμα x από το σύνολο δεδομένων του MNIST ανήκει στην κατηγορία «3» ή όχι, τότε ανάλογα με το πρόσημο της συνάρτησης διαχωρισμού, ο γραμμικός ταξινομητής SVM θα μπορέσει να αποφανθεί καταλλήλως· στην περίπτωση που η συνάρτηση g έχει θετική τιμή αυτό συνεπάγεται ότι το δείγμα x ανήκει στην κλάση «3», ενώ στην περίπτωση που η g έχει αρνητική τιμή ή μηδέν αυτό δηλώνει ότι το δείγμα x ανήκει στην κλάση «όχι 3».



Σχήμα 3.2: Επίθεση Biggio σε ταξινομητή SVM στον τομέα της αναγνώρισης ψηφίων (**Πηγή**:[Bigg13a]).

Υποθέτοντας τώρα ότι ένα δείγμα x έχει κατηγοριοποιηθεί σωστά από τον ταξινομητή στην κλάση «3» θα μελετήσουμε τη διαδικασία διεξαγωγής της επίθεσης. Σύμφωνα λοιπόν με την επίθεση Biggio, θα πρέπει να δημιουργήσουμε ένα νέο δείγμα x' για να ελαχιστοποιήσουμε την τιμή της συνάρτησης διαχωρισμού $g(x')$, ενώ συγχρόνως θα πρέπει να κρατήσουμε τη διαφορά $||x'-x||_1$ όσο μικρότερη γίνεται. Εάν η $g(x')$ είναι αρνητική τότε το δείγμα x' ταξινομείται στην κλάση «όχι 3», αλλά το x' είναι ακόμα κοντά στο x οπότε τελικά ο ταξινομητής παραπλανείται και η επίθεση είναι επιτυχημένη. Ένα οπτικό παράδειγμα της επίθεσης που μόλις περιγράψαμε δίνεται στο Σχήμα 3.2. Οι μελέτες σχετικά με αντιφατικά παραδείγματα για συμβατικά μοντέλα μηχανικής μάθησης [Bigg13b, Bigg13a, Dalv04] έχουν αποτελέσει πηγή έμπνευσης και κίνητρο για την πραγματοποίηση ερευνών πάνω σε θέματα ασφάλειας των μοντέλων βαθιάς μηχανικής μάθησης.

Επίθεση Szegedy L-BFGS

Η εργασία [Szeg14] συμπεριλαμβάνεται μεταξύ των πρώτων που υλοποιούν επίθεση σε ταξινομητές εικόνας βαθιών νευρωνικών δικτύων. Σύμφωνα με αυτή, το πρόβλημα βελτιστοποίησης διατυπώνεται ως μια αναζήτηση για την ελαχιστοποίηση του αλλοιωμένου δείγματος x' σύμφωνα με την Εξίσωση 3.4

$$\begin{aligned} &\min ||x - x'||_2^2 \\ &s.t. C(x') = t \\ &\quad x' \in [0,\ 1]^m \end{aligned} \tag{3.4}$$

Το πρόβλημα επιλύεται σχεδόν εξ' ολοκλήρου με την εισαγωγή της συνάρτησης σφάλματος όπως επισημαίνεται στην Εξίσωση 3.5.

$$\begin{aligned} &\min c||x - x'||_2^2 \ + \ \mathcal{L}(\theta,\ x',\ t) \\ &s.t. x' \in [0,\ 1]^m \end{aligned} \tag{3.5}$$

Στην παραπάνω σχέση, ο πρώτος όρος επιβάλλει την ομοιότητα μεταξύ x' και x. Ο δεύτερος όρος ενθαρρύνει τον αλγόριθμο να βρει x', το οποίο έχει μικρή τιμή σφάλματος για την κλάση t, οπότε ο ταξινομητής C είναι πολύ πιθανό να προβλέψει ότι το x' ανήκει στην t. Αλλάζοντας συνεχώς την τιμή της σταθεράς c, μπορούμε να βρούμε ένα x' που να έχει ελάχιστη απόσταση από το x, και συγχρόνως να ξεγελά τον ταξινομητή C. Για την επίλυση του τελευταίου προβλήματος, οι συγγραφείς χρησιμοποίησαν τον αλγόριθμο L-BFGS [Liu89].

Επίθεση με Fast Gradient Sign Method

Στην εργασία [Good15] παρουσιάζεται μια μέθοδος μέσω της οποίας δημιουργούνται γρήγορα τα αντιφατικά δείγματα μέσα σε ένα βήμα. Η μέθοδος αυτή δίνεται αναλυτικά από τη μαθηματική Εξίσωση 3.6. Στην ουσία, μέσω της fast gradient sign method (FGSM), ο εισβολέας προσπαθεί να προσθέσει θόρυβο στο δείγμα x, του οποίου η κατεύθυνση είναι ίδια με την κατεύθυνση της κλίσης της συνάρτησης κόστους. Στη συνέχεια ο θόρυβος κανονικοποιείται από το ϵ, το οποίο εξ' ορισμού είναι ένας πολύ μικρός αριθμός. Τονίζεται ότι, η τιμή της κλίσης της συνάρτησης κόστους δεν έχει καμία σημασία στημ Εξίσωση 3.6, παρά μόνο το πρόσημό της, που ορίζει την κατεύθυνση της.

$$\begin{aligned} x' &= x \ + \ \epsilon sign(\nabla_x \mathcal{L}(\theta,\ x,\ y)) \quad \text{non target} \\ x' &= x - \epsilon sign(\nabla_x \mathcal{L}(\theta,\ x,\ t)) \quad \text{target on } t \end{aligned} \tag{3.6}$$

Στην περίπτωση της στοχευμένης επίθεσης, η μέθοδος αυτή μπορεί να θεωρηθεί ως ένα βήμα του αλγορίθμου κατάβασης κλίσης που επιλύει το πρόβλημα που διατυπώνεται στη Σχέση 3.7:

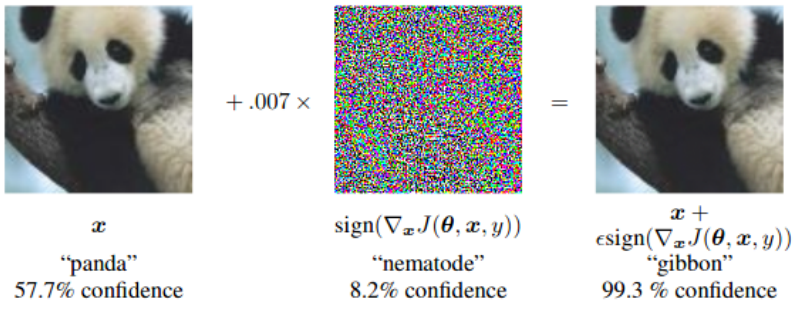
$$\begin{aligned} &\min \mathcal{L}(\theta,\ x',\ t) \\ &s.t.\ ||x' - x||_\infty \ \leq \epsilon \\ &x' \ \in \ [0,\ 1]^m \end{aligned} \tag{3.7}$$

Η συνάρτηση κόστους που περιγράφεται στην Εξίσωση 3.7 ψάχνει το σημείο εκείνο που παρουσιάζει την ελάχιστη τιμή σφάλματος για την κλάση t στα πλαίσια της γειτονιάς του x - δηλαδή στα πλαίσια μιας σφαίρας ακτίνας ϵ και κέντρου x - για το οποίο το μοντέλο F έχει τη μεγαλύτερη πιθανότητα να το κατηγοριοποιήσει στην κλάση t. Με αυτόν τον τρόπο, το κατασκευασμένο δείγμα x', που δημιουργήθηκε με ένα μόνο βήμα, είναι επίσης πιθανό να παραπλανήσει το μοντέλο.

Η επίθεση FGSM είναι απλή, υπολογιστικά αποτελεσματική συγκριτικά με άλλες μεθόδους και δύναται να δημιουργεί γρήγορα αντιφατικά παραδείγματα, αφού περιλαμβάνει μόνο τον υπολογισμό ενός βήματος προς τα πίσω διάδοσης. Το σημαντικό αυτό πλεονέκτημα την καθιστά την πλέον κατάλληλη σε περιπτώσεις όπου απαιτείται η δημιουργία αριθμητικά πολλών αντιφατικών δειγμάτων. Χαρακτηριστικό παράδειγμα μιας τέτοιας περίπτωσης αποτελεί η τεχνική της εκπαίδευσης με αντιπαλότητα, όπου χρησιμοποιείται η FGSM για την παραγωγή αντιφατικών παραδειγμάτων για κάθε ένα από τα δείγματα εισόδου του συνόλου δεδομένων εκπαίδευσης [Kura17b]. Στο Σχήμα 3.3 παρουσιάζεται ένα παράδειγμα εφαρμογής της επίθεσης FGSM στο σύνολο δεδομένων ImageNet.

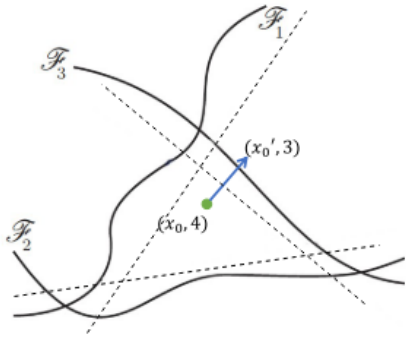
Επίθεση DeepFool

Στην εργασία [Moos16] οι συγγραφείς μελετούν έναν ταξινομητή με F_k συναρτήσεις k ορίων απόφασης γύρω από το σημείο x, όπου x είναι ένα δείγμα που έχει επιλεχθεί από τα δεδομένα εισόδου του μοντέλου. Προσπαθούν να βρουν ένα τρόπο έτσι ώστε το x να μπορεί να ξεπεράσει το όριο



Σχήμα 3.3: Εφαρμόζοντας την FGSM επίθεση ο ταξινομητής πλέον κατηγοριοποιεί το «πάντα» ως «γίββωνα» (**Πηγή**: [Good15]).

απόφασης, όπως φαίνεται στο Σχήμα 3.4, έτσι ώστε ο ταξινομητής να δώσει μια διαφορετική πρόβλεψη για το x από την πραγματική. Για παράδειγμα, έστω ότι επιλέγουμε ένα δείγμα-στόχο x_0 από το σύνολο δεδομένων του MNIST το οποίο ανήκει στην κλάση «4» και ο ταξινομητής του μοντέλου το κατηγοριοποιεί σωστά στην ίδια κλάση. Εάν υποθέσουμε ότι ο στόχος μας είναι να παραπλανήσουμε τον ταξινομητή ώστε να κατηγοριοποιήσει το x_0 στην κλάση t με $t \neq 4$, τότε το όριο απόφασης του ταξινομητή περιγράφεται ως $\mathcal{F}_t = z : F(x)_4 - F(x)_t = 0$. Σε κάθε βήμα της επίθεσης, χρησιμοποιώντας το ανάπτυγμα Taylor, το υπερεπίπεδο του ορίου απόφασης ευθυγραμμίζεται παίρνοντας τη μορφή $\mathcal{F}'_t = x : f(x) \approx f(x_0) + < \nabla_x f(x_0) \cdot (x - x_0) >= 0$, όπου χάριν συντομίας ορίζεται $f(x) = F(x)_4 - F(x)_t$. Στη συνέχεια, υπολογίζεται το κάθετο στο επίπεδο F'_t διάνυσμα ω, το οποίο έχει αρχή το x_0 και τέλος σημείο του F'_t. Το διάνυσμα αυτό μπορεί να θεωρηθεί ως η διαταραχή που θα προκαλέσει το x_0 να υπερβεί το όριο απόφασης. Έπειτα από επαρκή αριθμό βημάτων, μετατοπίζοντας το x'_0 κατά μήκος του ω, ο αλγόριθμος DeepFool μπορεί να βρει το αντιφατικό παράδειγμα x'_0 το οποίο θα οδηγήσει τον ταξινομητή να το κατηγοριοποιήσει στην κλάση t αντί της «4».



Σχήμα 3.4: Τα υπερεπίπεδα $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ διαχωρίζουν τα σημεία-δείγματα που ανήκουν στην κλάση «4» από τις κλάσεις «1»,«2» και «3» αντίστοιχα. Το δείγμα x_0 διασχίζει το όριο απόφασης $\mathcal{F}_3$ και έτσι το τροποποιημένο πλέον δείγμα x'_0 ταξινομείται στην κλάση «3». Το ευθυγραμμισμένο όριο απόφασης εμφανίζεται με διακεκομμένες γραμμές (**Πηγή**: [Moos16]).

Τα πειράματα της επίθεσης DeepFool [Moos16] δείχνουν ότι για τους κοινούς ταξινομητές ΒΝΔ στον τομέα της εικόνας, σχεδόν όλα τα δείγματα ελέγχου είναι πολύ κοντά στο όριο απόφασης τους. Για έναν καλά εκπαιδευμένο ταξινομητή LeNet [LeCu89] στο σύνολο δεδομένων MNIST, πάνω από το 90% των δειγμάτων ελέγχου μπορεί να προσβληθεί από μικρές διαταραχές των οποίων η l_∞ νόρμα είναι κάτω από $0,1$ κανονικοποιημένη στο $[0, 1]$. Αυτό υποδηλώνει ότι οι ταξινομητές των ΒΝΔ δεν είναι ισχυροί σε μικρές διαταραχές.

Επίθεση Jacobian-based saliency map

Στην εργασία [Pape15], οι συγγραφείς εισάγουν μια μέθοδο επίθεσης που βασίζεται στον υπολογισμό του Ιακωβιανού πίνακα της συνάρτησης F που υπολογίζει την έξοδο του μοντέλου. Η μέθοδος Jacobian-based saliency map (JSMA) μπορεί να θεωρηθεί ως ένας άπληστος αλγόριθμος επίθεσης, ο οποίος χειρίζεται επαναληπτικά το pixel που επηρεάζει περισσότερο την έξοδο του μοντέλου. Οι συγγραφείς χρησιμοποίησαν τον Ιακωβιανό πίνακα $J_F(x) = \frac{\partial F(x)}{\partial x} = \left\{\frac{\partial F_j(x)}{\partial x_i}\right\}_{i\times j}$ για να μοντελοποιήσουν την αλλαγή της $F(x)$ ως συνέπεια της αλλαγής της εισόδου της x. Στην περίπτωση της στοχευμένης επίθεσης όπου ο αντίπαλος επιδιώκει να δημιουργήσει ένα αντιφατικό δείγμα x' το οποίο να ταξινομείται στην κλάση-στόχο t, οι συγγραφείς προτείνουν την επανειλημμένη αναζήτηση και τροποποίηση x_i pixel, των οποίων η αύξηση της τιμής τους θα προκαλέσει αύξηση της $F_t(x)$ ή μείωση του αθροίσματος $\sum_{j\neq t} F_j(x)$. Κατά συνέπεια, η έξοδος του μοντέλου για το δείγμα x θα έχει μεγαλύτερη τιμή για την κλάση t και εν τέλει το x θα ταξινομηθεί στην κατηγορία t. Με άλλα λόγια, κατά τη διάρκεια της JSMA επίθεσης πραγματοποιείται επαναληπτικά προσθήκη θορύβου σε κανονικό δείγμα x μέχρις ότου ο ταξινομητής C να δώσει την ετικέτα t σε αυτό το δείγμα ή μέχρις ότου φτάσει στο μέγιστο αριθμό προσπαθειών.

Επίθεση Basic Iterative Method / Projected Gradient Descent

Η επίθεση Basic Iterative Method (BIM) είναι μια επαναληπτική εκδοχή της ενός βήματος επίθεσης FGSM που αναλύθηκε προηγουμένως [Kura17b, Kura17a]. Στην περίπτωση μη-στοχευμένης επίθεσης, η μαθηματική διατύπωση της επαναληπτικής αυτής μεθόδου για την κατασκευή του τροποποιημένου δείγματος x' δίνεται από την Εξίσωση 3.8.

$$\begin{aligned} &x_0 = x \\ &x^{t+1} = Clip_{x,\epsilon}\left(x^t + \alpha \text{sgn}\left(\nabla_x \mathcal{L}\left(\theta, x^t, y\right)\right)\right) \end{aligned} \tag{3.8}$$

Στην παραπάνω Εξίσωση, με τον όρο $Clip$ ορίζουμε τη συνάρτηση η οποία περιορίζει την παράμετρό της στην επιφάνεια της γειτονιάς του x. Η γειτονιά του x είναι μια σφαίρα η οποία περιγράφεται πλήρως από την εξής σχέση: $B_\epsilon(x) : \{x' : ||x' - x||_\infty \leq \epsilon\}$. Επιπλέον, η τιμή της παραμέτρου α, που αποτελεί το βήμα της μεθόδου, είναι συνήθως σχετικά μικρή. Αυτή η επαναληπτική μέθοδος επίθεσης είναι επίσης γνωστή ως Projected Gradient Method (PGD), εάν στον αλγόριθμο προστεθεί τυχαία αρχικοποίηση του x [Madr19].

Η επίθεση BIM (ή PGD) αναζητά μέσω ενός ευριστικού μηχανισμού τα δείγματα x' που έχουν τη μεγαλύτερη τιμή σφάλματος στα πλαίσια της σφαίρας γύρω από το αρχικό δείγμα x που έχει οριστεί μέσω της l_∞ νόρμας. Αυτά τα αντιφατικά παραδείγματα που δημιουργούνται μέσω της συγκεκριμένης επίθεσης ονομάζονται «most-adversarial» παραδείγματα, καθώς η πιθανότητα να εξαπατήσουν τον ταξινομητή, στα πλαίσια πάντα της περιορισμένης έντασης της διαταραχής που οριοθετούνται από κάποια l_p νόρμα, είναι μέγιστη. Σε αυτό το σημείο μπορούμε πλέον με βεβαιότητα να δεχτούμε ότι η εύρεση αυτών των αντιφατικών δειγμάτων είναι εξαιρετικά χρήσιμη στην εύρεση των αδυναμιών των μοντέλων βαθιάς μηχανικής μάθησης.

Επίθεση Carlini and Wagner

Η επίθεση Carlini και Wagner (C&W) [Carl17b] αντεπιτίθεται στην αμυντική στρατηγική της *απόσταξης* (distillation) [Pape16] που αποδείχθηκε επιτυχημένη κατά των επιθέσεων FGSM και L-BFGS. Η συγκεκριμένη επίθεση στοχεύει στην επίλυση του ίδιου προβλήματος όπως ορίζεται στην επίθεση L-BFGS μέσω της Σχέσης 3.4, δηλαδή στην προσπάθεια εύρεσης της ελάχιστης παραμόρφωσης. Οι συγγραφείς του άρθρου [Carl17b] αντιμετωπίζουν το πρόβλημα που περιγράφεται στην Εξίσωση 3.4 επιλύοντας αντ 'αυτού το πρόβλημα που ορίζεται στην Εξίσωση 3.9]:

$$\begin{aligned} &\min ||x - x'||_2^2 + c \cdot f(x', t) \\ &s.t.\ x' \in [0,\ 1]^m \end{aligned} \tag{3.9}$$

όπου η f ορίζεται ως $f(x', t) = (\max_{i \neq t} = Z(x')_i - Z(x')_t)^+$. Η ελαχιστοποίηση της $f(x', t)$ ενθαρρύνει τον αλγόριθμο να βρει ένα x' που έχει μεγαλύτερη τιμή εξόδου για την κλάση t από οποιαδήποτε άλλη κατηγορία, έτσι ώστε ο ταξινομητής να προβλέψει ότι το x ανήκει στην κλάση t. Στη συνέχεια, εφαρμόζοντας μια γραμμική αναζήτηση στη σταθερά c, μπορούμε να βρούμε το x' που έχει τη μικρότερη απόσταση από το x. Η συνάρτηση $f(x, y)$ μπορεί επίσης να θεωρηθεί ως συνάρτηση σφάλματος των δεδομένων (x, y) αφού λειτουργεί τιμωρητικά σε περιπτώσεις όπου υπάρχουν ορισμένες κλάσεις i με τιμές εξόδου $Z(x)_i$ μεγαλύτερες από $Z(x)_y$.

Η μόνη διαφορά μεταξύ αυτής της μεθόδου και της επίθεσης L-BFGS είναι ότι η επίθεση C&W χρησιμοποιεί την $f(x, t)$ ως συνάρτηση σφάλματος αντί της διασταυρούμενης εντροπίας $\mathcal{L}(x, t)$. Το πλεονέκτημα χρήσης της συγκεκριμένης συνάρτησης σφάλματος είναι ότι όταν $C(x') = t$, δηλαδή η τιμή της συνάρτησης θα είναι $f(x', t) = 0$, ο αλγόριθμος θα ελαχιστοποιήσει άμεσα την απόσταση από το x' στο x. Αυτή η διαδικασία είναι πιο αποτελεσματική στην εύρεση του αντιφατικού παραδείγματος που έχει υποστεί την λιγότερη παραμόρφωση.

Οι συγγραφείς ισχυρίζονται ότι η επίθεσή τους είναι μία από τις ισχυρότερες επιθέσεις, αντικρούοντας πολλές αμυντικές στρατηγικές που αποδείχθηκαν επιτυχημένες. Έτσι, η επίθεση τους μπορεί να χρησιμοποιηθεί ως σημείο αναφοράς για την εξέταση της ασφάλειας των ταξινομητών των ΒΝΔ ή αντίστοιχα της αποτελεσματικότητας άλλων αντιφατικών παραδειγμάτων.

Επίθεση ground truth

Χρόνο με τον χρόνο, τόσο οι επιθέσεις όσο και οι άμυνες βελτιώνονται συνεχώς προσπαθώντας να υπερνικήσει η μία την άλλη, δημιουργώντας έναν ατέρμονα βρόχο διαρκούς βελτίωσης. Προκειμένου να δοθεί μια λύση σε αυτό το αδιέξοδο, ο συγγραφέας της εργασίας [Carl18] προσπαθεί να βρει την «αποδεκτά ισχυρότερη επίθεση». Μπορεί να θεωρηθεί ως μια μέθοδος για την εύρεση των θεωρητικών ελάχιστων διαταραχών που χρειάζεται να προστεθούν σε κανονικά δείγματα προκειμένου να δημιουργηθούν αντιφατικά παραδείγματα. Αυτή η επίθεση βασίζεται στο Reluplex [Katz17], έναν αλγόριθμο για την επαλήθευση των ιδιοτήτων των νευρωνικών δικτύων. Ο αλγόριθμος αυτός κωδικοποιεί τις παραμέτρους του μοντέλου - F και τα δεδομένα (x, y) - ως μεταβλητές ενός γραμμικού προγραμματιστικού συστήματος και στη συνέχεια λύνει το σύστημα για να ελέγξει εάν υπάρχει κατάλληλο δείγμα x' στη γειτονιά $B_\epsilon(x)$ του x που μπορεί να ξεγελάσει το μοντέλο. Πιο συγκεκριμένα, ο αλγόριθμος αρχίζει να μειώνει την ακτίνα ϵ της περιοχής αναζήτησης $B_\epsilon(x)$, έως ότου το σύστημα διαπιστώσει ότι δεν υπάρχει τέτοιο x' που να μπορεί να εξαπατήσει το μοντέλο. Το τελευταίο τροποποιημένο παράδειγμα που προέκυψε από τον παραπάνω αλγόριθμο ονομάζεται παράδειγμα ground truth, επειδή έχει αποδειχθεί ότι έχει την ελάχιστη ανομοιομορφία με το x.

Η επίθεση ground truth είναι η πρώτη εργασία που υπολογίζει σοβαρά την ακριβή ανθεκτικότητα (ελάχιστη διαταραχή) των ταξινομητών. Ωστόσο, αυτή η μέθοδος περιλαμβάνει τη χρήση του SMT, ενός πολύπλοκου αλγορίθμου ο οποίος ελέγχει εάν ικανοποιείται μια σειρά θεωριών, κάτι που την καθιστά αργή και μη επεκτάσιμη σε μεγάλα δίκτυα. Οι πιο πρόσφατες έρευνες [Tjen19, Xiao19b] έχουν βελτιώσει την αποτελεσματικότητα της επίθεσης αυτής.

Άλλες l_p επιθέσεις

Οι προηγούμενες μελέτες που είδαμε επικεντρώνονται κυρίως σε διαταραχές που περιορίζονται από τις νόρμες l_2 ή l_∞. Ωστόσο, υπάρχουν επιστημονικές έρευνες οι οποίες εξετάζουν και άλλους τύπους επιθέσεων l_p.

1. **One-pixel επίθεση**: Στην εργασία [Su19b], οι συγγραφείς μελετούν ένα παρόμοιο πρόβλημα με αυτό της Εξίσωσης 3.4, αλλά περιορίζουν τον κανόνα της διαταραχής μέσω της l_0 νόρμας. Ο περιορισμός αυτός της διαταραχής $x' - x$ θα περιορίσει τον αριθμό των pixel που επιτρέπεται να τροποποιηθούν. Στην ουσία, η one-pixel τροποποίηση μπορεί να θεωρηθεί ως μια διαταραχή των δεδομένων κατά μήκος μιας κατεύθυνσης παράλληλης προς τον άξονα μίας από τις n διαστάσεις-pixel της εικόνας.

Για την επίλυση του παραπάνω προβλήματος, οι συγγραφείς χρησιμοποίησαν έναν αλγόριθμο βελτιστοποίησης βάσει πληθυσμού, αυτόν της *διαφορικής εξέλιξης* (differential evolution), ο οποίος ανήκει στην ευρύτερη κατηγορία των *εξελικτικών αλγορίθμων* (evolutionary algorithms). Στην πράξη, ο αλγόριθμος αυτός αναμένεται να βρίσκει υψηλότερης ποιότητας λύσεις από ότι οι βασισμένοι στην κλίση αλγόριθμοι ή ακόμη και άλλα είδη εξελικτικών αλγορίθμων. Πιο συγκεκριμένα, κατά τη διάρκεια κάθε επανάληψης του αλγορίθμου δημιουργείται ένα νέο σύνολο υποψήφιων λύσεων (παιδιά) σύμφωνα με τον τρέχοντα πληθυσμό (γονείς). Στη συνέχεια, τα παιδιά συγκρίνονται με τους αντίστοιχους γονείς τους και εάν είναι πιο κατάλληλα από τους γονείς τους επιβιώνουν. Η διαφορική εξέλιξη δεν χρησιμοποιεί πληροφορίες από την κλίση για βελτιστοποίηση και για αυτό μπορεί να χρησιμοποιηθεί σε ένα ευρύτερο φάσμα προβλημάτων βελτιστοποίησης.

Μέσα από την εργασία [Su19b] επισημαίνεται ότι στο σύνολο δεδομένων CIFAR10 [onlib], για ένα καλά εκπαιδευμένο CNN (π.χ. VGG16 [Simo15], με ακρίβεια $85,5\%$ στα δεδομένα ελέγχου), τα περισσότερα δείγματα του συνόλου ελέγχου ($63,5\%$) μπορούν να υποστούν επιτυχώς μια μη στοχευμένη επίθεση αλλάζοντας την τιμή μόνο ενός pixel. Το γεγονός αυτό τονίζει επίσης την κακή ευρωστία των μοντέλων βαθιάς μάθησης.

2. **Elastic-Net επίθεση**: Στην εργασία [Chen18], ο συγγραφέας μελετά ένα παρόμοιο πρόβλημα με αυτό της Εξίσωσης 3.4, αλλά περιορίζει τον κανόνα της διαταραχής μέσω των νορμών l_1 και l_2 εφαρμόζοντάς τες συγχρόνως. Όπως φαίνεται στην εργασία [Shar18a], ορισμένα ισχυρά αμυντικά μοντέλα που στοχεύουν στην απόρριψη των επιθέσεων που βασίζονται στις l_∞ και l_2 νόρμες [Madr19] εξακολουθούν να είναι ευάλωτα στην επίθεση Elastic-Net με βάση τον περιορισμό της l_1 νόρμας.

Καθολική (universal) επίθεση

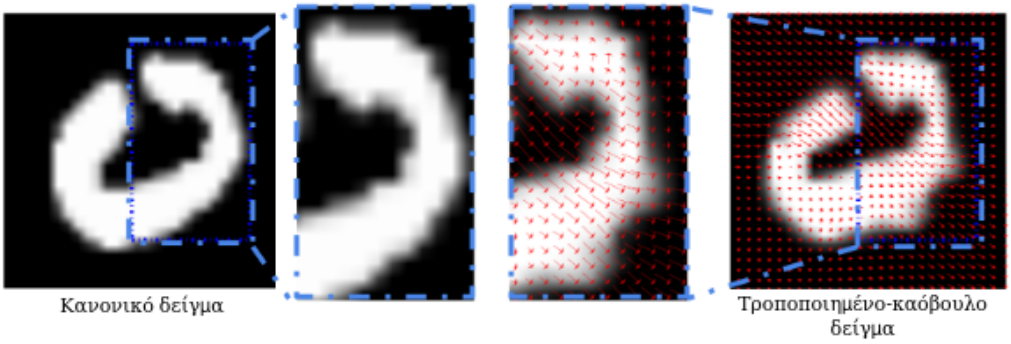
Οι μέθοδοι που αναλύθηκαν προηγουμένως εξετάζουν μόνο ένα συγκεκριμένο στοχευμένο δείγμα-στόχο x. Ωστόσο, στην εργασία [Moos17], οι συγγραφείς επινοούν έναν αλγόριθμο που παραπλανά επιτυχώς την απόφαση ενός ταξινομητή για όλες σχεδόν τις εικόνες του συνόλου ελέγχου. Προσπαθούν να βρουν μια διαταραχή δ που να ικανοποιεί τις εξής απαιτήσεις που ορίζονται στην Εξίσωση 3.10]:

$$\begin{aligned} &1.\ \ \delta||_p \leq\ \epsilon \\ &2.\ \ \mathbb{P}_{x\sim D(x)}\ \ (C(x+\delta) \neq C(x)) \leq 1{-}\sigma \end{aligned} \tag{3.10}$$

Αυτή η μέθοδος στοχεύει να βρει μια διαταραχή δ τέτοια ώστε ο ταξινομητής να δίνει λανθασμένες αποφάσεις στα περισσότερα από τα δείγματα. Στα πειράματά τους, για παράδειγμα, βρήκαν μια διαταραχή που μπορεί να εφαρμοστεί επιτυχώς στο $85,4\%$ των δειγμάτων ελέγχου του συνόλου δεδομένων ILSVRC 2012 [Russ14] και να επιτεθεί σε έναν ταξινομητή ResNet-152 [He16]. Η ύπαρξη *«καθολικών» αντιφατικών παραδειγμάτων* (universal adversarial examples) αποκαλύπτει την εγγενή αδυναμία των ταξινομητών των ΒΝΔ που παρουσιάζουν σε όλα τα δείγματα εισόδου.

Επίθεση χωρικής μετατροπής

Οι παραδοσιακοί αλγόριθμοι επίθεσης τροποποιούν άμεσα την τιμή των pixel μιας εικόνας, η οποία αλλάζει την ένταση του χρώματος του συγκεκριμένου pixel. Στην εργασία [Xiao18] επινοείται μια άλλη μέθοδος σύμφωνα με την οποία η εικόνα διαταράσσεται και τροποποιείται εφαρμόζοντας ελαφρύ *χωρικό μετασχηματισμό* (spatial transformation), όπου τα τοπικά χαρακτηριστικά της εικόνας περιστρέφονται και παραμορφώνονται ελαφρά. Η διαταραχή είναι αρκετά μικρή για να γίνει αντιληπτή από την ανθρώπινο παράγοντα, αλλά μπορεί να ξεγελάσει τους ταξινομητές όπως φαίνεται στο Σχήμα 3.5.



Σχήμα 3.5: Δημιουργία αντιφατικού δείγματος από το σύνολο δεδομένων MNIST. Το πάνω μέρος του ψηφίου «0» είναι τροποποιημένο και πιο λεπτό από ότι στην αρχική εικόνα. Ο ταξινομητής εσφαλμένα κατηγοριοποιεί το ψηφίο «0» ως «2» (**Πηγή**: [Xiao18]).

Επίθεση χωρίς περιορισμούς στη δημιουργία αντιφατικών δειγμάτων

Οι προηγούμενες μέθοδοι επίθεσης εξετάζουν μόνο την προσθήκη ελάχιστων και απαρατήρητων διαταραχών σε εικόνες. Στην εργασία [Song18b] οι συγγραφείς εισάγουν μια μέθοδο για τη δημιουργία απεριόριστων αντιφατικών παραδειγμάτων. Αυτά τα παραδείγματα δεν είναι ακριβώς τα ίδια με τα δείγματα-στόχους, αλλά εξακολουθούν να κατηγοριοποιούνται από τον ανθρώπινο παράγοντα στη σωστή κλάση και δύναται να εξαπατήσουν τον ταξινομητή. Παλαιότερες επιτυχημένες αμυντικές στρατηγικές, οι οποίες στοχεύουν σε επιθέσεις που βασίζονται σε περιορισμένες διαταραχές, αποτυγχάνουν να αναγνωρίσουν τέτοιου είδους επιθέσεις.

Για να πραγματοποιηθεί μια επίθεση απέναντι σε έναν ταξινομητή C, οι συγγραφείς της παραπάνω μελέτης αρχικά προεκπαίδευσαν έναν βοηθητικό ταξινομητή μέσω ενός GAN, [Oden17], έτσι ώστε να μπορούν να δημιουργήσουν ένα κανονικό δείγμα x από ένα διάνυσμα θορύβου z_0, που να ανήκει στην κλάση y. Στη συνέχεια, για να δημιουργήσουν ένα αντιφατικό παράδειγμα, βρήκαν ένα διάνυσμα θορύβου z κοντά στο z_0, αλλά απαίτησαν η έξοδος του GAN $G(z)$ να ταξινομηθεί εσφαλμένα από τον ταξινομητή C του μοντέλου-στόχου. Επειδή το z είναι κοντά στο z_0 στον *λανθάνοντα χώρο* (latent space) του AC-GAN, η έξοδος του πρέπει να ανήκει στην ίδια κατηγορία y. Με αυτόν τον τρόπο, το παραγόμενο δείγμα $G(z)$ είναι διαφορετικό από το x, παραπλανεί τον ταξινομητή C, αλλά εξακολουθεί να είναι ένα δείγμα που ανήκει στην y κλάση στα μάτια ενός ανθρώπου.

3.2.2 Επιθέσεις στο φυσικό κόσμο

Όλες οι μέθοδοι επίθεσης που αναπτύχθηκαν στις προηγούμενες ενότητες εφαρμόζονται ψηφιακά, όπου ο αντίπαλος παρέχει εικόνες εισόδου απευθείας στο μοντέλο μηχανικής εκμάθησης. Ωστόσο, αυτό δεν ισχύει πάντα καθώς υπάρχουν ορισμένα σενάρια, όπως αυτά που χρησιμοποιούν κάμερες, μικρόφωνα ή άλλους αισθητήρες για τη λήψη των σημάτων εισόδου τους, στα οποία τα δείγματα εισόδου προέρχονται από τον φυσικό κόσμο και όχι απευθείας από τον ψηφιακό. Σε αυτές τις περιπτώσεις, τίθεται το ερώτημα αν ακόμη υπάρχει η δυνατότητα της επίθεσης σε αυτά τα συστήματα δημιουργώντας αντικείμενα του φυσικού κόσμου. Πρόσφατα μελέτες δείχνουν ότι τέτοιες επιθέσεις υπάρχουν όπως για παράδειγμα, στο άρθρο [Eykh18] περιγράφεται η προσάρτηση αυτοκόλλητων σε πινακίδες οδικής κυκλοφορίας που μπορεί να παραπλανήσει σοβαρά το σύστημα αναγνώρισης πινακίδων των αυτόνομων οχημάτων. Αυτά τα είδη αντιφατικών αντικειμένων είναι ακόμα πιο καταστροφικά για τα μοντέλα βαθιάς μάθησης, αφού μπορούν να επιτεθούν άμεσα σε πολλές πρακτικές εφαρμογές των ΒΝΔ, όπως στην αναγνώριση προσώπου και στα συστήματα αυτόνομων οχημάτων.

Επιθέσεις αντιφατικών παραδειγμάτων στο φυσικό κόσμο

Στην εργασία [Kura17a], οι συγγραφείς ερευνούν τη δυνατότητα δημιουργίας *αντιφατικών αντικειμένων* (adversarial objects) στο φυσικό κόσμο, ελέγχοντας εάν οι παραγόμενες από τις μεθόδους FGSM και BIM αντιφατικές εικόνες είναι *ανθεκτικές* όταν υποβληθούν σε κάποιο φυσικό μετασχηματισμό, όπως αλλαγή της οπτικής γωνίας ή του φωτισμού. Με τον όρο «ανθεκτικές» προσπαθούμε να αποδώσουμε την ιδιότητα που επιδιώκουμε να έχουν οι δημιουργημένες εικόνες, ώστε να παραμένουν αντιφατικές και μετά τον μετασχηματισμό.

Για να εφαρμοστεί αυτός ο μετασχηματισμός, πρέπει πρώτα να εκτυπωθούν οι κατασκευασμένες εικόνες και έπειτα η ομάδα των ατόμων που έχει καθοριστεί για την εκτέλεση του πειράματος να τραβήξει φωτογραφίες αυτών των εκτυπώσεων κάνοντας χρήση κινητών τηλεφώνων. Κατά τη διάρκεια αυτής της διαδικασίας, δεν υπάρχει κανένας περιορισμός ως προς τη γωνία λήψης ή το φωτισμό του εξωτερικού περιβάλλοντος, επομένως οι ληφθείσες φωτογραφίες αποτελούν μετασχηματισμένα δείγματα των αντιφατικών παραδειγμάτων που δημιουργήθηκαν προηγουμένως. Τα πειραματικά αποτελέσματα που προέκυψαν από την παραπάνω έρευνα καταδεικνύουν ότι μετά τον μετασχηματισμό, ένα μεγάλο μέρος αυτών των αντιφατικών παραδειγμάτων, ειδικά εκείνων που δημιουργήθηκαν μέσω του FGSM αλγορίθμου, παρέμειναν αντιφατικά εξαπατώντας τον ταξινομητή του μοντέλου-θύματος. Τα συμπεράσματα αυτά υποδηλώνουν τη ρεαλιστική πιθανότητα δημιουργίας φυσικών αντιφατικών αντικειμένων που μπορούν να παραπλανήσουν τον αισθητήρα κάτω από διαφορετικά περιβάλλοντα.

Επίθεση Eykholt σε πινακίδες οδικής σήμανσης

Στην εργασία [Eykh18], οι συγγραφείς περιγράφουν τη δημιουργία φυσικών αντιφατικών αντικειμένων, τα οποία τροποποιούν οδικές πινακίδες για να εξαπατήσουν συστήματα αναγνώρισης πινακίδων οδικής κυκλοφορίας. Η επίθεση επιτυγχάνεται μετά την τοποθέτηση αυτοκόλλητων στα σήματα στοπ σε συγκεκριμένες επιθυμητές θέσεις. Αναλυτικότερα, για την εκτέλεση της εν λόγω επίθεσης πρέπει να ακολουθηθούν τα εξής 3 βήματα:

1. Αρχικά, ο επιτιθέμενος καλείται να εκτελέσει μια επίθεση με βάση την l_1 νόρμα - περιορίζοντας δηλαδή τη διαταραχή σε $||x' - x||_1 < \epsilon$ - σε ψηφιακές εικόνες οδικών πινακίδων για την εύρεση περίπου της περιοχής που θα εισαχθεί η διαταραχή. Αυτές οι περιοχές θα αποτελέσουν αργότερα τα σημεία τοποθέτησης των αυτοκόλλητων. Γενικά, οι επιθέσεις με βάση την l_1 νόρμα δημιουργούν αραιή διαταραχή, η οποία βοηθά στην εύρεση της κατάλληλης τοποθεσίας για την πραγματοποίηση της επίθεσης.

2. Στη συνέχεια, οφείλει να επικεντρωθεί στις περιοχές που υπολογίστηκαν στο προηγούμενο βήμα και να εφαρμόσει μια επίθεση με βάση την l_2 νόρμα για να δημιουργήσει το χρώμα για τα αυτοκόλλητα.

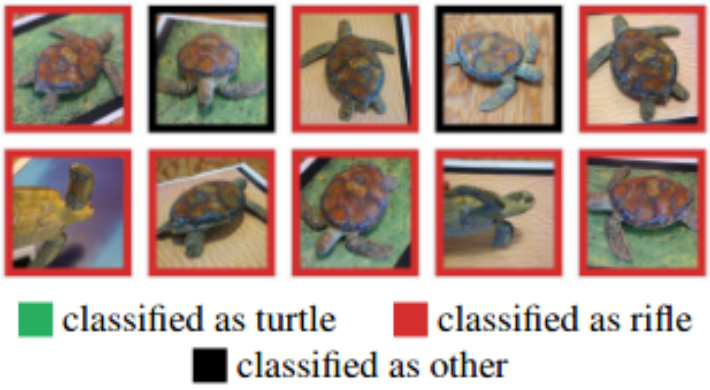
3. Τέλος, θα πρέπει να εκτυπώσει τα κομμάτια από τη διαταραχή που υπολογίστηκε στα προηγούμενα βήματα και να τα κολλήσει στην πινακίδα, όπως φαίνεται και στο Σχήμα 3.6. Το τροποποιημένο πλέον σήμα στοπ δύναται να παραπλανήσει ένα αυτόνομο όχημα από οποιαδήποτε απόσταση και οπτική γωνία.

Δημιουργία 3D αντιφατικού αντικειμένου Athalye

Η εργασία [Atha18b] αποτελεί την πρώτη δημοσίευση που παρουσιάζει τη δημιουργία επιτυχημένων *φυσικών αντιφατικών 3D αντικειμένων* (physical 3D adversarial objects). Όπως φαίνεται στο Σχήμα 3.7, οι συγγραφείς χρησιμοποιούν τρισδιάστατη εκτύπωση για την κατασκευή μιας «αντιφατικής» χελώνας. Για την επίτευξη του στόχου, εφαρμόζουν μια 3D τεχνική κατασκευής των αντικειμένων. Με δεδομένο ένα τρισδιάστατο αντικείμενο με χαρακτηριστική υφή, βελτιστοποιούν πρώτα την

Σχήμα 3.6: Επίθεση σε πινακίδα στοπ (**Πηγή**:[Eykh18]).

υφή του αντικειμένου έτσι ώστε οι κατασκευασμένες εικόνες να είναι αντιφατικές από οποιαδήποτε οπτική γωνία. Κατά τη διαδικασία αυτή, διασφαλίζεται επίσης ότι η διαταραχή παραμένει αντιφατική κάτω από διαφορετικές συνθήκες και περιβάλλοντα, αλλάζοντας για παράδειγμα την απόσταση ή τον προσανατολισμό της κάμερας, τις συνθήκες φωτισμού και το φόντο. Αφού εντοπιστεί η διαταραχή στην 3D εκδοχή της, τότε εκτυπώνεται ένα στιγμιότυπο του τρισδιάστατου αντικειμένου.



Σχήμα 3.7: Τυχαίες πόζες μιας «αντιφατικής» χελώνας σε τρισδιάστατη εκτύπωση, η οποία κατηγοριοποιείται πλέον από τον ταξινομητή στην κλάση «τουφέκι» (**Πηγή**: [Oden17]).

3.2.3 Επιθέσεις black box

Υποκατάστατο μοντέλο

Η εργασία [Pape17] ήταν η πρώτη που εισήγαγε έναν αποτελεσματικό αλγόριθμο για να επιτεθεί στους ταξινομητές ΒΝΔ, με την υπόθεση ότι ο αντίπαλος δεν έχει πρόσβαση στις παραμέτρους του ταξινομητή ούτε στο σύνολο δεδομένων εκπαίδευσης. Ένας αντίπαλος μπορεί να τροφοδοτήσει μόνο την είσοδο x για να λάβει την ετικέτα εξόδου y από τον ταξινομητή. Επιπλέον, ο αντίπαλος μπορεί να έχει μόνο μερική γνώση σχετικά με: (i) τον τομέα δεδομένων του ταξινομητή (π.χ. χειρόγραφα ψηφία, φωτογραφίες, ανθρώπινα πρόσωπα) και (ii) την αρχιτεκτονική του ταξινομητή (π.χ. CNN, κ.λ.π.).

Η παραπάνω μελέτη εκμεταλλεύεται τη *δυνατότητα μεταφοράς* που παρουσιάζουν τα αντιφατικά παραδείγματα. Με τον όρο δυνατότητα μεταφοράς εννοούμε το γεγονός ότι ένα δείγμα x' που μπορεί να επιτεθεί στο F_1, είναι επίσης πιθανό να μπορεί να επιτεθεί επιτυχώς και στο F_2, το οποίο έχει παρόμοια δομή με το F_1. Έτσι, χρησιμοποιώντας την παραπάνω ιδιότητα, οι συγγραφείς εισάγουν μια μέθοδο κατά την οποία αρχικά εκπαιδεύουν ένα υποκατάστατο μοντέλο F' για να μιμηθεί τον ταξινομητή-θύμα F, και στη συνέχεια δημιουργούν το αντιφατικό παράδειγμα μέσω της επίθεσης στο υποκατάστατο μοντέλο F'. Τα κύρια βήματα της μεθόδου αυτής είναι τα εξής:

1. *Σύνθεση υποκατάστατου συνόλου δεδομένων εκπαίδευσης* (substitute training dataset): Το πρώτο βήμα περιλαμβάνει τη δημιουργία ενός αντιγράφου του συνόλου δεδομένων εκπαίδευσης. Για

παράδειγμα, στην περίπτωση που ένας αντίπαλος επιθυμεί να επιτεθεί σε έναν ταξινομητή-θύμα αναγνώρισης χειρόγραφων ψηφίων, τότε θα πρέπει πρώτα να δημιουργήσει ένα υποκατάστατο σύνολο δεδομένων εκπαίδευσης: (i) χρησιμοποιώντας δείγματα από το σύνολο ελέγχου ή (ii) κατασκευάζοντας δείγματα «με το χέρι».

2. *Εκπαίδευση του υποκατάστατου μοντέλου*: Στο δεύτερο βήμα ο αντίπαλος καλείται να εισάγει το υποκατάστατο σύνολο δεδομένων εκπαίδευσης X, που δημιούργησε στο προηγούμενο βήμα, στον ταξινομητή-θύμα προκειμένου να αποκτήσει τις ετικέτες του Y. Στη συνέχεια, οφείλει να επιλέξει ένα υποκατάστατο μοντέλο ΒΝΔ, ώστε να το εκπαιδεύσει πάνω στα (X, Y) ζεύγη και να λάβει το F'. Με βάση τις γνώσεις του εισβολέα, το επιλεγμένο ΒΝΔ θα πρέπει να έχει παρόμοια δομή με το μοντέλο-θύμα.

3. *Αύξηση συνόλου δεδομένων*: Στο τρίτο βήμα ο αντίπαλος αυξάνει το σύνολο δεδομένων (X, Y) και επανεκπαιδεύει το υποκατάστατο μοντέλο F' επαναληπτικά. Αυτή η διαδικασία συμβάλλει στην αύξηση της ποικιλομορφίας του αντιγράφου του συνόλου δεδομένων εκπαίδευσης και στη βελτίωση της ακρίβειας του υποκατάστατου μοντέλου F'.

4. *Επίθεση στο υποκατάστατο μοντέλο*: Χρησιμοποιώντας τις μεθόδους επίθεσης που αναπτύχθηκαν σε προηγούμενες ενότητες, όπως την FGSM, ο εισβολέας επιτίθεται στο μοντέλο F'. Τα αντιφατικά παραδείγματα που δημιουργούνται είναι επίσης πολύ πιθανό να καταφέρουν να παραπλανήσουν το μοντέλο-θύμα F λόγω της δυνατότητας μεταφοράς τους.

Για να απαντήσουμε στο ερώτημα ποιος αλγόριθμος είναι ο καταλληλότερος για την υλοποίηση της παραπάνω επίθεσης, θα πρέπει πρώτα να αναλογιστούμε το γεγονός ότι η επιτυχία της επίθεσης black- box βασίζεται στη δυνατότητα μεταφοράς των αντιφατικών παραδειγμάτων. Για το λόγο αυτό, κατά την υλοποίηση μιας επίθεσης black- box, επιλέγουμε αλγορίθμους που έχουν υψηλή δυνατότητα μεταφοράς, όπως ο FGSM και ο PGD, καθώς και επαναληπτικούς αλγορίθμους που χρησιμοποιούν το μέγεθος της ορμής στους υπολογισμούς της συνάρτησης σφάλματος [Dong18].

Επίθεση ZOO (Zeroth Order Optimization based)

Σε αντίθεση με την περίπτωση που αναλύσαμε στην προηγούμενη ενότητα, όπου ο αντίπαλος μπορεί να λάβει μόνο τις πληροφορίες της ετικέτας από τον ταξινομητή, στην εργασία [Chen17] οι συγγραφείς υποθέτουν ότι μέσα από την έξοδο του ταξινομητή-θύματος, ο εισβολέας έχει πρόσβαση στην τιμή που αφορά στην εμπιστοσύνη της πρόβλεψης. Σε αυτήν την περίπτωση, ο εισβολέας δεν χρειάζεται να δημιουργήσει το υποκατάστατο σύνολο δεδομένων εκπαίδευσης ούτε να χτίσει το υποκατάστατο μοντέλο. Οι συγγραφείς του παραπάνω έργου παρουσιάζουν έναν αλγόριθμο για να καταφέρει ο αντίπαλος να εξάγει τις απαραίτητες πληροφορίες γύρω από το δείγμα-θύμα x, παρατηρώντας τις αλλαγές στην εμπιστοσύνη πρόβλεψης $F(x)$ που προκαλούνται λόγω των μεταβολών στις τιμές των pixel του x.

$$\frac{\partial F(x)}{\partial x_i} \approx \frac{F(x + he_i) - F(x - he_i)}{2h} \tag{3.11}$$

Αναλύοντας την Εξίσωση 3.11 παρατηρούμε ότι για κάθε δείκτη i του δείγματος x, προσθέτουμε στο x_i (ή αφαιρούμε από αυτό) το h. Εάν το h είναι αρκετά μικρό, μπορούμε να αποκόψουμε τις πληροφορίες σχετικά με την κλίση του $F(x)$ από την έξοδο του $F(\cdot)$. Χρησιμοποιώντας την κατά προσέγγιση μερική παράγωγο της Εξίσωσης 3.11, μπορούμε να εφαρμόσουμε τις μεθόδους επίθεσης Carlini & Wagner και FGSM που παρουσιάστηκαν σε προηγούμενες ενότητες. Το ποσοστό επιτυχίας της επίθεσης ZOO είναι υψηλότερο από αυτό της τεχνικής του υποκατάστατου μοντέλου επειδή μπορεί να χρησιμοποιήσει τις πληροφορίες της εμπιστοσύνης πρόβλεψης, αντί τις προβλεπόμενες ετικέτες αποκλειστικά.

Επίθεση με χρήση αποτελεσματικών ερωτημάτων

Οι επιθέσεις black-box που παρουσιάστηκαν προηγουμένως απαιτούν την υποβολή πολλών ερωτημάτων στην είσοδο του ταξινομητή-θύματος, κάτι που στην πράξη μπορεί να είναι απαγορευτικό σε πραγματικές εφαρμογές. Ο παραπάνω προβληματισμός οδήγησε στην ύπαρξη ορισμένων ερευνών που μελετούν τη βελτίωση της αποτελεσματικότητας των αντιφατικών παραδειγμάτων μέσω περιορισμένου αριθμού ερωτημάτων. Για παράδειγμα στην εργασία [Ilya18], εισάγεται ένας πιο αποτελεσματικός τρόπος για την εκτίμηση των πληροφοριών της κλίσης του $F(x)$ από την έξοδο $F(x)$ του μοντέλου. Πιο συγκεκριμένα, οι συγγραφείς του [Ilya18] χρησιμοποιούν τις στρατηγικές φυσικής εξέλιξης [Wier11] για να δειγματοληπτήσουν τα αποτελέσματα της εξόδου του μοντέλου με βάση τα ερωτήματα που τέθηκαν γύρω από το δείγμα-θύμα x και να εκτιμήσουν την τιμή της κλίσης του F στο x. Αυτή η διαδικασία απαιτεί την υποβολή λιγότερων ερωτήματα στο μοντέλο. Επιπλέον, ένας γενετικός αλγόριθμος εφαρμόζεται για την αναζήτηση αντιφατικών παραδειγμάτων στους γείτονες μιας έγκυρης εικόνας [Alza19].

3.2.4 Επιθέσεις grey (semi-white) box

Στην εργασία [Xiao19a], παρουσιάζεται ένα πλαίσιο επιθέσεων semi-white box. Στην αρχή ο αντίπαλος οφείλει να εκπαιδεύσει ένα GAN [Good14], στοχεύοντας το μοντέλο που επιθυμεί. Στη συνέχεια, ο εισβολέας μπορεί να δημιουργήσει αντιφατικά παραδείγματα απευθείας από το δίκτυο που δημιούργησε ο ίδιος. Το πλεονέκτημα της επίθεσης που βασίζεται στο μοντέλο GAN είναι ότι πρώτον επιταχύνει τη διαδικασία κατασκευής αντιφατικών παραδειγμάτων και δεύτερον τα αντιφατικά αυτά δείγματα είναι πιο φυσικά και λιγότερο ανιχνεύσιμα από αμυντικές στρατηγικές. Στην εργασία [Deb19], το GAN χρησιμοποιείται για τη δημιουργία αντιφατικών προσώπων, ώστε ο εισβολέας να καταφέρει να παραπλανήσει το λογισμικό αναγνώρισης προσώπου. Από τα πειραματικά αποτελέσματα της παραπάνω έρευνας συμπεραίνουμε ότι οι κατασκευασμένες εικόνες προσώπου φαίνεται να είναι πιο φυσικές και η διαφορά τους από τις πραγματικές εικόνες προσώπου-στόχου είναι οριακά διακριτή.

3.2.5 Επιθέσεις δηλητηρίασης

Οι επιθέσεις που αναπτύξαμε μέχρι στιγμής είναι επιθέσεις διαφυγής, οι οποίες εφαρμόζονται αφού έχει προηγηθεί η εκπαίδευση του μοντέλου ταξινόμησης. Αντ' αυτού ορισμένες μελέτες δημιουργούν αντιφατικά παραδείγματα πριν από την εκπαίδευση του ταξινομητή. Αυτά τα εχθρικά δείγματα εισάγονται στο σύνολο δεδομένων εκπαίδευσης για να υπονομεύσουν τη συνολική ακρίβεια του ταξινομητή ή να επηρεάσουν την πρόβλεψή του σε συγκεκριμένα δείγματα ελέγχου. Αυτή η διαδικασία αποτελεί τη βασική μέθοδο που ακολουθείται στις επιθέσεις δηλητηρίασης.

Επίθεση δηλητηρίασης Biggio σε SVM

Η εργασία [Bigg13b] εισάγει μια μέθοδο δηλητηρίασης του συνόλου δεδομένων εκπαίδευσης προκειμένου να μειώσει την ακρίβεια του μοντέλου SVM. Σύμφωνα με τη μέθοδο αυτή, οι συγγραφείς προσπαθούν να δημιουργήσουν ένα δηλητηριασμένο δείγμα x_c το οποίο, όταν εισαχθεί στο σύνολο δεδομένων εκπαίδευσης, θα οδηγήσει το εκπαιδευμένο SVM μοντέλο F_{x_c} να έχει ένα μεγάλο συνολικό σφάλμα σε ολόκληρο το *σύνολο επικύρωσης* (validation set). Για την επίτευξη αυτού του στόχου, χρησιμοποιούν τεχνικές σταδιακής μάθησης για SVMs [Tvei03], οι οποίες μπορούν να μοντελοποιήσουν την επίδραση των δειγμάτων εκπαίδευσης στο εκπαιδευμένο μοντέλο SVM.

Μια επίθεση δηλητηρίασης με βάση την παραπάνω διαδικασία είναι αρκετά επιτυχής για τα μοντέλα SVM. Ωστόσο, στα μοντέλα βαθιάς μάθησης, δεν είναι εύκολο να καταλάβουμε ρητά πως ακριβώς επιδρούν τα δείγματα εκπαίδευσης στο εκπαιδευμένο μοντέλο. Στη συνέχεια παρουσιάζουμε ορισμένες προσεγγίσεις σύμφωνα με τις οποίες εφαρμόζονται οι επιθέσεις δηλητηρίασης εξειδικευμένα πάνω σε μοντέλα ΒΝΔ.

Επίθεση σύμφωνα με το μοντέλο Koh

Στην εργασία [Koh20], γεννάται το ερώτημα πώς θα αλλάξουν οι προβλέψεις του μοντέλου εάν τροποποιηθεί ένα δείγμα εκπαίδευσης. Προκειμένου να δοθούν απαντήσεις στο παραπάνω ερώτημα, οι συγγραφείς του άρθρου εισάγουν μια μέθοδο για την ερμηνεία των βαθιών νευρωνικών δικτύων. Η μέθοδος αυτή μπορεί να ποσοτικοποιήσει ρητά την αλλαγή στο τελικό σφάλμα χωρίς να χρειαστεί να επανεκπαιδεύσει το μοντέλο στην περίπτωση που τροποποιείται μόνο ένα δείγμα εκπαίδευσης. Αυτή η διαδικασία μπορεί να υιοθετηθεί σε επιθέσεις δηλητηρίασης, βρίσκοντας εκείνα τα δείγματα εκπαίδευσης που έχουν μεγάλη επιρροή στην πρόβλεψη του μοντέλου.

Δηλητηριώδεις βατράχια (poison frogs)

Στην εργασία [Shaf18], παρουσιάζεται μια μέθοδος κατά την οποία μια αντιφατική εικόνα με την πραγματική της ετικέτα εισάγεται στο σύνολο δεδομένων εκπαίδευσης, προκειμένου να αναγκάσει το εκπαιδευμένο μοντέλο να ταξινομήσει εσφαλμένα ένα συγκεκριμένο δείγμα-στόχο. Σε αυτή τη μελέτη, δεδομένου ενός δείγματος-στόχου x_t με την πραγματική ετικέτα y_t, ο εισβολέας χρησιμοποιεί πρώτα ένα δείγμα x_b από την κλάση y_b ως βάση. Έπειτα, επιλύει την Εξίσωση 3.12 για να υπολογίσει το x'.

$$x' = \arg_x \min ||Z(x) - Z(x_t)||_2^2 + \beta||x - x_b||_2^2 \tag{3.12}$$

Μετά την εισαγωγή του δηλητηριασμένου δείγματος x στο σύνολο δεδομένων εκπαίδευσης, το νέο μοντέλο που εκπαιδεύτηκε στο $X_{train} + x'$ θα ταξινομήσει το x στην κλάση y_b, λόγω της μικρής απόστασης μεταξύ x' και x_b. Χρησιμοποιώντας ένα νέο εκπαιδευμένο μοντέλο για να προβλέψει την ταξινόμηση του x_t, το διάνυσμα βαθμολογίας x_t και x' εξαναγκάζονται μέσω της Εξίσωσης 3.12 να βρίσκονται σε κοντινή απόσταση. Συνεπώς, το αποτέλεσμα της πρόβλεψης του x' και του x_t θα είναι το ίδιο. Με αυτόν τον τρόπο, το νέο εκπαιδευμένο μοντέλο θα προβλέψει ότι το δείγμα-στόχος x_t ανήκει στην κλάση y_b.

Κεφάλαιο 4

Στρατηγικές αντιμετώπισης επιθέσεων

4.1 Ασφάλεια νευρωνικών δικτύων

Στο προηγούμενο Κεφάλαιο είδαμε αναλυτικά διάφορες κατηγορίες καθώς και πρακτικά παραδείγματα επιθέσεων που πραγματοποιούνται σε ΒΝΔ. Ο στόχος, τα κίνητρα και οι γνώσεις που διαθέτει ο εισβολέας γύρω από το μοντέλο που έχει επιλέξει να επιτεθεί, αποτελούν βασικούς παράγοντες για την οργάνωση και την εκτέλεση μιας επιτυχημένης επίθεσης. Επιπλέον, το ίδιο το μοντέλο-στόχος μπορεί να αποτελέσει κίνητρο ή αντικίνητρο για μια επίθεση, καθώς έχει παρατηρηθεί ότι ορισμένες δομές και αρχιτεκτονικές μοντέλων είναι περισσότερο ευάλωτες σε εισβολές που εμπεριέχουν επιθέσεις με εχθρικά παραδείγματα. Εφόσον πλέον έχουμε καλύψει το ζήτημα των επιθέσεων από την σκοπιά του εισβολέα αλλά και του ίδιου του μοντέλου-θύματος, το επόμενο ερώτημα που μας δημιουργείται είναι με ποιόν τρόπο θα μπορούσαμε να αξιολογήσουμε την ασφάλεια ενός μοντέλου-θύματος όταν αυτό έρχεται αντιμέτωπο με αντιφατικά παραδείγματα. Ποσοτικοποιώντας την έννοια της ασφάλειας κάθε μοντέλου διευκολύνεται το έργο της ισχυροποίησης της και κατά συνέπεια η διαδικασία αντιμετώπισης των εισβολών. Συνεπώς, κρίνεται απαραίτητη η εισαγωγή ορισμένων μεγεθών μέτρησης της ασφάλειας των νευρωνικών δικτύων καθώς και οι πιθανές σχέσεις ή διαφορές που παρουσιάζουν μεταξύ τους. Στη συνέχεια, θα παρουσιάσουμε ορισμένες χρήσιμες μετρικές που χρησιμοποιούνται για τον προσδιορισμό της ασφάλειας των δικτύων βαθιάς μηχανικής μάθησης.

4.1.1 Αξιολόγηση ασφάλειας νευρωνικών δικτύων

Το βασικότερο ίσως μέγεθος στον τομέα της ασφάλειας των ΒΝΔ αποτελεί η αντίσταση που παρουσιάζει ένα μοντέλο όταν δέχεται επίθεση από κακόβουλα αντιφατικά παραδείγματα. Όσο μεγαλύτερη αντίσταση παρουσιάζει ένα μοντέλο σε αντιφατικά παραδείγματα τόσο μεγαλύτερη ανθεκτικότητα έχει σε εχθρικές επιθέσεις. Οι δύο όροι που χρησιμοποιούνται ως επί το πλείστον για να περιγράψουν αυτήν την αντίσταση των μοντέλων ΒΝΔ είτε σε ένα μόνο δείγμα είτε στο συνολικό πληθυσμό αντίστοιχα είναι η *ευρωστία* (robustness) και το *αντιφατικό ρίσκο ή σφάλμα* (adversarial risk-loss). Οι δύο βασικές αυτές έννοιες θα ορισθούν πλήρως και θα αναλυθούν στη συνέχεια.

Ευρωστία

Ορισμός 1. *Ελάχιστη Διαταραχή (minimal perturbation):*

Δεδομένου ενός ταξινομητή F και του ζεύγους δεδομένων (x, y), η εχθρική διαταραχή θα έχει την ελάχιστη νόρμα - επιθυμώντας να πετύχουμε μια σχεδόν απαρατήρητη διαταραχή - όπως φαίνεται στην Εξίσωση 4.1:

$$\delta_{min} \; = \; \arg_x \min ||\delta|| \qquad s.\,t.\; F(x+\delta) \; \neq y \tag{4.1}$$

όπου με τον όρο $||\cdot||$ συνήθως αναφερόμαστε στην l_p νόρμα.

Ορισμός 2. *Ευρωστία (robustness)*

Η ευρωστία ορίζεται στη Εξίσωση 4.2 ως η νόρμα της ελάχιστης διαταραχής:

$$r(x, F) = ||\delta_{min}|| \tag{4.2}$$

Ορισμός 3. *Ολική ευρωστία (global robustness):*

Η ολική ευρωστία αποτελεί μια εκτίμηση της ευρωστίας σε ολόκληρο τον πληθυσμό D και ορίζεται μέσα από την Εξίσωση 4.3:

$$\rho(F) = \mathop{\mathbb{E}}_{x \sim D} r(x, F) \tag{4.3}$$

Χρησιμοποιώντας την ελάχιστη διαταραχή μπορούμε να κατασκευάσουμε αντιφατικό παράδειγμα που είναι όσο το δυνατόν πιο όμοιο με το x στο μοντέλο F που εξετάζουμε. Επομένως, όσο μεγαλύτερο είναι το $r(x, F)$ ή το $\rho(F)$, ο αντίπαλος πρέπει να θυσιάσει περισσότερη ομοιότητα για να δημιουργήσει αντιφατικά δείγματα. Με άλλα λόγια όταν οι τιμές των $r(x, F)$ και $\rho(F)$ είναι υψηλές, τα κακόβουλα δείγματα που δημιουργεί ο εισβολέας έχουν αναγκαστικά λιγότερη ομοιότητα με το αρχικό δείγμα x, γεγονός το οποίο δηλώνει ότι ο ταξινομητής F είναι πιο ανθεκτικός σε επιθέσεις, δηλαδή πιο ασφαλής.

Αντιφατικό σφάλμα

Ορισμός 4. *Το πιο αντιφατικό παράδειγμα (most adversarial example):*

Δεδομένου ενός ταξινομητή F και του δείγματος εισόδου x, το δείγμα x_adv με τη μεγαλύτερη τιμή σφάλματος στη γειτονιά του x δίνεται από την Εξίσωση 4.4:

$$x_{adv} = \arg_{x'} \max \mathcal{L}(x', F) \qquad s.\,t.\ ||x' - x|| \leq \epsilon \tag{4.4}$$

Ορισμός 5. *Αντιφατικό σφάλμα (adversarial loss):*

Ως αντιφατικό σφάλμα ορίζεται η τιμή σφάλματος του πιο αντιφατικού παραδείγματος, όπως φαίνεται στην Εξίσωση 4.5:

$$\mathcal{L}_{adv}(x) = \mathcal{L}(x_{adv}) = \max_{||x' - x|| < \epsilon} \mathcal{L}(\theta, x', y) \tag{4.5}$$

Ορισμός 6. *Ολικό αντιφατικό σφάλμα (global adversarial loss):*

Το ολικό αντιφατικό σφάλμα αποτελεί μια εκτίμηση του αντιφατικού σφάλματος πάνω στα δεδομένα της κατανομής D και ορίζεται από την Εξίσωση 4.6:

$$\mathcal{R}_{adv}(F) = \mathop{\mathbb{E}}_{x \sim D} \max_{||x' - x|| < \epsilon} \mathcal{L}(\theta, x', y) \tag{4.6}$$

Το πιο αντιφατικό παράδειγμα είναι το σημείο όπου το μοντέλο έχει τη μεγαλύτερη πιθανότητα να ξεγελαστεί μέσα στη γειτονιά του x. Μια χαμηλότερη τιμή σφάλματος $\mathcal{L}_{adv}$ υποδηλώνει ένα πιο στιβαρό και ανθεκτικό μοντέλο F.

Διαφορές μεταξύ αντιφατικού ρίσκου και ρίσκου

Ο ορισμός του αντιφατικού ρίσκου προέρχεται από τον ορισμό του ρίσκου του ταξινομητή (εμπειρικός κίνδυνος) που δίνεται από την Εξίσωση 4.7:

$$\mathcal{R}(F) = \mathop{\mathbb{E}}_{x \sim D} \mathcal{L}(\theta, x, y) \tag{4.7}$$

Το ρίσκο αφορά στη μελέτη της απόδοσης ενός ταξινομητή σε δείγματα από φυσική κατανομή D, σε αντίθεση με το αντιφατικό ρίσκο που όπως φαίνεται και από την Εξίσωση 4.6 αφορά στη μελέτη της απόδοσης ενός ταξινομητή σε αντιφατικό παράδειγμα x'. Σε αυτό το σημείο είναι σημαντικό να σημειωθεί ότι το x' δεν ακολουθεί απαραιτήτως την κατανομή D.

Υπάρχουν διάφορες μελέτες που αναφέρονται στη σχέση μεταξύ αυτών των δύο ιδιοτήτων [Tsip19, Su19a, Stut19, Zhan19b], ωστόσο ο στόχος μας στα πλαίσια αυτής της εργασίας είναι να υπάρξει σαφής διάκριση μεταξύ των δύο εννοιών και κυρίως να σημειωθεί η σημασία της μελέτης όλων των παραπάνω μετρικών.

4.2 Αντίμετρα κατά των αντιφατικών παραδειγμάτων

Προκειμένου να διασφαλιστεί η ασφάλεια των μοντέλων βαθιάς μηχανικής μάθησης, διαφορετικές στρατηγικές έχουν αναπτυχθεί και έχουν θεωρηθεί ως αντίμετρα έναντι των αντιφατικών παραδειγμάτων. Πιο συγκεκριμένα, υπάρχουν τρεις κύριες κατηγορίες αυτών των αντιμέτρων:

1. *Κάλυψη κλίσης* (gradient masking/obfuscation): Δεδομένου ότι οι περισσότεροι αλγόριθμοι επίθεσης - προκειμένου να διασφαλίσουν την επιτυχία τους - βασίζονται στις πληροφορίες που προκύπτουν από τον υπολογισμό του διανύσματος μερικών παραγόντων του ταξινομητή, η απόκρυψη της κλίσης της συνάρτησης κόστους θα μπορούσε να λειτουργήσει ως τροχοπέδη για τους αντιπάλους.

2. *Εύρωστη βελτιστοποίηση* (robust optimization): Η εκ νέου εκπαίδευση του δικτύου και εκμάθηση των παραμέτρων ενός ταξινομητή ΒΝΔ μπορεί να αυξήσει την ευρωστία του απέναντι σε κακόβουλες επιθέσεις. Ο εκπαιδευμένος για δεύτερη φορά ταξινομητής θα μπορεί πλέον να ταξινομήσει σωστά τα αντιφατικά παραδείγματα που μετέπειτα πιθανόν να δημιουργηθούν.

3. *Ανίχνευση αντιφατικών παραδειγμάτων* (Adversarial examples detection): Σε πρώτο στάδιο η συγκεκριμένη τεχνική αφορά στην ενδελεχή μελέτη των φυσικών καλοήθων παραδειγμάτων, έτσι ώστε σε δεύτερο στάδιο να είναι δυνατός ο εντοπισμός των αντιφατικών παραδειγμάτων και τελικώς να απαγορεύεται η ταξινόμησή τους.

4.2.1 Κάλυψη κλίσης

Η μέθοδος κάλυψης της κλίσης του ταξινομητή-στόχου αποτελεί μια αμυντική στρατηγική κατά την οποία ο αμυνόμενος κρύβει σκόπιμα τις πληροφορίες του διανύσματος μερικών παραγόντων του μοντέλου. Μέσω της τακτικής αυτής ο αμυνόμενος επιδιώκει να προκαλέσει σύγχυση στους αντιπάλους, καθώς όπως είδαμε αναλυτικά σε προηγούμενο Κεφάλαιο, οι περισσότεροι αλγόριθμοι επίθεσης βασίζονται στις πληροφορίες κλίσης της συνάρτησης κόστους του ταξινομητή.

Αμυντική απόσταξη

Ο όρος *απόσταξη* χρησιμοποιήθηκε για πρώτη φορά στο [Hint15] και αποτελεί μια τεχνική εκπαίδευσης δικτύου που στοχεύει στη μείωση του μεγέθους των αρχιτεκτονικών των ΒΝΔ. Για την επίτευξη του στόχου αυτού, απαιτείται η εκπαίδευση ενός μικρότερου σε μέγεθος μοντέλου ΒΝΔ πάνω στην έξοδο του τελευταίου επιπέδου πριν από το softmax. Στην εργασία [Pape16], οι συγγραφείς αναδιατυπώνουν τη διαδικασία της απόσταξης για να εκπαιδεύσουν ένα μοντέλο ΒΝΔ το οποίο δύναται να αντισταθεί σε επιθέσεις που χρησιμοποιούν αντιφατικά παραδείγματα, σε επιθέσεις δηλαδή τύπου FGSM, Szegedy's L-BFGS ή DeepFool. Η εκπαιδευτική διαδικασία της συγκεκριμένης τεχνικής αναλύεται στα εξής βήματα:

1. Αρχικά, εκπαιδεύουμε ένα δίκτυο F στο δοσμένο σύνολο δεδομένων εκπαιδευσης (X, Y) ρυθμίζοντας τη θερμοκρασία Boltzman της softmax σε T. Με την προσθήκη της υπερπαραμέτρου T, η συνάρτηση softmax παίρνει τη μορφή που φαίνεται στην Εξίσωση 4.8:

$$softmax(x, T)_i = \frac{e^{x_i/T}}{\sum_j e^{x_i/T}} \tag{4.8}$$

όπου $i = 0, 2, ..., K-1$

2. Στη συνέχεια, υπολογίζουμε ξανά τα αποτελέσματα που εξάγονται από το $F(X)$ αφού περάσουν από softmax και τα αξιολογούμε με βάση τη θερμοκρασία .

3. Έπειτα, χρησιμοποιώντας πάλι τη συνάρτηση ενεργοποίησης softmax με θερμοκρασία T, εκπαιδεύουμε ένα άλλο δίκτυο F'_T στο σύνολο δεδομένων $(X, F(X))$. Θα αναφερόμαστε πλέον στο μοντέλο F'_T με τον όρο *αποσταγμένο μοντέλο* (distilled model).

4. Τέλος, κατά τη διαδικασία πρόβλεψης του συνόλου δεδομένων ελέγχου X_test και κατά συνέπεια των αντιφατικών παραδειγμάτων, χρησιμοποιούμε το αποσταγμένο μοντέλο F'_T του προηγούμενου βήματος δίνοντας στη θερμοκρασία της softmax την τιμή 1.

Στην εργασία [Carl17b], οι συγγραφείς εξηγούν τον τρόπο με τον οποίο ο παραπάνω αλγόριθμος επιτυγχάνει τον σκοπό του. Χαρακτηριστικά αναφέρουν ότι, κατά τη διαδικασία της εκπαίδευσης του αποσταγμένου μοντέλου η τιμή της θερμοκρασίας Boltzman είναι T, ενώ στη φάση ελέγχου του δικτύου η τιμή της θερμοκρασίας είναι 1. Με αυτόν τον τρόπο, εξαναγκάζουμε τις εισόδους της softmax να μεγαλώσουν κατά T. Έστω για παράδειγμα ότι επιλέγουμε $T = 100$, οπότε οι έξοδοι του τελευταίου κρυφού επιπέδου $Z(\cdot)$ για το δείγμα x και τα γειτονικά του σημεία x' θα είναι 100 φορές μεγαλύτεροι. Η $F1(\cdot) = softmax(Z(\cdot), 1)$ θα εξάγει ως αποτέλεσμα ένα διάνυσμα της μορφής $(\epsilon, \epsilon, ..., 1 - (m-1)\epsilon, \epsilon, ..., \epsilon)$, όπου η τιμή του διανύσματος που αντιστοιχεί στην κλάση-στόχο θα βρίσκεται πολύ κοντά στο 1 ενώ όλες οι υπόλοιπες θα κυμαίνονται κοντά στο 0. Στην πράξη, η τιμή του ϵ είναι τόσο μικρή όπου η τυπική αναπαράστασή του στον υπολογιστή ως αριθμός κινητής υποδιαστολής 32 δυαδικών ψηφίων στρογγυλοποιείται στο 0. Συνεπώς, με αυτές τις τιμές εξόδου ο υπολογιστής δεν μπορεί να υπολογίσει το διάνυσμα μερικών παραγώγων, γεγονός το οποίο αναστέλλει τις επιθέσεις που βασίζονται στην κλίση.

«Θρυμματισμένες» κλίσεις

Ορισμένες μελέτες [Buck18, Guo18] προσπαθούν να προστατεύσουν το μοντέλο μέσω προεπεξεργασίας των δεδομένων εισόδου. Αρχικά, προσθέτουν έναν *μη ομαλό* ή *μη διαφορίσιμο* προεπεξεργαστή $g(\cdot)$ και στη συνέχεια εκπαιδεύουν ένα ΒΝΔ f στο $g(X)$. Ο εκπαιδευμένος ταξινομητής $f(g(\cdot))$ δεν είναι διαφορίσιμος στο x, προκαλώντας την αποτυχία των εχθρικών επιθέσεων.

Πιο συγκεκριμένα, η μέθοδος thermometer encoding [Buck18] χρησιμοποιεί έναν προεπεξεργαστή για να διακριτοποιήσει την τιμή pixel x_i μιας εικόνας μέσω ενός l-διαστάσεων διανύσματος (x_i). Το διάνυσμα (x_i) λειτουργεί ως «θερμόμετρο» για την καταγραφή της τιμής του pixel x_i. Για παράδειγμα, όταν $l = 10$, $(0, 66) = 1111110000$. Στη συνέχεια, ένα μοντέλο ΒΝΔ εκπαιδεύεται πάνω σε αυτά τα διανύσματα. Η εργασία [Guo18] μελετά διάφορα εργαλεία επεξεργασίας εικόνας, όπως περικοπή εικόνας, συμπίεση ή ελαχιστοποίηση συνολικής διακύμανσης, για να προσδιορίσει εάν αυτές οι τεχνικές βοηθούν στην προστασία του μοντέλου από αντιφατικά παραδείγματα. Όλες αυτές οι προσεγγίσεις εμποδίζουν την ομαλή σύνδεση μεταξύ της εξόδου του μοντέλου και των αρχικών δειγμάτων εισόδου, οπότε ο εισβολέας δεν μπορεί να βρει εύκολα το διάνυσμα μερικών παραγώγων $\frac{\partial F(x)}{\partial x}$ και να εξάγει τις απαραίτητες πληροφορίες για την επίθεση.

Στοχαστικά διανύσματα μερικών παραγώγων

Ορισμένες αμυντικές στρατηγικές προσπαθούν να προσδώσουν μια στοχαστικότητα και μια τυχαιότητα στο μοντέλο ΒΝΔ προκειμένου να μπερδέψουν τον αντίπαλο. Για παράδειγμα, μια χαρακτηριστική τεχνική αυτών των στρατηγικών είναι να εκπαιδεύσουμε για αρχή ένα σύνολο ταξινομητών $s = \{F_t : t = 1, 2, ..., k\}$. Κατά την αξιολόγηση των δεδομένων x, επιλέγουμε τυχαία έναν ταξινομητή από το σύνολο s και προβλέπουμε την κλάση y. Επειδή ο αντίπαλος δεν έχει ιδέα ποιος ταξινομητής χρησιμοποιείται από το μοντέλο πρόβλεψης, το ποσοστό επιτυχίας επίθεσης θα μειωθεί σημαντικά. Χαρακτηριστικά παραδείγματα αυτής της στρατηγικής αποτελούν η εργασία [Dhil18], σύμφωνα με την οποία σε κάθε επίπεδο του μοντέλου ΒΝΔ απενεργοποιούνται τυχαία ορισμένοι νευρώνες, και η εργασία [Xie18] που μεταβάλλει το μέγεθος των εικόνων εισόδου σε ένα τυχαίο μέγεθος και συμπληρώνει με μηδενικά γύρω από την εικόνα εισόδου.

Εξαφάνιση διανύσματος μερικών παραγώγων

Τόσο η μέθοδος PixelDefend [Song18a] όσο και η τεχνική Defense-GAN [Sama18] προτείνουν τη χρήση παραγωγικών μοντέλων για να προβάλουν ένα πιθανό αντιφατικό παράδειγμα στο σύνολο των καλοήθων δεδομένων προτού αυτά ταξινομηθούν. Το PixelDefend χρησιμοποιεί το παραγωγικό μοντέλο PixelCNN [Jawa19], ενώ το Defense-GAN χρησιμοποιεί μια GAN αρχιτεκτονική [Good14]. Τα παραγωγικά μοντέλα μπορούν να θεωρηθούν ως ένα είδος «καθαριστών» που μετατρέπουν τα κακόβουλα παραδείγματα σε καλοήθη.

Και οι δύο αυτές μέθοδοι προσθέτουν ένα παραγωγικό δίκτυο πριν από τον ταξινομητή και έτσι το τελικό μοντέλο ταξινόμησης προκύπτει να είναι ένα εξαιρετικά βαθύ νευρωνικό δίκτυο. Η επιτυχία των παραπάνω μεθόδων οφείλεται κυρίως στο γεγονός ότι το άθροισμα των μερικών παραγώγων από κάθε επίπεδο επηρεάζει τη συνολική κλίση του δικτύου $\frac{\partial \mathcal{L}(x)}{\partial x}$, η οποία αποκτά είτε εξαιρετικά μικρή τιμή είτε ακανόνιστα μεγάλη, εμποδίζοντας έτσι τον εισβολέα να εκτιμήσει με ακρίβεια τη θέση των αντιφατικών παραδειγμάτων.

Μέθοδοι κάλυψης της κλίσης

Η εργασία [Carl17b] δείχνει ότι η μέθοδος της αμυντικής απόσταξης εξακολουθεί να είναι ευάλωτη στα αντιφατικά παραδείγματα. Η εργασία [Atha18a] παρουσιάζει διαφορετικούς αλγόριθμους επίθεσης προκειμένου να σπάσει τις αμυντικές στρατηγικές που βασίζονται σε τεχνικές απόκρυψης κλίσης. Κύρια αδυναμία των στρατηγικών αυτών αποτελεί το γεγονός ότι οι μέθοδοι αυτοί δύνανται μόνο να «μπερδέψουν» τους αντιπάλους, αλλά δυστυχώς δεν μπορούν να εξαλείψουν την ύπαρξη αντιφατικών παραδειγμάτων.

4.2.2 Εύρωστη βελτιστοποίηση

Οι μέθοδοι εύρωστης βελτιστοποίησης στοχεύουν στην ενδυνάμωση της αντοχής του ταξινομητή σε επιθέσεις, αλλάζοντας τον τρόπο μάθησης του μοντέλου ΒΝΔ. Μελετούν πώς θα μπορούσαμε να μάθουμε εκείνες τις παραμέτρους του μοντέλου που παρουσιάζουν υποσχόμενες προβλέψεις για πιθανά αντιφατικά παραδείγματα. Σε αυτή την κατηγορία αντιμέτρων, οι επιστημονικές έρευνες εστιάζουν κυρίως στην εκμάθηση των παραμέτρων $\theta*$ του μοντέλου με στόχο:

1. την ελαχιστοποίηση του μέσου αντιφατικού σφάλματος, όπως ορίζεται στην Εξίσωση 4.9.

$$\theta* = \arg\min_{\theta\in\Theta} \mathbb{E}_{x\sim D} \max_{||x'-x||\leq\epsilon} \mathcal{L}(\theta, x', y) \tag{4.9}$$

2. τη μεγιστοποίηση της μέσης *ελάχιστης διαταραχής* (minimal perturbation), όπως ορίζεται στην Εξίσωση 4.10.

$$\theta* = \arg\max_{\theta\in\Theta} \mathbb{E}_{x\sim D} \min_{C(x')\neq y} ||x'-x|| \tag{4.10}$$

Τυπικά, ένας αλγόριθμος εύρωστης βελτιστοποίησης γνωρίζει εκ των προτέρων την πιθανή απειλή ή την πιθανή επίθεση που πρόκειται να δεχθεί (αντιφατικός χώρος D). Με αυτόν τον τρόπο, οι αμυνόμενοι δημιουργούν ταξινομητές που είναι ασφαλείς έναντι της συγκεκριμένης επίθεσης. Οι περισσότερες σχετικές έρευνες [Good15, Kura17a, Madr19] στοχεύουν στην υπεράσπιση μοντέλων ΒΝΔ στις περιπτώσεις που δέχονται επιθέσεις μέσω αντιφατικών παραδειγμάτων, τα οποία έχουν δημιουργηθεί υπό τον περιορισμό μικρής l_p νόρμας (συγκεκριμένα l_∞ και l_2). Παρόλο που υπάρχει πιθανότητα οι άμυνες αυτές να είναι ευάλωτες σε επιθέσεις από άλλους μηχανισμούς [Xiao18], η μελέτη της ασφάλειας κατά της επίθεσης l_p είναι θεμελιώδης και μπορεί να γενικευτεί σε άλλες επιθέσεις.

Σε αυτήν την ενότητα, επικεντρωνόμαστε σε αμυντικές προσεγγίσεις που στοχεύουν στην αντιμετώπιση l_p επιθέσεων χρησιμοποιώντας αλγορίθμους εύρωστης βελτιστοποίησης. Για δική μας διευκόλυνση και καλύτερη κατανόηση, διαχωρίζουμε την υπάρχουσα σχετική έρευνα σε τρεις κατηγορίες: (i) στις *μεθόδους κανονικοποίησης* (regularization methods), (ii) στην *αντιφατική (εκ νέου) εκπαίδευση* (adversarial (re)training) και (iii) στις *πιστοποιημένες άμυνες* (certified defenses).

Μέθοδοι κανονικοποίησης

Κάποιες πρώιμες μελέτες για την αντιμετώπιση των επιθέσεων που χρησιμοποιούν αντιφατικά παραδείγματα επικεντρώνονται στη διερεύνηση ορισμένων ιδιοτήτων που πρέπει να έχει ένα ισχυρό ΒΝΔ προκειμένου να αντισταθεί σε τέτοιους είδους επιθέσεις. Για παράδειγμα, στην εργασία [Szeg14], προτείνεται ότι ένα εύρωστο μοντέλο θα πρέπει να είναι σταθερό ακόμη και όταν οι είσοδοί του παραμορφώνονται. Οι συγγραφείς της παραπάνω εργασίας επιτυγχάνουν την επιβολή αυτής της «σταθερότητας» της εξόδου του μοντέλου μέσω του περιορισμού της σταθερά Lipschitz. Η εκπαίδευση του δικτύου υπό αυτές τις κανονικοποιήσεις δύναται να βοηθήσει το μοντέλο να παρουσιάζει μεγαλύτερη ευρωστία.

1. *Επιβολή κυρώσεων στα επίπεδα εξαιτίας της σταθεράς Lipschitz*: Όταν μελετήθηκε για πρώτη φορά το ζήτημα του πόσο ευάλωτα εμφανίζονται τα μοντέλα ΒΝΔ απέναντι σε επιθέσεις αντιφατικών παραδειγμάτων, οι συγγραφείς του [Szeg14] πρότειναν ότι η προσθήκη όρων κανονικοποίησης στις παραμέτρους του δικτύου κατά τη διάρκεια της εκπαίδευσης μπορεί να αναγκάσει το εκπαιδευμένο μοντέλο να γίνει πιο σταθερό. Πιο συγκεκριμένα, εισήγαγαν τον περιορισμό της σταθεράς Lipschitz L_k μεταξύ οποιωνδήποτε δύο επιπέδων, όπως δίνεται στην Εξίσωση 4.11:
$$\forall x,\ \delta,\ ||h_k(x;W_k) - h_k(x\ +\ \delta;W_k)||\ \leq\ L_k||\delta|| \tag{4.11}$$
έτσι ώστε το αποτέλεσμα κάθε επιπέδου να μην επηρεάζεται εύκολα από τη μικρή πιθανή παραμόρφωση της εισόδου του. Οι συγγραφείς της εργασίας [Ciss17] επισημοποίησαν την παραπάνω ιδέα, ισχυριζόμενοι ότι το αντιφατικό σφάλμα του μοντέλου, που ορίστηκε στην Εξίσωση 4.9, εξαρτάται απευθείας από αυτή την αστάθεια L_k, όπως αποδεικνύεται από την Εξίσωση 4.12:
$$\begin{aligned} &\mathop{\mathbb{E}}_{x\sim D} \mathcal{L}_{adv}(x) \leqslant \mathop{\mathbb{E}}_{x\sim D} \mathcal{L}(x) \\ &+ \mathop{\mathbb{E}}_{x\sim D}\Big[\max_{||x'-x||\leqslant\epsilon} |\mathcal{L}(F(x'),y) - \mathcal{L}(F(x),y)|\Big] \\ &\leqslant \mathop{\mathbb{E}}_{x\sim D} \mathcal{L}(x) + \lambda_p \prod_{k=1}^{K} L_k \end{aligned} \tag{4.12}$$
όπου λ_p είναι η σταθερά Lipschitz της συνάρτησης κόστους. Αυτή η φόρμουλα που διατυπώνεται στην Εξίσωση 4.12 δηλώνει ότι κατά τη διάρκεια της εκπαίδευσης του δικτύου, η επιβολή κυρώσεων σε κάθε κρυφό επίπεδο εξαιτίας της μεγάλης αστάθειας μπορεί να συμβάλει στη μείωση του αντιφατικού σφάλματος του μοντέλου και κατά συνέπεια, στην αύξηση της, που αφορά σε άμυνες χωρίς εποπτεία ή με μερική εποπτεία [Miya16].

2. *Επιβολή κυρώσεων στις μερικές παραγώγου των επιπέδων*: Η εργασία [Gu15] εισήγαγε τον αλγόριθμο Deep Contractive Network για την κανονικοποίηση της διαδικασίας της εκπαίδευσης. Η μέθοδος αυτή προτείνει την προσθήκη ποινής στις μερικές παραγώγους του μοντέλου σε κάθε επίπεδο κατά τη διαδικασία εκτέλεσης του αλγορίθμου της προς τα πίσω διάδοσης, έτσι ώστε μια αλλαγή των δεδομένων εισόδου να μην προκαλεί μεγάλη αλλαγή στην έξοδο κάθε επιπέδου. Συνεπώς με αυτόν τον τρόπο, καθίσταται πλέον πιο δύσκολο για τον ταξινομητή να δώσει εσφαλμένες προβλέψεις στα δείγματα που περιέχουν κάποια διαταραχή.

Αντιφατική (εκ νέου) εκπαίδευση με FGSM

Η εργασία [Good15] είναι η πρώτη που προτείνει την εισαγωγή παραγόμενων αντιφατικών παραδειγμάτων στη διαδικασία της εκπαίδευσης. Προσθέτοντας τα αντιφατικά παραδείγματα με την αληθινή τους ετικέτα (x', y) στο σύνολο δεδομένων εκπαίδευσης, στην ουσία εκπαιδεύουμε τον ταξινομητή να κατηγοριοποιεί το δείγμα x' στην κλάση y, έτσι ώστε το μετέπειτα εκπαιδευμένο μοντέλο να προβλέπει σωστά την ετικέτα παρόμοιων μελλοντικών αντιφατικών παραδειγμάτων.

Οι συγγραφείς της παραπάνω μελέτης χρησιμοποιούν μη στοχευμένη FGSM επίθεση (Ενότητα 3.2.1) προκειμένου να δημιουργήσουν αντιφατικά παραδείγματα x' για το σύνολο δεδομένων εκπαίδευσης, όπως φαίνεται και από την Εξίσωση 4.13.

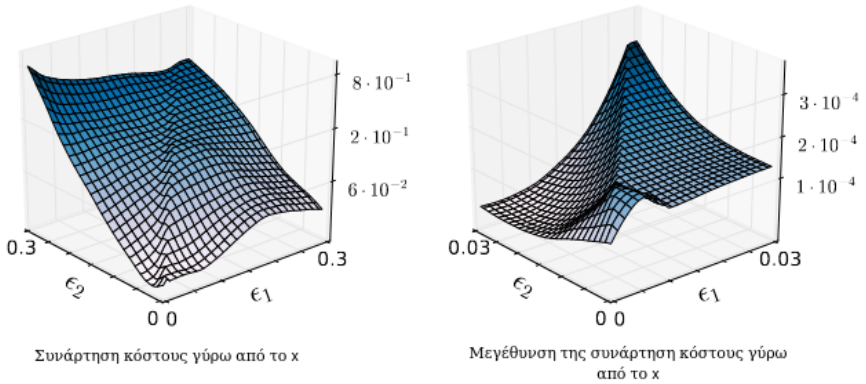
$$x' = x + \epsilon \text{sgn}(\nabla_x \mathcal{L}((\theta,\ x,\ y)) \tag{4.13}$$

Εκπαιδεύοντας το μοντέλο σε καλοήθη δείγματα επαυξημένα με αντιφατικά παραδείγματα, αυξάνεται η ανθεκτικότητα του δικτύου έναντι κακόβουλων δειγμάτων που δημιουργούνται από την FGSM μέθοδο. Η στρατηγική εκπαίδευσης αυτής της μεθόδου αλλάζει στην εργασία [Kura17a] έτσι ώστε να είναι δυνατή η κλιμάκωση του μοντέλου σε μεγαλύτερα σύνολα δεδομένων, όπως το ImageNet [onlid]. Οι συγγραφείς αυτής της εργασίας προτείνουν ότι η εφαρμογή κανονικοποίησης ανά δέσμες θα βελτιώσει την αποτελεσματικότητα της αντιφατικής εκπαίδευσης [Ioff15]. Στον Αλγόριθμο 1 παρουσιάζουμε τα βασικά σημεία αυτής της μεθόδου.

Algorithm 1: Adversarial Training with FGSM by batches

Randomly initialize network F
repeat
1. Read minibatch $B = \{x^1, ..., x^m\}$ from training set.
2. Generate k adversarial examples $\{x^1_{adv}, ...x^k_{adv}\}$ for corresponding benign examples
using current state of the network F.
3.Update $B' = \{x^1_{adv}, ..., x^k_{adv}, x^{k+1}, ..., x^m\}$.
Do one training step of network F using minibatch B'
until training converged

Ο εκπαιδευμένος ταξινομητής παρουσιάζει υψηλή αντοχή στις επιθέσεις FGSM, αλλά εξακολουθεί να είναι ευάλωτος σε επιθέσεις που χρησιμοποιούν επαναληπτικές διαδικασίες. Η εργασία [Tram20] έρχεται να καταρρίψει το παραπάνω συμπέρασμα υποστηρίζοντας ότι η άμυνα αυτή είναι επίσης ευάλωτη σε επιθέσεις με ένα βήμα. Σε κάθε περίπτωση, η αντιφατική εκπαίδευση με FGSM δύναται να προκαλέσει απόκρυψη της κλίσης, καθώς ο εκπαιδευμένος ταξινομητής F παρουσιάζει έντονη μη ομαλότητα κοντά στο δείγμα ελέγχου x, όπως φαίνεται και στο Σχήμα 4.1.



Σχήμα 4.1: Απεικόνιση της απόκρυψης κλίσης στην αντιφατική εκπαίδευση με FGSM (**Πηγή**:[Tram20]).

Αντιφατική εκπαίδευση με PGD

Στην εργασία [Madr19] προτείνεται η χρήση της επίθεσης PGD για την αντιφατική εκπαίδευση, αντί της χρήσης επιθέσεων ενός βήματος όπως η FGSM. Οι επιθέσεις PGD μπορούν να θεωρηθούν ως μια ευριστική μέθοδος για την εύρεση του πιο αντιφατικού παραδείγματος στην περιορισμένη από

την l_∞ σφαίρα γύρω από το x, $B_\epsilon(x)$ και συνοψίζονται μέσα από την Εξίσωση 4.14:

$$x_{adv} = \arg\max_{x' \in B\epsilon(x)} \mathcal{L}(x', F) \tag{4.14}$$

Εδώ, το πιο αντιφατικό παράδειγμα x_{adv} αποτελεί το σημείο εκείνο όπου ο ταξινομητής F έχει τη μεγαλύτερη πιθανότητα να παραπλανηθεί. Όταν το μοντέλο ΒΝΔ εκπαιδεύεται σε αυτά τα πιο αντιφατικά παραδείγματα, στην πραγματικότητα επιλύεται το πρόβλημα εκμάθησης των παραμέτρων θ του μοντέλου που ελαχιστοποιούν το αντιφατικό σφάλμα, όπως έχει οριστεί στην Εξίσωση 4.9. Εάν το εκπαιδευμένο μοντέλο έχει μικρή τιμή σφάλματος σε αυτά τα πιο αντιφατικά παραδείγματα, τότε είναι ασφαλές παντού στη γειτονιά του x που ορίζεται από τη σφαίρα $B_\epsilon(x)$. Επισημαίνεται ότι, η μέθοδος αυτή εκπαιδεύει το μοντέλο μόνο σε αντιφατικά παραδείγματα, και όχι στο συνδυασμό καλοήθων και κακόβουλων παραδειγμάτων που είδαμε σε προηγούμενη παράγραφο. Ο αλγόριθμος εκπαίδευσης παρουσιάζεται στον Αλγόριθμο 2.

Algorithm 2: Adversarial Training with PGD

Randomly initialize network F

repeat

1. Read minibatch $B = \{x^1, ..., x^m\}$ from training set.
2. Generate m adversarial examples $\{x^1_{adv}, ...x^m_{adv}\}$ by PGD attack using current state of the network F.
3. Update $B' = \{x^1_{adv}, ..., x^m_{adv}\}$.

Do one training step of network F using minibatch B'

until training converged

Το εκπαιδευμένο μοντέλο βάσει αυτής της μεθόδου παρουσιάζει υψηλή ανθεκτικότητα απέναντι σε επιθέσεις ενός βήματος αλλά και σε επιθέσεις που χρησιμοποιούν επαναληπτικές μεθόδους πάνω στα σύνολα δεδομένων MNIST [LeCu] και CIFAR10 [onlib]. Ωστόσο, η μέθοδος αυτή περιλαμβάνει μια επαναληπτική επίθεση για κάθε δείγμα του συνόλου δεδομένων εκπαίδευσης. Λόγω αυτού, η χρονική πολυπλοκότητα αυτής της αντιφατικής εκπαίδευσης θα είναι - χρησιμοποιώντας k-βήματα στη PGD μέθοδο - k φορές μεγαλύτερη από την φυσική εκπαίδευση και κατά συνέπεια η κλιμάκωση αυτής της μεθόδου σε μεγάλα σύνολα δεδομένων όπως το ImageNet [onlid] καθίσταται εξαιρετικά δύσκολη.

Συνδυαστική αντιφατική εκπαίδευσης

Η εργασία [Tram20] εισάγει μια μέθοδο αντιφατικής εκπαίδευσης, η οποία μπορεί να προστατεύσει μοντέλα συνελικτικών νευρωνικών δικτύων από επιθέσεις ενός βήματος και μπορεί επίσης να εφαρμοστεί σε μεγάλα σύνολα δεδομένων όπως το ImageNet [onlid]. Η κεντρική ιδέα της μεθόδου αυτής βασίζεται στην αύξηση του συνόλου δεδομένων εκπαίδευσης του ταξινομητή μέσω της προσθήκης κατασκευασμένων αντιφατικών παραδειγμάτων που έχουν δημιουργηθεί μέσα από άλλους προεκπαιδευμένους ταξινομητές. Για παράδειγμα, εάν σκοπεύουμε να εκπαιδεύσουμε έναν εύρωστο ταξινομητή F, μπορούμε πρώτα να προεκπαιδεύσουμε τους ταξινομητές F_1, F_2 και F_3 ως βάσεις. Αυτοί οι ταξινομητές θα έχουν διαφορετικές υπερπαραμέτρους από το μοντέλο F. Στη συνέχεια, για κάθε δείγμα x, χρησιμοποιούμε την επίθεση ενός βήματος FGSM για να δημιουργήσουμε αντιφατικά παραδείγματα για τα μοντέλα F_1, F_2 και F_3 και να πάρουμε αντίστοιχα x^1_{adv}, x^2_{adv}, x^3_{adv}. Χάρη στην ιδιότητα μεταφοράς μεταξύ διαφορετικών μοντέλων που παρουσιάζουν οι επιθέσεις ενός βήματος, τα δείγματα x^1_{adv}, x^2_{adv}, x^3_{adv} είναι πιθανό να παραπλανήσουν και τον ταξινομητή F. Αυτό σημαίνει ότι τα κατασκευασμένα αυτά αντιφατικά παραδείγματα αποτελούν μια καλή προσέγγιση του πιο αντιφατικού παραδείγματος για το μοντέλο F στο x, που ορίστηκε μέσα από την Εξίσωση 4.14. Η εκπαίδευση του μοντέλου πάνω στο συνδυασμό αυτών των δειγμάτων θα ελαχιστοποιήσει το αντιφατικό σφάλμα, που δίνεται από την Εξίσωση 4.9.

Αυτός ο αλγόριθμος *συνδυαστικής αντιφατικής εκπαίδευσης* (ensemble adversarial training) είναι πιο αποτελεσματικός από αυτούς που είδαμε στις δύο προηγούμενες παραγράφους, καθώς αποσυνδέει τη διαδικασία εκπαίδευσης των μοντέλων από τη δημιουργία των αντιφατικών παραδειγμάτων.

Επιτάχυνση της αντιφατικής εκπαίδευσης (Accelerate adversarial training)

Η αντιφατική εκπαίδευση με PGD που αναλύσαμε προηγουμένως, αν και πρόκειται για μια από τις πιο υποσχόμενες και αξιόπιστες στρατηγικές άμυνας, στη γενική περίπτωση είναι αργή μέθοδος με μεγάλο υπολογιστικό κόστος. Στην εργασία [Shaf19], προτείνεται ένας αλγόριθμος αντιφατικής εκπαίδευσης ο οποίος βελτιώνει την αποτελεσματικότητά του επαναχρησιμοποιώντας τους υπολογισμούς των πίσω περασμάτων. Σε αυτόν τον αλγόριθμο, το διάνυσμα μερικών παραγώγων του σφάλματος στο δείγμα εισόδου $\frac{\partial \mathcal{L}(x+\delta,\theta)}{\partial x}$ και το διάνυσμα μερικών παραγώγων του σφάλματος στις παραμέτρους του μοντέλου $\frac{\partial \mathcal{L}(x+\delta,\theta)}{\partial \theta}$ μπορούν να υπολογιστούν μαζί σε ένα βήμα οπισθοδιάδοσης, μέσω κοινής χρήσης των ίδιων όρων του κανόνα αλυσίδας. Ο αλγόριθμος αυτός αντιφατικής εκπαίδευσης παρουσιάζεται στον Αλγόριθμο 3.

Algorithm 3: Free Adversarial Training

Randomly initialize network F

repeat

1. Read minibatch $B = \{x^1, ..., x^m\}$ from training set.
2. for $i = 1...m$ do
   (a) Update model parameter θ
   $g_\theta \leftarrow \mathop{\mathbb{E}}_{(x,y)\in B}[\nabla_\theta \mathcal{L}((x_{+\delta}, y, \theta)]$
   $g_{adv} \leftarrow \nabla_x \mathcal{L}((x_{+\delta}, y, \theta)$
   $\theta \leftarrow \theta - \alpha g_\theta$
   (b) Generate adversarial examples
   $\delta \leftarrow \delta + \epsilon sign(g_{adv})$
   $\delta \leftarrow clip(\delta, -\epsilon, \epsilon)$
3. Update minibatch B with adversarial examples $x + \delta$.

until training converged

Στην εργασία [Zhan19a] υποστηρίζεται ότι όταν οι παράμετροι του μοντέλου είναι σταθερές, το αντιφατικό παράδειγμα που δημιουργείται μέσω της μεθόδου PGD συνδυάζεται μόνο με τα βάρη του πρώτου επιπέδου του ΒΝΔ. Στην έρευνα αυτή αναπτύσσεται ο αλγόριθμος You Only Propagate Once (YOPO), ο οποίος επαναχρησιμοποιεί την κλίση του σφάλματος στην έξοδο του πρώτου επιπέδου του μοντέλου $\frac{\partial \mathcal{L}(x+\delta,\theta)}{\partial Z_1(x)}$ κατά τη διάρκεια των PGD επιθέσεων. Με αυτόν τον τρόπο, το YOPO αποφεύγει πολλές φορές τον υπολογισμό της κλίσης και συνεπώς μειώνει το υπολογιστικό κόστος.

Ευαπόδεικτες άμυνες

Η αντιφατική εκπαίδευση έχει αποδειχθεί αποτελεσματική στην προστασία των μοντέλων έναντι αντιφατικών παραδειγμάτων. Βέβαια, η μέθοδος αυτή εξακολουθεί να μην αποτελεί εγγύηση για την ασφάλεια των εκπαιδευμένων ταξινομητών. Ποτέ δεν θα ξέρουμε εάν υπάρχουν πιο ισχυρές επιθέσεις που μπορούν να σπάσουν αυτές τις άμυνες, οπότε η άμεση εφαρμογή αυτών των αλγορίθμων αντιφατικής εκπαίδευσης σε εργασίες κρίσιμες ως προς την ασφάλεια δεν αποτελεί ρεαλιστική επιλογή.

Όπως αναφέραμε σε προηγούμενη ενότητα, η εργασία [Carl18] ήταν η πρώτη που εισήγαγε τον αλγόριθμο Reluplex για να εξετάσει την ευρωστία των μοντέλων ΒΝΔ. Σύμφωνα με αυτή την έρευνα, όταν δίνεται ένα μοντέλο F, ο αλγόριθμος υπολογίζει την ακριβή τιμή της απόστασης της ελάχιστης διαταραχής $r(x; F)$. Με άλλα λόγια, ο ταξινομητής θεωρείται ασφαλής έναντι οποιασδήποτε διαταραχής με νόρμα μικρότερη από $r(x; F)$. Εάν εφαρμόσουμε τον αλγόριθμο Reluplex σε ολόκληρο το σύνολο δεδομένων ελέγχου, μπορούμε να εξάγουμε το ποσοστό των δειγμάτων που είναι απολύτως ασφαλές έναντι διαταραχών με νόρμα μικρότερη από r_0. Με αυτόν τον τρόπο, κερδίζουμε εμπιστοσύνη και μειώνουμε τον αναμενόμενο κίνδυνο κατά την κατασκευή μοντέλων ΒΝΔ.

Η μέθοδος Reluplex - όπως επισημάναμε προηγουμένως - επιδιώκει να βρει την ακριβή τιμή του $r(x; F)$ που μπορεί να επαληθεύσει την αντοχή του μοντέλου F στο x. Αντίθετα, εργασίες όπως οι [Ragh20, Wong18a, Hein17] προσπαθούν να βρουν εκπαιδεύσιμα «πιστοποιητικά» $C(x; F)$ για να επαληθεύσουν την ευρωστία του μοντέλου. Για παράδειγμα στην εργασία [Hein17], ένα πιστοποι-

ητικό $C(x, F)$ υπολογίζεται για το μοντέλο F στο x. Το πιστοποιητικό αυτό είναι στην ουσία ένα κατώτερο όριο της απόστασης της ελάχιστης διαταραχής: $C(x, F) \leqslant r(x, F)$. Επομένως, το μοντέλο πρέπει να είναι ασφαλές έναντι οποιασδήποτε διαταραχής με νόρμα που περιορίζεται από το $C(x, F)$. Επιπλέον, αυτά τα πιστοποιητικά δύνανται να εκπαιδευτούν, πράγμα το οποίο σημαίνει ότι η εκπαίδευση για τη βελτιστοποίηση τους θα προσφέρει καλύτερη ευρωστία στον ταξινομητή. Σε αυτήν την ενότητα, θα παρουσιάσουμε εν συντομία ορισμένες μεθόδους για το σχεδιασμό αυτών των πιστοποιητικών.

1. *Κατώτερο όριο ελάχιστης διαταραχής*: Η εργασία [Hein17] προτείνει ένα κατώτερο όριο $C(x, F)$ για την απόσταση της ελάχιστης διαταραχής του F στο x κάνοντας χρήση του θεωρήματος Cross-Lipschitz και ορίζεται από την Εξίσωση 4.15]:

$$\max_{\epsilon>0} \min \left\{ \min_{i \neq y} \frac{Z_y(x) - Z_i(x)}{\max_{x' \in B\epsilon(x)} ||\nabla Z_y(x') - \nabla Z_i(x')||} , \epsilon \right\} \tag{4.15}$$

   Αξίζει να σημειωθεί ότι η διατύπωση του $C(x, F)$ εξαρτάται μόνο από τα F και x, και είναι εύκολο να υπολογιστεί για ένα νευρωνικό δίκτυο με ένα μόνο κρυφό επίπεδο. Έτσι, το μοντέλο F μπορεί να αποδειχθεί ασφαλές στην περιοχή που ορίζεται εντός της απόστασης $C(x, F)$. Η εκπαίδευση για τη μεγιστοποίηση αυτού του κατώτερου ορίου θα κάνει τον ταξινομητή πιο εύρωστο.

2. *Άνω όριο αντιφατικού σφάλματος*: Οι εργασίες [Ragh20, Wong18a] προσπαθούν να βρουν ένα άνω όριο $\mathcal{U}(x, F)$ που είναι μεγαλύτερο από το αντιφατικό σφάλμα $\mathcal{L}_{adv}(x, F)$ και δίνεται από την Εξίσωση 4.16:

$$\begin{aligned} \mathcal{L}_{adv}(x) = & \max_{x'} \left\{ \max_{i \neq y} Z_i(x') - Z_y(x') \right\} \\ & s.\,t.\ x' \in B_\epsilon(x) \end{aligned} \tag{4.16}$$

   Το πιστοποιητικό $\mathcal{U}(x, F)$ λειτουργεί με τον εξής τρόπο: εάν $\mathcal{U}(x, F) < 0$, τότε το αντιφατικό σφάλμα είναι $\mathcal{L}(x, F) < 0$. Επομένως, ο ταξινομητής δίνει πάντα τη μεγαλύτερη βαθμολογία στην πραγματική κλάση y εντός της σφαίρας $B_\epsilon(x)$ και έτσι το μοντέλο είναι ασφαλές σε αυτήν την περιοχή. Για να αυξήσουμε την ευρωστία του μοντέλου, πρέπει να μάθουμε τις παραμέτρους που έχουν τη μικρότερη τιμή $\mathcal{U}$, έτσι ώστε όλο και περισσότερα δείγματα να έχουν αποκτούν αρνητικές τιμές $\mathcal{U}$.

   Η εργασία [Ragh20] χρησιμοποιεί *ημι-ορισμένο προγραμματισμό* (semi-definite programming) [Vand96] για την επίλυση του πιστοποιητικού. Αντιθέτως, η εργασία [Wong18a] μετατρέπει το πρόβλημα της Εξίσωσης 4.16 σε πρόβλημα γραμμικού προγραμματισμού και το επιλύει μέσω της εκπαίδευσης ενός εναλλακτικού νευρωνικού δικτύου. Και οι δύο μέθοδοι εξετάζουν μόνο τα νευρωνικά δίκτυα με ένα κρυφό επίπεδο. Υπάρχουν επίσης μελέτες [Ragh18, Wong18b] που βελτιώνουν την αποτελεσματικότητα και την κλιμάκωση των αλγορίθμων αυτών.

   Επιπλέον, στην εργασία [Sinh20] οι συγγραφείς συνδυάζουν την αντιφατική εκπαίδευση μαζί με την ευαπόδεικτη άμυνα. Πιο συγκεκριμένα, εκπαιδεύουν τον ταξινομητή τροφοδοτώντας τον με αντιφατικά παραδείγματα τα οποία έχουν δειγματοληπτηθεί από την κατανομή της χειρότερης περίπτωσης διαταραχής και εξάγουν τα πιστοποιητικά μελετώντας τη δυαδικότητα της μεθόδου Lagrange πάνω στο αντιφατικό σφάλμα.

4.2.3 Ανίχνευση αντιφατικών παραδειγμάτων

Η ανίχνευση αντιφατικών παραδειγμάτων αποτελεί άλλη μια κύρια προσέγγιση για την προστασία των ταξινομητών των ΒΝΔ. Αντί να προβλέπουν απευθείας την είσοδο του μοντέλου, αυτές οι μέθοδοι διακρίνουν πρώτα εάν η είσοδος είναι καλοήθης ή κακόβουλη. Στη συνέχεια, εάν εντοπιστεί αντιφατική είσοδος, ο ταξινομητής ΒΝΔ αρνείται να προβλέψει την ετικέτα του δείγματος. Στην εργασία [Carl17a], τα μοντέλα-απειλές κατατάσσονται σε 3 κατηγορίες τις οποίες καλούνται να αντιμετωπίσουν οι τεχνικές ανίχνευσης:

1. Στην κατηγορία *μηδενικής γνώσης του αντιπάλου* (zero-knowledge adversary), ο αντίπαλος έχει πρόσβαση μόνο στις παραμέτρους του ταξινομητή F και δεν γνωρίζει το μοντέλο ανίχνευσης D.

2. Στην κατηγορία *άριστης γνώσης του αντιπάλου* (perfect-knowledge adversary), ο αντίπαλος γνωρίζει, πέρα από το μοντέλο F, το σχήμα ανίχνευσης D και τις παραμέτρους του.

3. Στην κατηγορία *περιορισμένης γνώσης του αντιπάλου* (limited-knowledge adversary), ο εισβολέας γνωρίζει το μοντέλο F και το σχήμα ανίχνευσης D, αλλά δεν έχει πρόσβαση στις παραμέτρους του D. Δηλαδή, αυτός ο αντίπαλος δεν γνωρίζει το σύνολο δεδομένων εκπαίδευσης του μοντέλου.

Σε όλες τις παραπάνω κατηγορίες, απαιτούνται εργαλεία ανίχνευσης για τη σωστή κατηγοριοποίηση των αντιφατικών παραδειγμάτων, τα οποία έχουν χαμηλή πιθανότητα εσφαλμένης ταξινόμησης καλοήθων παραδειγμάτων. Στη συνέχεια, θα εξετάσουμε ορισμένες βασικές μεθόδους για την ανίχνευση αντιφατικών παραδειγμάτων.

Βοηθητικό μοντέλο για την κατηγοριοποίηση αντιφατικών παρδειγμάτων

Ορισμένες εργασίες επικεντρώνονται στο σχεδιασμό βοηθητικών μοντέλων που στοχεύουν στη διάκριση μεταξύ κακόβουλων και καλοήθων παραδείγματα. Στην εργασία [Gros17] οι συγγραφείς εκπαιδεύουν ένα μοντέλο ΒΝΔ με $|Y| = K + 1$ ετικέτες, με μια πρόσθετη ετικέτα για όλα τα αντιφατικά παραδείγματα, έτσι ώστε το δίκτυο να κατηγοριοποιεί κάθε αντιφατικό παράδειγμα στην κλάση $K + 1$. Παρομοίως, στην εργασία [Gong17] αρχικά εκπαιδεύεται ένα μοντέλο δυαδικής ταξινόμησης προκειμένου να διακρίνει όλα τα αντιφατικά από τα καλοήθη παραδείγματα και στη συνέχεια ένας ταξινομητής εκπαιδεύεται στα αναγνωρισμένα καλοήθη δείγματα. Η εργασία [Metz17] προτείνει μια μέθοδο ανίχνευσης κατασκευάζοντας ένα βοηθητικό νευρωνικό δίκτυο D, το οποίο λαμβάνει εισόδους από τις τιμές των κρυφών κόμβων H του φυσικού εκπαιδευμένου ταξινομητή. Ο εκπαιδευμένος ταξινομητής ανίχνευσης $D : H \rightarrow [0, 1]$ είναι ένα μοντέλο δυαδικής ταξινόμησης που διακρίνει τα κακόβουλα παραδείγματα από τα καλοήθη, αξιοποιώντας πληροφορίες που εξάγει από τα κρυμμένα επίπεδα του δικτύου.

Χρήση στατιστικών για τη διάκριση των αντιφατικών παραδειγμάτων

Ορισμένες μελέτες ερευνούν τις διαφορές που παρουσιάζουν στις στατιστικές ιδιότητές τους τα αντιφατικά από τα καλοήθη παραδείγματα. Για παράδειγμα, στην εργασία [Hend17], παρατηρήθηκε ότι τα αντιφατικά παραδείγματα παρουσιάζουν υψηλότερο βάρος στις μεγαλύτερες κύριες συνιστώσες, σε αντίθεση με τις φυσικές εικόνες οι οποίες παρουσιάζουν μεγαλύτερο βάρος στις μικρότερες κύριες συνιστώσες. Συνεπώς, τα δείγματα μπορούν να διαχωριστούν κάνοντας χρήση της μεθόδου ανάλυσης σε κύριες συνιστώσες.

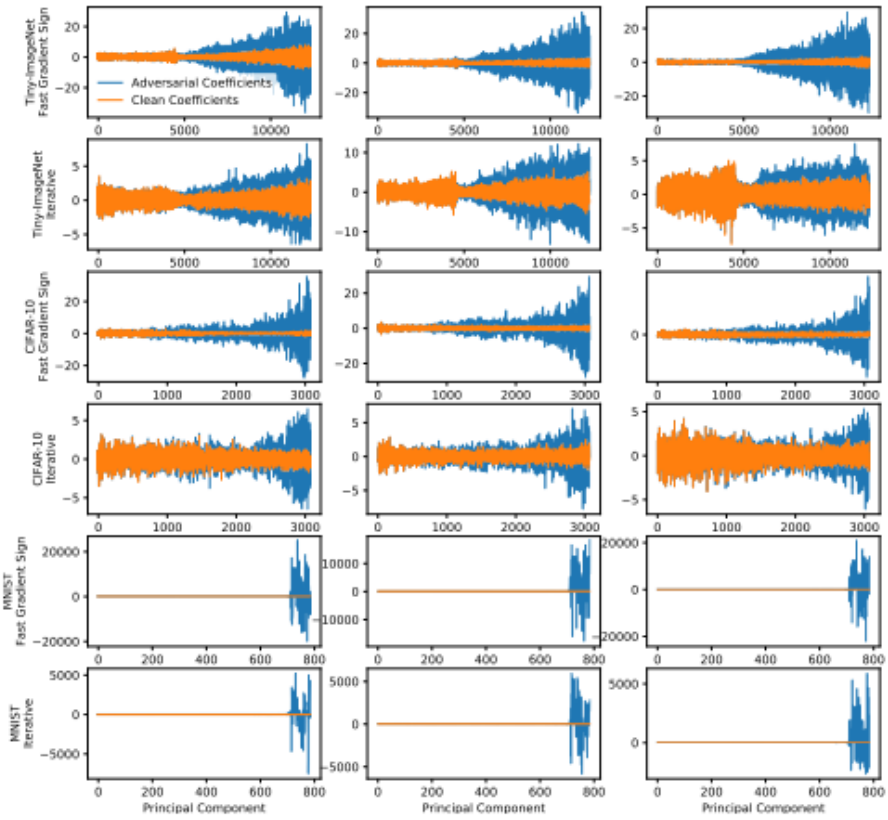
Πιο συγκεκριμένα, στην εργασία [Hend17] παρουσιάζεται η PCA-whitening, μια τεχνική ανίχνευσης αντιφατικών παραδειγμάτων βασισμένη στη μέθοδο ανάλυσης σε κύριες συνιστώσες. Για την επιτυχή εκτέλεση της παραπάνω μεθόδου χρειάζεται να ακολουθήσουμε τα εξής βήματα:

1. Κεντράρουμε τα δεδομένα εκπαίδευσης γύρω από το μηδέν.

2. Υπολογίζουμε τον πίνακα συνδιακύμανσης C των δεδομένων που προέκυψαν από το προηγούμενο βήμα.

3. Πραγματοποιούμε ανάλυση ιδιάζουσων τιμών στον C, η οποία υπολογίζεται από τον τύπο $C = U\Sigma V^T$, όπου οι στήλες του U είναι τα αριστερά-ιδιάζοντα διανύσματα του C, οι στήλες του V είναι τα δεξιά-ιδιάζοντα διανύσματα του C και ο Σ είναι ένας διαγώνιος πίνακας που περιλαμβάνει τις ιδιάζουσες τιμές του C.

4. Τέλος, εφαρμόζουμε την PCA-whitening τεχνική λαμβάνοντας μια είσοδο x και υπολογίζοντας το $\Sigma^{1/2}U^T x$.



Σχήμα 4.2: Κάθε γραφική παράσταση αντιστοιχεί σε συντελεστές διακύμανσης για ένα τυχαία επιλεγμένο ζεύγος καλοήθων και κακόβουλων εικόνων (**Πηγή**: [Hend17]).

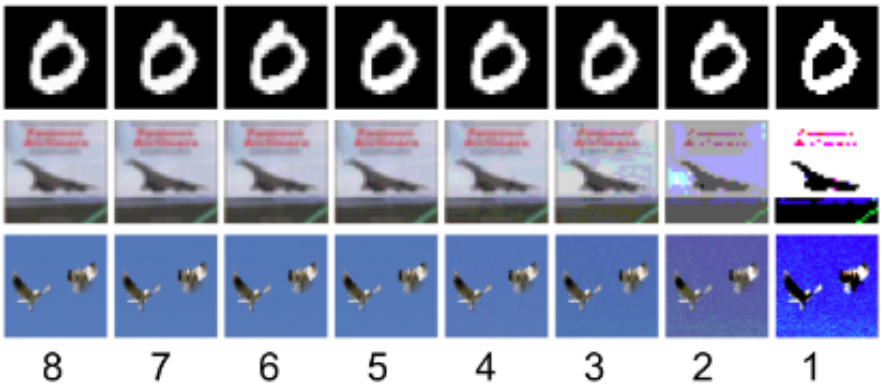
Όπως φαίνεται στο Σχήμα 4.2, τα αντιφατικά παραδείγματα παρουσιάζουν σταθερά μεγαλύτερους συντελεστές διακύμανσης στις μεγαλύτερες τιμές κύριων συνιστωσών από ότι τα καλοήθη δείγματα. Στην εργασία [Gros17], εφαρμόζεται μια στατιστική δοκιμή, η μέγιστη μέση απόκλιση [Gret12], η οποία χρησιμοποιείται για να ελέγξει εάν δύο σύνολα δεδομένων προέρχονται από την ίδια κατανομή. Οι συγγραφείς της συγκεκριμένης εργασίας χρησιμοποιούν αυτό το εργαλείο δοκιμών για να ελέγξουν εάν μια ομάδα σημείων από το σύνολο δεδομένων είναι καλοήθης ή αντιφατική.

Έλεγχος της συνέπειας των προβλέψεων του ταξινομητή

Υπάρχουν αρκετές μελέτες που επικεντρώνονται στον έλεγχο της συνέπειας του αποτελέσματος της πρόβλεψης του δείγματος x. Συνήθως σε αυτές περιπτώσεις οι αμυνόμενοι χειρίζονται τις παραμέτρους του μοντέλου ή τα ίδια τα παραδείγματα εισόδου, προκειμένου να ελέγξουν εάν οι έξοδοι του ταξινομητή έχουν υποστεί σημαντικές μεταβολές. Τα παραπάνω βασίζονται στην πεποίθηση ότι ο ταξινομητής θα έχει σταθερές προβλέψεις για τα φυσικά παραδείγματα υπό από αυτές τις συνθήκες.

Η έρευνα [Fein17] προσθέτει τυχαιότητα στον ταξινομητή χρησιμοποιώντας την τεχνική Dropout [Sriv14]. Εάν αυτοί οι ταξινομητές δώσουν πολύ διαφορετικά αποτελέσματα πρόβλεψης στο x μετά την τυχαιοποίηση, αυτό το δείγμα x είναι πολύ πιθανό να είναι κακόβουλο προϊόν κάποιας επίθεσης.

Η έρευνα [Xu18] χειρίζεται το ίδιο το δείγμα εισόδου για να ελέγξει τη συνέπεια. Μια τυπική ψηφιακή εικόνα αναπαρίσταται από έναν πίνακα από pixel, καθένα από τα οποία ισοδυναμεί συνήθως με έναν αριθμό που αντιπροσωπεύει ένα συγκεκριμένο χρώμα. Οι κοινές αναπαραστάσεις εικόνας χρησιμοποιούν αρκετά μεγάλο βάθος χρώματος, δημιουργώντας έτσι ένα μεγάλου εύρους πεδίο τιμών για την κατασκευή των εικόνων. Υποθέτοντας ότι η μείωση του βάθους του χρώματος των εικόνων



Σχήμα 4.3: Εικόνες από το MNIST, το CIFAR10 και το ImageNet. Από αριστερά προς τα δεξιά, το βάθος του χρώματος μειώνεται από 8-bit, σε 7-bit, ..., σε 2-bit, σε 1-bit (**Πηγή**: [Xu18]).

μπορεί να μειώσει την πιθανότητα δημιουργίας αντιφατικών παραδειγμάτων χωρίς να βλάψει την ακρίβεια του ταξινομητή, οι συγγραφείς της δεδομένης μελέτης προτείνουν τη μέθοδο squeezing color bits, όπου για κάθε δείγμα εισόδου x πραγματοποιείται μείωση του βάθους του χρώματους του. Για παράδειγμα, μία εικόνα σε κλίμακα του γκρι 8-bit με 256 πιθανές τιμές για κάθε pixel μετατρέπεται σε 7-bit εικόνα με 128 πιθανές τιμές, όπως φαίνεται στο Σχήμα 4.3. Επιπλέον, υποθέτουν ότι για τις φυσικές εικόνες, η μείωση του βάθους του χρώματος δεν θα αλλάξει το αποτέλεσμα της πρόβλεψης του ταξινομητή, αλλά η πρόβλεψη για τα αντιφατικά παραδείγματα θα αλλάξει. Με αυτόν τον τρόπο, μπορούν να εντοπίσουν τα αντιφατικά παραδείγματα. Παρόμοια με την τεχνική μείωσης του βάθους του χρώματος, η εργασία [Fein17] εισήγαγε επίσης άλλες μεθόδους συμπίεσης χαρακτηριστικών, όπως τη χωρική εξομάλυνση (spatial smoothing).

Επιθέσεις που ξεφεύγουν από τις τεχνικές ανίχνευσης

Οι επιθέσεις που περιγράφονται στην εργασία [Carl17a] παρακάμπτουν πολλές από τις μεθόδους ανίχνευσης που εμπίπτουν στις τρεις παραπάνω κατηγορίες. Οι μέθοδοι συμπίεσης χαρακτηριστικών αποδείχθηκαν ανεπαρκείς στην εργασία [Shar18b], όπου εισήχθη μια «ισχυρότερη» αντιφατική επίθεση. Οι συγγραφείς της εργασίας [Carl17a] ισχυρίζονται ότι οι ιδιότητες που είναι εγγενείς στα αντιφατικά παραδείγματα δεν είναι πολύ εύκολο να βρεθούν. Επίσης, προτείνουν διάφορες μελλοντικές έρευνες που μπορούν να πραγματοποιηθούν στα πλαίσια της τεχνικής της ανίχνευσης αντιφατικών παραδειγμάτων. Στη συνέχεια παραθέτουμε ορισμένες κομβικές προτάσεις τους.

1. Η τυχαιοποίηση μπορεί να αυξήσει την απαιτούμενη παραμόρφωση της επίθεσης.
2. Οι άμυνες που χειρίζονται απευθείας τις τιμές των raw pixel είναι αναποτελεσματικές.
3. Αναφορά ψευδώς θετικών και αληθινά θετικών ποσοστών στη διαδικασία της ανίχνευσης.
4. Η χρήση μιας ισχυρής επίθεσης και η εστίαση μόνο σε επιθέσεις white-box είναι επικίνδυνη.

Κεφάλαιο 5

Επιθέσεις και Άμυνες στην πράξη

5.1 Εισαγωγή στην πειραματική διαδικασία

Στα πλαίσια της παρούσας διπλωματικής εργασίας, επιλέξαμε να διερευνήσουμε περισσότερο μια ειδική υποπερίπτωση επίθεσης, την επίθεση one-pixel [Su19b], εφαρμόζοντάς την σε ένα δημοφιλές βαθύ νευρωνικό δίκτυο, το VGG16 [Simo15], χρησιμοποιώντας τα σύνολα δεδομένων MNIST [LeCu], CIFAR10 [Kriz12a] και kaggle CIFAR10 [onlib]. Αρχικά, εκτελέσαμε την επίθεση δημιουργώντας ορισμένα αντιφατικά παραδείγματα και μελετήσαμε πόσο ανθεκτικό είναι το μοντέλο μας. Έπειτα, επιλέξαμε να εφαρμόσουμε την PCA-whitening [Hend17] τεχνική ανίχνευσης αντιφατικών παραδειγμάτων για να ερευνήσουμε τη συμπεριφορά της απέναντι στη συγκεκριμένη επίθεση. Προκειμένου να έχουμε μια βάση σύγκρισης για την αξιολόγηση των αποτελεσμάτων μας, εφαρμόσαμε μια πολύ διαδεδομένη μέθοδο επίθεσης των ΒΝΔ, την FGSM [Good15], χρησιμοποιώντας την ως σημείο αναφοράς τόσο για την αξιολόγηση της αποτελεσματικότητας της επίθεσης one-pixel όσο και για τον έλεγχο της αποδοτικότητας της ανίχνευσης PCA-whitening. Στη συνέχεια, μελετήσαμε την περίπτωση συνδυασμού δύο μεθόδων επιθέσεων, εφαρμόσαμε δηλαδή μαζί και τις δύο τεχνικές επιθέσεων που αναφέραμε προηγουμένως, δημιουργώντας νέα αντιφατικά παραδείγματα. Τα κακόβουλα αυτά παραδείγματα τα εισάγαμε στον ανιχνευτή PCA-whitening και μελετήσαμε τη δική του απόδοση στη συνδυαστική αυτή επίθεση. Τέλος, πειραματιστήκαμε με μια δεύτερη τεχνική ανίχνευσης, την squeezing color bits [Xu18], συνδυάζοντας την με την PCA-whitening με σκοπό να μελετήσουμε την αποδοτικότητα μιας συνδυαστικής τεχνικής ανίχνευσης έναντι στις παραπάνω επιθέσεις.

5.2 Σύνολα δεδομένων

Η εκπαίδευση του VGG16 δικτύου πραγματοποιήθηκε πάνω στα δημοφιλή στον τομέα επεξεργασία εικόνας σύνολα δεδομένων MNIST [LeCu] και CIFAR10 [Kriz12a]. Για λόγους πληρότητας και αποσαφήνισης των επιλογών μας παραθέτουμε στη συνέχεια τα βασικά χαρακτηριστικά των συνόλων δεδομένων που χρησιμοποιήθηκαν.

Το MNIST είναι ένα σύνολο δεδομένων από εικόνες με χειρόγραφα ψηφία από 0-9 σε κλίμακα του γκρι. Πρόκειται για ένα υποσύνολο ενός μεγαλύτερου συνόλου δεδομένων που διατίθεται από το NIST. Οι πρωτότυπες ασπρόμαυρες εικόνες από το NIST κανονικοποιήθηκαν σε μέγεθος ώστε να μην ξεπερνούν σε διαστάσεις τα 20x20 pixel, ενώ συγχρόνως διατήρησαν την αναλογία ύψους και πλάτους της αρχικής εικόνας. Επιπλέον, οι κανονικοποιημένες αυτές εικόνες τοποθετήθηκαν στο κέντρο ενός πλαισίου 28x28 pixel, υπολογίζοντας αρχικά το κέντρο μάζας του συνόλου των pixel κάθε εικόνας και μεταφέροντας τες στη συνέχεια μία προς μία στο πλαίσιο με τέτοιο τρόπο ώστε το κέντρο μάζας κάθε 20x20 εικόνας να συμπίπτει με το κέντρο του αντίστοιχου 28x28 πλαισίου. Η γκρι απόχρωση των τελικών δειγμάτων οφείλεται στην τεχνική *ελαχιστοποίησης της παραμόρφωσης της εικόνας* (anti-aliasing) που εφαρμόστηκε, η οποία χρησιμοποιείται συχνά σε περιπτώσεις όπου εικόνες υψηλής ανάλυσης χρειάζεται να αναπαρασταθούν σε χαμηλότερη ανάλυση. Το σύνολο δεδομένων εκπαίδευσης του MNIST αποτελείται από 60.000 εικόνες με τις αντίστοιχες ετικέτες τους, ενώ το σύνολο ελέγχου περιλαμβάνει 10.000 δείγματα με τις αντίστοιχες κλάσεις τους.

Το σύνολο δεδομένων CIFAR10 αποτελείται από 60.000 έγχρωμες εικόνες διαστάσεων 32x32

pixel κατανεμημένες σε 10 κλάσεις - αεροπλάνο, αυτοκίνητο, πουλί, γάτα, ελάφι, σκύλος, βάτραχος, άλογο, πλοίο, φορτηγό -, με 6.000 εικόνες ανά κλάση. Υπάρχουν 50.000 εικόνες με τις αντίστοιχες ετικέτες τους που χρησιμεύουν στην εκπαίδευση του εκάστοτε μοντέλου και 10.000 εικόνες μαζί με τις ετικέτες τους που χρησιμεύουν στον έλεγχο της αποδοτικότητας του εκπαιδευμένου μοντέλου. Επιπλέον, το CIFAR10 χωρίζεται σε πέντε δέσμες εκπαίδευσης και μια δέσμη ελέγχου, όπου η κάθε δέσμη αποτελείται από 10.000 εικόνες. Πιο συγκεκριμένα, η δέσμη ελέγχου περιέχει ακριβώς 1.000 τυχαία επιλεγμένες εικόνες από κάθε κλάση, ενώ οι υπόλοιπες εικόνες έχουν μοιραστεί με τυχαίο τρόπο στις δέσμες της εκπαίδευσης. Αυτό πρακτικά σημαίνει ότι ορισμένες δέσμες εκπαίδευσης είναι πιθανό να περιέχουν περισσότερες εικόνες από τη μία κατηγορία από ότι από κάποια άλλη, ωστόσο συνολικά οι δέσμες εκπαίδευσης περιέχουν 5.000 εικόνες από κάθε κλάση, δηλαδή υπάρχει πλήρης ισοκατανομή στις κατηγορίες. Επίσης, οι 10 κατηγορίες που περιλαμβάνει το CIFAR10 είναι αμοιβαία αποκλειστικά σύνολα, καθώς δεν υπάρχει καμία αλληλοεπικάλυψη στις εικόνες που ανήκουν σε μια κλάση με τις εικόνες που ανήκουν σε άλλη.

Πρόκειται λοιπόν για δύο ανοιχτά σύνολα δεδομένων όπου ο κάθε ενδιαφερόμενος εύκολα μπορεί να κατεβάσει από τις επίσημες ιστοσελίδες τους [LeCu, Kriz12a]. Σε ότι αφορά στο σύνολο δεδομένων ελέγχου που χρειάστηκε στα πειράματα μας, ως βάση του συνόλου των αντιφατικών παραδειγμάτων που δημιουργήσαμε, χρησιμοποιήσαμε το σύνολο δεδομένων του kaggle CIFAR10 αντί του πρωτότυπου CIFAR10. Το σύνολο δεδομένων εκπαίδευσης του kaggle CIFAR10 είναι το ίδιο με του CIFAR10. Η διαφορά τους έγκειται στο σύνολο των δεδομένων ελέγχου, όπου οι δημιουργοί του kaggle CIFAR10 τροποποίησαν τις 10.000 εικόνες ελέγχου της αρχικής συλλογής δεδομένων, προσθέτοντας ακόμη 290.000 εικόνες. Η περιστροφή, η αποκοπή, η θόλωση και η προσθήκη μερικών τυχαίων pixel αποτελούν ορισμένες από τις τεχνικές τροποποίησης που χρησιμοποιήθηκαν και είναι εμφανείς οπτικά, ωστόσο ο ακριβής αλγόριθμος τροποποίησης των εικόνων δεν έχει δημοσιευτεί. Η διαφορά αυτή που παρουσιάζουν οι δύο συλλογές δεδομένων καθιστά το kaggle CIFAR10 ένα πιο πρακτικό σύνολο δεδομένων που προσομοιώνει συνήθη σενάρια, όπου οι εικόνες περιέχουν άγνωστο τυχαίο θόρυβο.

5.3 Μοντέλο VGG16

Για τη δημιουργία των δύο μοντέλων-στόχων - ένα για κάθε σύνολο δεδομένων εκπαίδευσης - χρησιμοποιήσαμε το προεκπαιδευμένο μοντέλο VGG16 [Simo15]. Το VGG16 είναι μια αρχιτεκτονική συνελικτικού νευρωνικού δικτύου εκπαιδευμένου στο σύνολο δεδομένων ImageNet [onlid], που χρησιμοποιήθηκε για να κερδίσει το διαγωνισμό ILSVR το 2014 [Russ15]. Θεωρείται ως μια από τις καλύτερες αρχιτεκτονικές μοντέλων όρασης. Το σύνολο δεδομένων Imagenet πάνω στο οποίο πραγματοποιήθηκε η διαδικασία της εκπαίδευσης του μοντέλου, αποτελείται από λίγο παραπάνω από 14 εκατομμύρια εικόνες οι οποίες ταξινομούνται σε μία ή περισσότερες από τις περίπου 22 χιλιάδες κατηγορίες.

Ο αριθμός 16 στην ονομασία του συγκεκριμένου μοντέλου αναφέρεται στο πλήθος των επιπέδων που διαθέτουν βάρη. Πρόκειται για ένα αρκετά μεγάλο δίκτυο που περιλαμβάνει περίπου 138 εκατομμύρια παραμέτρους, ενώ έχει συνολικά 23 επίπεδα βάθος. Το σημαντικότερο ίσως σημείο στην αρχιτεκτονική του μοντέλου αποτελεί το γεγονός ότι οι δημιουργοί του, αντί να εστιάσουν στην αύξηση του αριθμού των υπερπαραμέτρων, επικεντρώθηκαν στο να σχεδιάσουν συνελικτικά επίπεδα φίλτρων διαστάσεων 3x3, με το ίδιο *γέμισμα* (padding) και ένα επίπεδο *σμίκρυνσης μεγίστου* (max pooling) διάστασης 2x2. Όπως μπορούμε να διακρίνουμε στον Πίνακα 5.1, τα συνελικτικά επιπέδα ακολουθούνται από επίπεδα σμίκρυνσης μεγίστου, με την διαδικασία να επαναλαμβάνεται αλλάζοντας κάθε φορά τον αριθμό και τις παραμέτρους των συνελικτικών επιπέδων.

Στο τελικό στάδιο του δικτύου χρησιμοποιήσαμε ένα πλήρως συνδεδεμένο επίπεδο ακολουθούμενο από τη συνάρτηση ενεργοποίησης softmax. Συνήθως, ένα μοντέλο VGG16 περιλαμβάνει στο τέλος δύο πλήρως συνδεδεμένα επίπεδα. Ωστόσο παρατηρήθηκε πειραματικά ότι ένα πλήρως συνδεδεμένο επίπεδο εξυπηρετεί καλύτερα τους σκοπούς της παρούσας διπλωματικής εργασίας, καθώς έτσι σημειώσαμε σημαντικά υψηλότερο ποσοστό επιτυχίας πάνω στο MNIST dataset, εκτελώντας τον

Layer (type)	Output Shape	Number of parametres
InputLayer	(None, 32, 32, 3)	0
Conv2D	(None, 32, 32, 64)	1792
Conv2D	(None, 32, 32, 64)	36928
MaxPooling2D	(None, 16, 16, 64)	0
Conv2D	(None, 16, 16, 128)	73856
Conv2D	(None, 16, 16, 128)	147584
MaxPooling2D	(None, 8, 8, 128)	0
Conv2D	(None, 8, 8, 256)	295168
Conv2D	(None, 8, 8, 256)	590080
Conv2D	(None, 8, 8, 256)	590080
MaxPooling2D	(None, 4, 4, 256)	0
Conv2D	(None, 4, 4, 512)	1180160
Conv2D	(None, 4, 4, 512)	2359808
Conv2D	(None, 4, 4, 512)	2359808
MaxPooling2D	(None, 2, 2, 512)	0
Conv2D	(None, 2, 2, 512)	2359808
Conv2D	(None, 2, 2, 512)	2359808
Conv2D	(None, 2, 2, 512)	2359808
MaxPooling2D	(None, 1, 1, 512)	0
Flatten	(None, 512)	0
FC	(None, 512)	262656
Dense	(None, 10)	5130

Πίνακας 5.1: Αρχιτεκτονική VGG16 μοντέλου.

ίδιο αριθμό επαναλήψεων στο στάδιο της εκπαίδευσης. Επιπλέον, επιλέξαμε να χρησιμοποιήσουμε μια κοινή αρχιτεκτονική μοντέλου και για τα δύο σύνολα δεδομένων, καθώς κύριος στόχος μας είναι να διερευνήσουμε πως επιδρούν οι επιθέσεις και οι άμυνες που εφαρμόσαμε στα πειράματά μας σε ένα επιτυχημένο βαθύ νευρωνικό δίκτυο, όπως το VGG16, χρησιμοποιώντας δύο σύνολα δεδομένων για επαλήθευση και μερική γενίκευση των συμπερασμάτων μας. Ως συνάρτηση κόστους χρησιμοποιήσαμε την αραιή διασταυρούμενη εντροπία αντί της διασταυρούμενης εντροπίας που συνηθίζεται σε προβλήματα ταξινόμησης με συνάρτηση ενεργοποίησης softmax, καθώς οι τιμές των ετικετών των δειγμάτων εισάγονταν ως ακέραιοι αριθμοί και όχι σε μορφή κωδικοποίησης διανύσματος one-hot.

Τέλος, τα δεδομένα εισόδου που εισάγουμε στο VGG16 μοντέλο πρέπει να έχουν ακριβώς 3 κανάλια εισόδου και το πλάτος και το ύψος τους δεν πρέπει να είναι μικρότερο από 32 pixel. Συνεπώς, οι εικόνες από το σύνολο δεδομένων του MNIST, πριν εισαχθούν στο μοντέλο VGG16, υποβάλλονται σε επεξεργασία, μετατρέποντας τις εικόνες από 28x28 pixel σε 32x32 pixel και από γκρι κλίμακα σε RGB, ώστε να πληρούν τις παραπάνω προϋποθέσεις. Συγχρόνως, μετατρέψαμε τις εικόνες εισόδου και από τις 3 συλλογές δεδομένων από RGB σε BGR και επίσης κεντράραμε κάθε κανάλι χρώματος γύρω από το μηδέν χωρίς να εφαρμόσουμε κάποια τεχνική κλιμάκωσης. Έπειτα από 20 εποχές εκπαίδευσης το ποσοστό επιτυχίας που σημείωσε το μοντέλο στο σύνολο των δεδομένων ελέγχου του CIFAR10 ήταν $63,37\%$ και έπειτα από 3 εποχές εκπαίδευσης στο σύνολο δεδομένων του MNIST το αντίστοιχο ποσοστό επιτυχίας ήταν $94,72\%$.

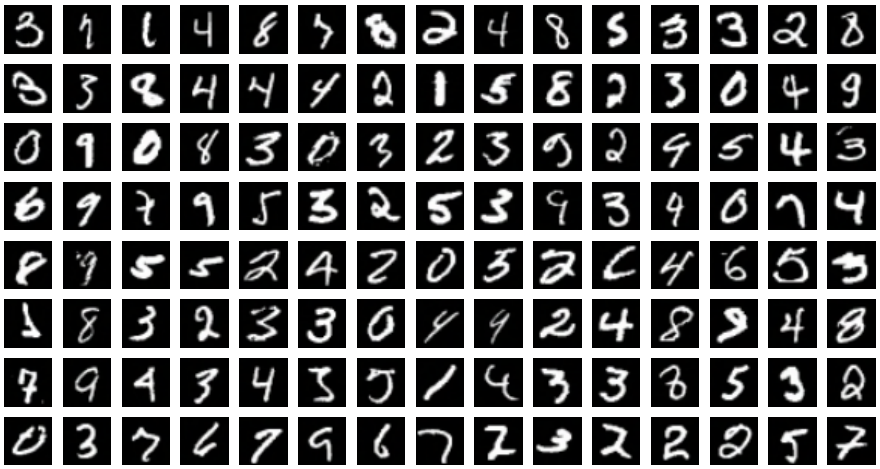
5.4 Εφαρμογή επιθέσεων

5.4.1 Επίθεση one-pixel

Στην επίθεση one-pixel βασικός στόχος του επιτιθέμενου είναι να δημιουργήσει την κατάλληλη διαταραχή σε ένα μόνο pixel της εικόνας-στόχου έτσι ώστε να ξεγελάσει τον ταξινομητή του μοντέλου. Στο πείραμά μας, κωδικοποιήσαμε τη διαταραχή σε έναν πίνακα, ο οποίος αποτελεί μια υποψήφια λύση και βελτιστοποιείται-εξελίσσεται μέσω του αλγορίθμου της διαφορικής εξέλιξης. Στην περίπτωση της επίθεσης one-pixel, μία υποψήφια λύση περιέχει μόνο μια διαταραχή που τροποποιεί ένα pixel και συνίσταται από μια πλειάδα που περιέχει πέντε στοιχεία: τις συντεταγμένες $x - y$ που αντιστοιχούν στη θέση του pixel-θύματος μέσα στο 32x32 πλέγμα της εικόνας και την τιμή της διαταραχής για κάθε ένα από τα 3 κανάλια RGB. Ο αρχικός αριθμός των υποψηφίων λύσεων (πληθυσμός) είναι 10 και σε κάθε επανάληψη άλλες 10 υποψήφιες λύσεις (παιδιά) θα παραχθούν χρησιμοποιώντας αλγόριθμο διαφορικής εξέλιξης, που περιγράφεται από την Εξίσωση 5.1:

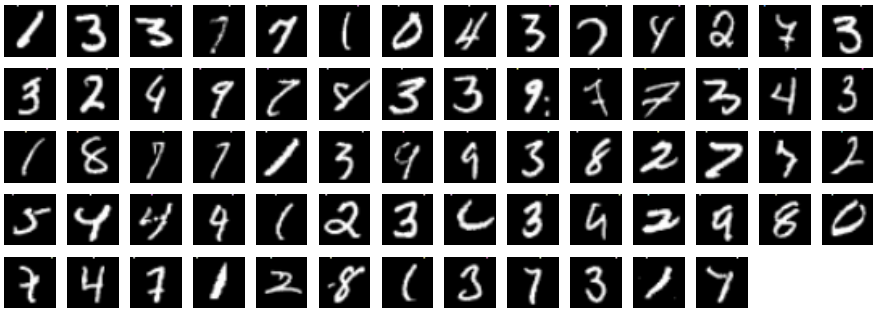
$$\begin{aligned} x_i(g+1) =& x_{r1}(g) + F(x_{r2}(g) - x_{r3}(g)), \\ & r1 \neq r2 \neq r3 \end{aligned} \tag{5.1}$$

όπου x_i είναι μια πιθανή υποψήφια λύση, $r1, r2, r3$ είναι τυχαίοι αριθμοί μεταξύ του 0 και 9 διαφορετικοί από το i, F είναι το διαφορικό βάρος που παίρνει τιμές από την κανονική κατανομή $U(0.5, 1)$ και διαφέρει από γενιά σε γενιά και g είναι ο τρέχων δείκτης γενιάς.



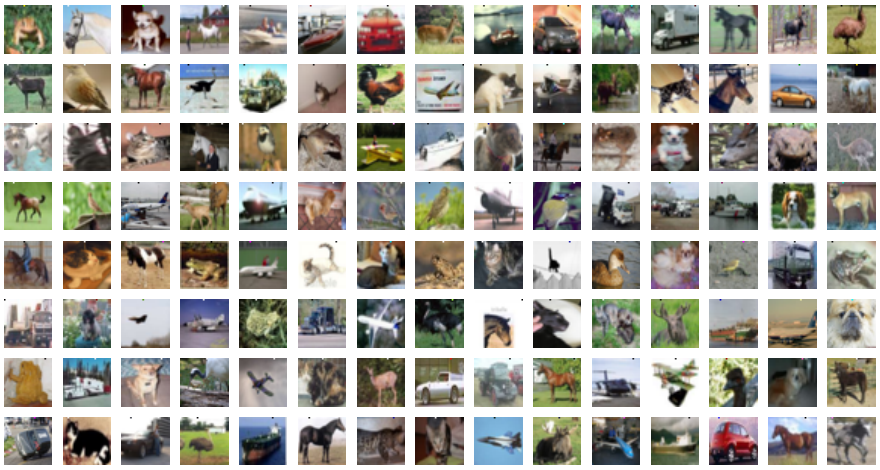
Σχήμα 5.1: «Μη ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της επίθεσης one-pixel, χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων MNIST.

Μόλις δημιουργηθεί για κάθε υποψήφια λύση $x_i(g)$ του πληθυσμού η νέα γενιά $x_i(g+1)$, τότε κάθε παιδί ανταγωνίζεται τους αντίστοιχους γονείς τους σύμφωνα με τον δείκτη i του πληθυσμού και ο νικητής επιβιώνει για την επόμενη επανάληψη. Ο μέγιστος αριθμός επαναλήψεων-γενεών ορίζεται σε 600 και το κριτήριο πρώιμης διακοπής της επαναληπτικής διαδικασίας ενεργοποιείται όταν ο ταξινομητής του μοντέλου μας προβλέψει εσφαλμένη κλάση για το τροποποιημένο, με την προσθήκη της διαταραχής, δείγμα με πιθανότητα άνω του 85% στην περίπτωση του MNIST και άνω του 75% στην περίπτωση του kaggle CIFAR10. Να σημειωθεί ότι σε αυτό το σημείο έγινε προσπάθεια εφαρμογής της επίθεσης και στη συλλογή δεδομένων CIFAR10, ωστόσο παρατηρήθηκε ότι, λόγω της υψηλής αρχικής πιθανότητας η εκάστοτε εικόνα να ανήκει σε συγκεκριμένη κατηγορία, ο αλγόριθμος διαφορικής εξέλιξης απαιτούσε περισσότερες επαναλήψεις καθώς και διαφορετική παραμετροποίηση. Μοναδική προϋπόθεση για την επιλογή ενός δείγματος ως στόχου για την εκτέλεση της επίθεσης είναι η εικόνα αυτή να ταξινομείται ορθά από τον ταξινομητή του VGG16 πριν την επίθεση. Άλλωστε, η εφαρμογή της επίθεσης σε ένα δείγμα που εξ' αρχής «ξεγελά» τον ταξινομητή δεν θα είχε κάποιο νό-



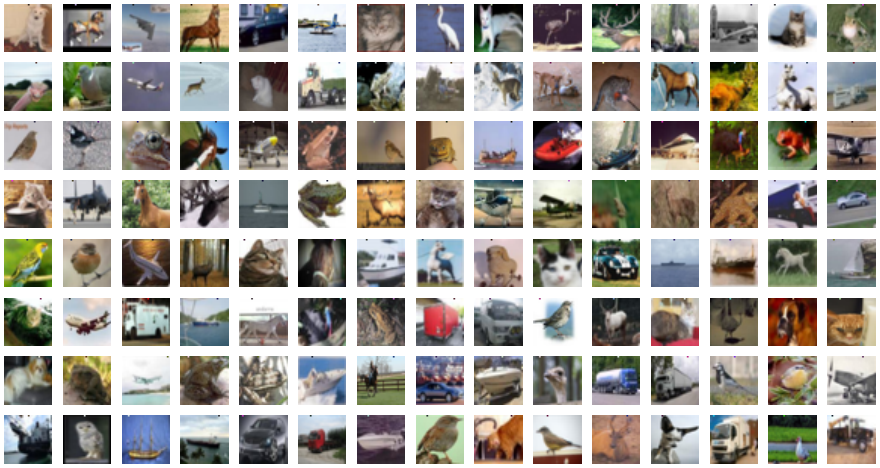
Σχήμα 5.2: «Ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της επίθεσης one-pixel, χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων MNIST

ημα. Ενδεικτικά, στα Σχήματα 5.1,5.2,5.3,5.4 παρατίθενται ορισμένα αντιφατικά παραδείγματα που δημιουργήσαμε, εφαρμόζοντας την παραπάνω μη στοχευμένη επίθεση.

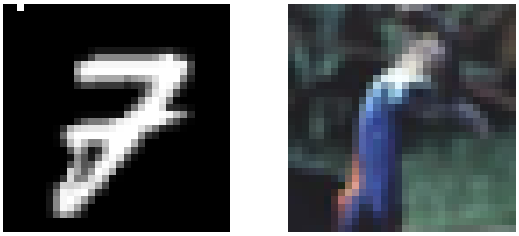


Σχήμα 5.3: «Μη ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της επίθεσης one-pixel χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων kaggle CIFAR10.

Αρχικά ως επιτυχημένα παραδείγματα θεωρήσαμε εκείνα για τα οποία ο ταξινομητής VGG16 προβλέπει εσφαλμένη κλάση με πιθανότητα άνω του 85% στην περίπτωση του MNIST και άνω του 75% στην περίπτωση του kaggle CIFAR10. Ωστόσο, εξ' ορισμού, μια επίθεση είναι πετυχημένη εφόσον καταφέρει να παραπλανήσει τον ταξινομητή. Ένα 24% ποσοστό εμπιστοσύνης του ταξινομητή στο τελικό αποτέλεσμα, αν και φαίνεται σχετικά χαμηλό, δηλώνει ότι οι υπόλοιπες 9 κλάσεις, περιλαμβανομένης και της πραγματικής, έχουν ακόμη χαμηλότερα ποσοστά σε μια ομοιόμορφη κατανομή κατηγοριών. Για το λόγο αυτό, κρίναμε ότι ένας διαχωρισμός ανάμεσα στις επιτυχημένες επιθέσεις είναι απαραίτητος και έτσι κατηγοριοποιήσαμε τα επιτυχημένα αντιφατικά παραδείγματα σε «ισχυρά» - όταν το ποσοστό εμπιστοσύνης είναι άνω του 85% στην περίπτωση του MNIST και άνω του 75% στην περίπτωση του kaggle CIFAR10 - και «μη ισχυρά» - όταν το ποσοστό εμπιστοσύνης είναι λιγότερο από 85% στην περίπτωση του MNIST και 75% στην περίπτωση του kaggle CIFAR10, αλλά τα δείγματα εξακολουθούν τα παραπλανούν τον ταξινομητή. Αποτυχημένες επιθέσεις θεωρήθηκαν εκείνες στις οποίες τα τροποποιημένα με διαταραχή δείγματα δεν κατάφεραν να ξεγελάσουν τον ταξινομητή, παρ' ότι σημειώθηκε μείωση στο ποσοστό εμπιστοσύνης του ταξινομητή. Στο Σχήμα 5.5 παραθέτονται εικόνες δειγμάτων από αποτυχημένες επιθέσεις.



Σχήμα 5.4: «Ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της επίθεσης one-pixel χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων kaggle CIFAR10.



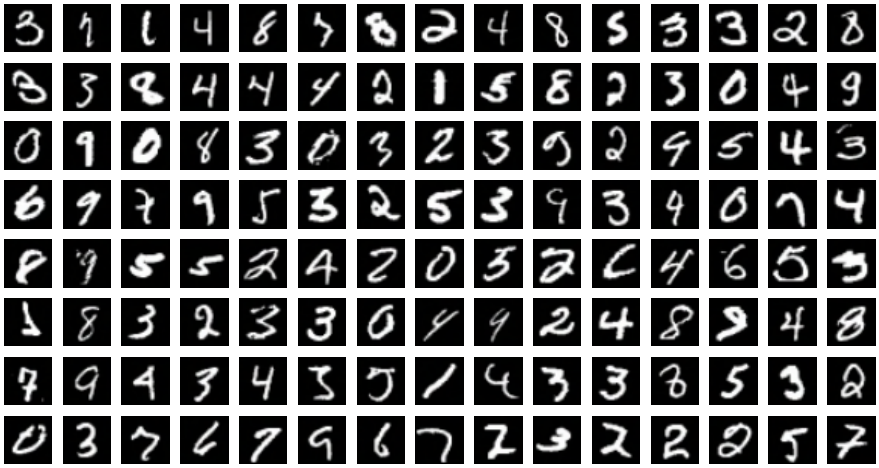
Σχήμα 5.5: Ενδεικτικό δείγμα μη επιτυχημένης επίθεσης one-pixel στο MNIST (αριστερά) και FGSM επίθεσης στο kaggle CIFAR10 (δεξιά). (α) Πριν την επίθεση ο ταξινομητής του μοντέλου κατηγοριοποιεί το δείγμα στην κλάση «7» με 81% εμπιστοσύνης. Μετά την one-pixel επίθεση ο ταξινομητής κατηγοριοποιεί το κακόβουλο αυτό δείγμα πάλι στην κλάση «7», αυτή τη φορά όμως με 53% εμπιστοσύνης. (β) Πριν την επίθεση ο ταξινομητής του μοντέλου κατηγοριοποιεί το δείγμα στην κλάση «πουλί» με 100% εμπιστοσύνης, ποσοστό το οποίο διατηρείται για την ίδια κλάση και μετά την FGSM επίθεση.

5.4.2 Επίθεση FGSM

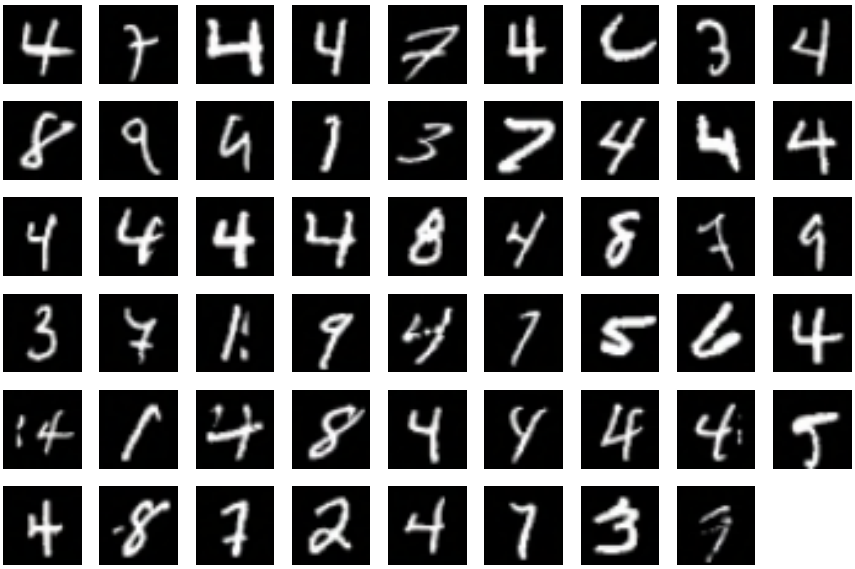
Η επίθεση FGSM αποτελεί μια από τις πιο διαδεδομένες και επιτυχημένες τεχνικές επιθέσεων white-box που χρησιμοποιεί τις πληροφορίες κλίσης της συνάρτησης κόστους του ταξινομητή για βελτίωση της. Συναντάται συχνά στη βιβλιογραφία ως σημείο αναφοράς, κυρίως για την αποτελεσματικότητα των αμυντικών στρατηγικών αλλά και για την αποδοτικότητα άλλων επιθέσεων. Στην παρούσα διπλωματική εργασία εφαρμόσαμε την επίθεση αυτή έτσι ώστε στο μετέπειτα συνδυαστικό βήμα των επιθέσεων να την χρησιμοποιήσουμε ως σημείο αναφοράς και να συγκρίνουμε την απόδοση της συνδυαστικής επίθεσης σε σχέση με τις μονές καθώς και με το πόσο εύκολα ανιχνεύσιμη είναι η διπλή επίθεση από τις αμυντικές τεχνικές που θα δούμε στη συνέχεια.

Θεωρώντας ότι η συνάρτηση κόστους του μοντέλου μας είναι γνωστή στο εισβολέα υλοποιήσαμε την μη στοχευμένη επίθεση FGSM σύμφωνα με την Εξίσωση 3.6. Έπειτα από αρκετές πειραματικές δοκιμές καταλήξαμε ότι στη δική μας περίπτωση η τιμή 1 αποτελεί μια καλή τιμή για την παράμετρο ϵ της FGSM επίθεσης. Καθοριστικό ρόλο στην επιλογή της τιμής αυτής είχε η απόδοση της διπλής επίθεσης που εφαρμόσαμε σε επόμενη φάση και θα δούμε αναλυτικότερα στη συνέχεια. Τα δείγματα-στόχοι επιλέχθηκαν με μοναδικό κριτήριο εάν πριν από οποιαδήποτε επίθεση ο ταξινομητής τα κατηγοριοποιεί ορθά, περιορισμός που εφαρμόστηκε και στην περίπτωση της επίθεσης one-pixel, όπως είδαμε προηγουμένως. Επίσης, για τον διαχωρισμό των επιτυχών αντιφατικών παραδειγμάτων σε «ισχυρών» και «μη» ακολουθήθηκε η ίδια ακριβώς διαδικασία με αυτήν που περιγράφηκε προη-

γουμένως. Στο Σχήμα 5.5 δίνεται για λόγους πληρότητας ένα παράδειγμα αποτυχημένης επίθεσης FGSM χρησιμοποιώντας δείγμα από το kaggle CIFAR10, ενώ στα Σχήματα 5.6,5.7,5.8,5.9 παρατίθενται ορισμένα «ισχυρά» και «μη» αντιφατικά παραδείγματα που δημιουργήθηκαν έπειτα από την εκτέλεση της επίθεσης FGSM.



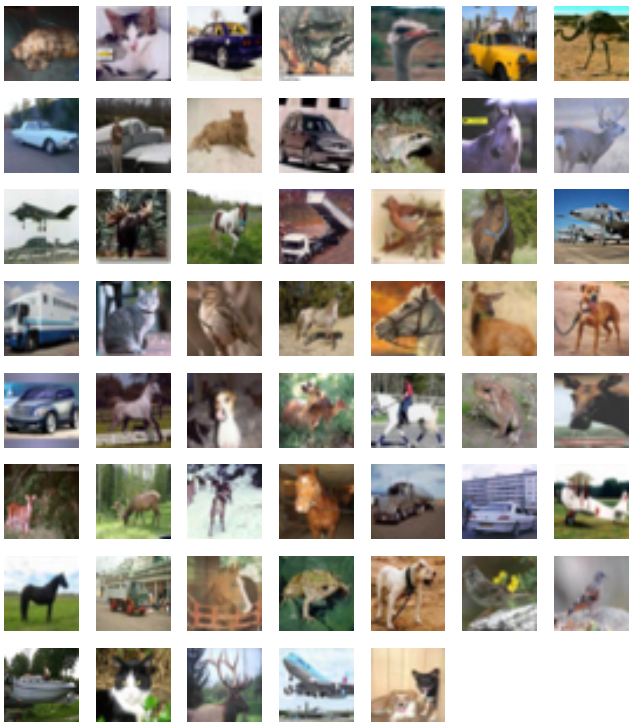
Σχήμα 5.6: «Μη ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της επίθεσης FGSM χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων MNIST.



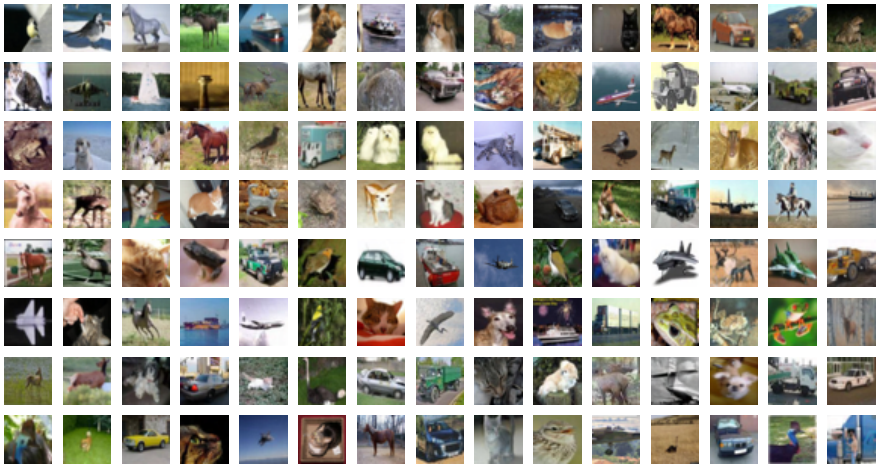
Σχήμα 5.7: «Ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της επίθεσης FGSM χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων MNIST.

5.4.3 Διπλή επίθεση

Όπως είδαμε αναλυτικά σε προηγούμενο κεφάλαιο, υπάρχει πλούσια βιβλιογραφία πάνω σε επιστημονικές έρευνες που μελετούν διάφορες μεμονωμένες επιθέσεις σε ΒΝΔ. Παρατηρήσαμε ωστόσο ότι δεν υπάρχουν αναφορές σχετικές με συνδυασμό μεθόδων για την διεξαγωγή μιας επίθεσης. Στα πλαίσια αυτής της παρατήρησης, αποφασίσαμε να μελετήσουμε μια συνδυαστική μη στοχευμένη επίθεση white-box αποτελούμενη από δύο μεθόδους επιθέσεων, την one-pixel και την FGSM που είχαμε



Σχήμα 5.8: «Μη ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της επίθεσης FGSM χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων kaggle CIFAR10.

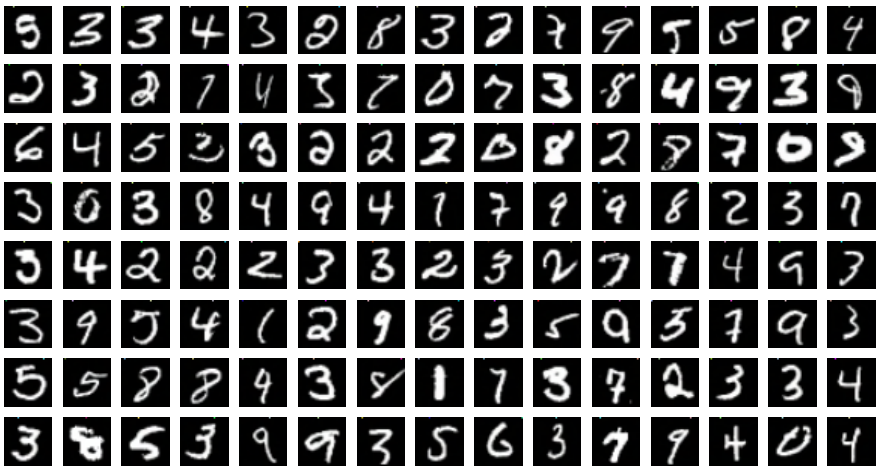


Σχήμα 5.9: «Ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της επίθεσης FGSM χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων kaggle CIFAR10.

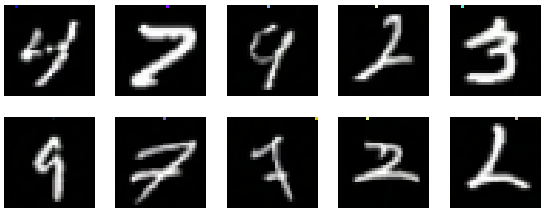
εφαρμόσει ήδη σε προηγούμενη φάση. Κατά τη διάρκεια της πειραματικής διαδικασίας ήρθαμε αντιμέτωποι με ένα κρίσιμο για την διεκπεραίωση του στόχου μας πρόβλημα, παρατηρώντας ότι στη γενική περίπτωση οι επιθέσεις που εφαρμόζαμε αλληλοαναιρούνταν και συνεπώς η σύνθεση επιτυχημένων αντιφατικών παραδειγμάτων δεν ήταν εύκολη.

Συγκεκριμένα, όταν εφαρμόσαμε πρώτα την επίθεση one-pixel και έπειτα την FGSM στα δείγματα-στόχο, το ποσοστό των επιτυχημένων αντιφατικών παραδειγμάτων που είχαν δεχτεί τη διπλή επίθεση ήταν πολύ χαμηλό. Χαρακτηριστικά αναφέρουμε ότι για το σύνολο δεδομένων του kaggle CIFAR10, περίπου το 9% των δειγμάτων που είχαν ήδη υποστεί επιτυχώς την επίθεση one-pixel κατάφεραν με επιτυχία να δεχθούν την επίθεση FGSM, έτσι ώστε να δημιουργηθούν αντιφατικά παραδείγματα διπλής επίθεσης, εκ των οποίων σχεδόν τα μισά μόνο (περίπου $4,7\%$ του συνόλου) περνούσαν το όριο του 75% της εμπιστοσύνης του ταξινομητή που έχουμε ορίσει για της υψηλής ποιότητας επιθέσεις. Αντίθετα, όταν εφαρμόσαμε πρώτα την επίθεση FGSM και έπειτα την one-pixel, το ποσοστό των επιτυχημένων αντιφατικών παραδειγμάτων που είχαν δεχτεί τη διπλή επίθεση ήταν σχετικά ικανοποιητικό, με το $87,6\%$ των δειγμάτων που είχαν ήδη υποστεί την επίθεση FGSM να δέχονται με επιτυχία την επίθεση one-pixel και το $69,7\%$ να περνούν το όριο του ποσοστού εμπιστοσύνης του ταξινομητή για το σύνολο δεδομένων του kaggle CIFAR10. Σε ότι αφορά στο σύνολο δεδομένων του MNIST, και στις δύο περιπτώσεις διπλής επίθεσης το ποσοστό των επιτυχημένων αντιφατικών παραδειγμάτων είναι ακόμα πιο χαμηλό, με το $44,1\%$ των δειγμάτων που είχαν ήδη υποστεί επιτυχώς την FGSM επίθεση να δέχονται με επιτυχία την επίθεση one-pixel και μόνο το $1,7\%$ να περνούν το όριο του 85% της εμπιστοσύνης του ταξινομητή.

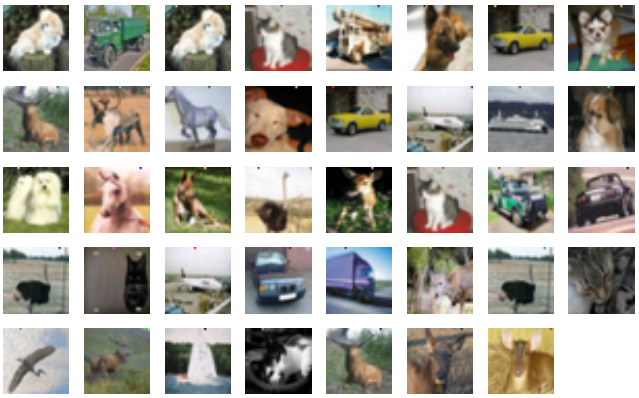
Έπειτα από αρκετές δοκιμές καταφέραμε να ολοκληρώσουμε τη διπλή επίθεση εφαρμόζοντας πρώτα την επίθεση FGSM στα δύο σύνολα των δεδομένων ελέγχου και στη συνέχεια χρησιμοποιώντας μόνο τα επιτυχημένα αντιφατικά παραδείγματα εφαρμόσαμε την επίθεση one-pixel, παράγοντας τελικά έναν ικανό - για την υπόλοιπη μελέτη μας - αριθμό αντιφατικών παραδειγμάτων. Στην περίπτωση του kaggle CIFAR10, για την εκτέλεση της δεύτερης επίθεσης χρησιμοποιήσαμε μόνο τα «ισχυρά» αντιφατικά παραδείγματα που προέκυψαν από την επίθεση FGSM, καθώς το πλήθος τους ήταν επαρκές, ενώ στην περίπτωση του MNIST, η επίθεση one-pixel εφαρμόστηκε σε όλα τα επιτυχημένα αντιφατικά παραδείγματα της πρώτης επίθεσης, «ισχυρά» και «μη». Στα Σχήματα 5.10,5.11,5.12,5.13 παραθέτουμε ορισμένα από τα τελικά αποτελέσματα της διπλής αυτής συνδυαστικής επίθεσης. Στο Παράρτημα Α.1 παραθέτουμε για λόγους πληρότητας ορισμένα αποτελέσματα από τη διπλή επίθεση που εκτελέστηκε στο σύνολο δεδομένων του kaggle CIFAR10 εφαρμόζοντας πρώτα την επίθεση one-pixel και έπειτα την FGSM.



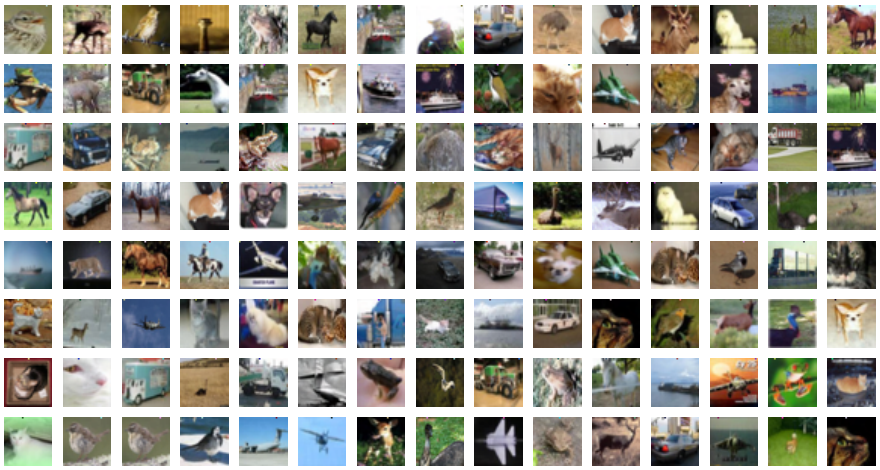
Σχήμα 5.10: «Μη ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της διπλής επίθεσης χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων MNIST.



Σχήμα 5.11: «Ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της διπλής επίθεσης χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων MNIST.



Σχήμα 5.12: «Μη ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της διπλής επίθεσης χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων kaggle CIFAR10.

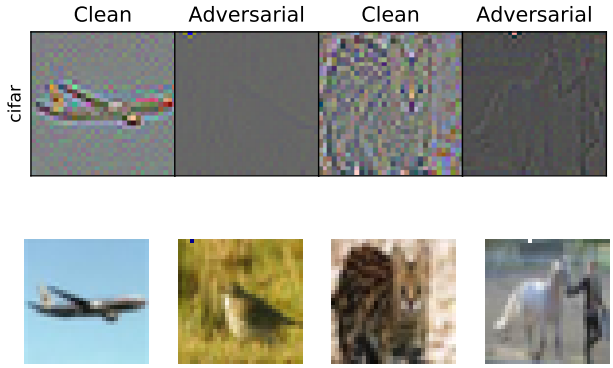


Σχήμα 5.13: «Ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της διπλής επίθεσης χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων kaggle CIFAR10.

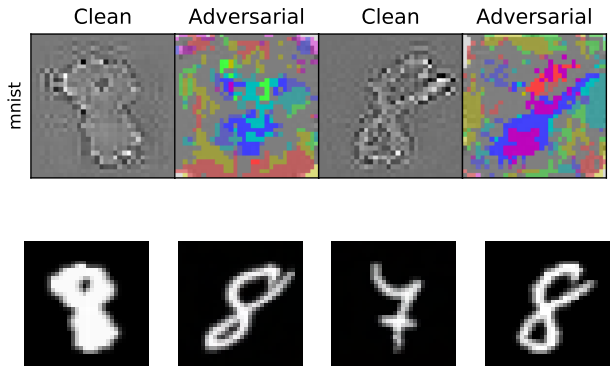
5.5 Εφαρμογή τεχνικών ανίχνευσης επιθέσεων

5.5.1 Τεχνική ανίχνευσης PCA-whitening

Για να ελέγξουμε την ισχύ των αντιφατικών παραδειγμάτων που δημιουργήσαμε στην προηγούμενη ενότητα αναπτύξαμε αρχικά την τεχνική ανίχνευσης PCA-whitening. Πρώτος μας στόχος υπήρξε η μελέτη της δυναμικής του ζεύγους *επίθεση one-pixel - άμυνα PCA-whitening*. Στη συνέχεια, επεκτείναμε τη μελέτη αυτή και στο ζεύγος *επίθεση FGSM - άμυνα PCA-whitening* έτσι ώστε να υπάρχει ένα μέτρο σύγκρισης και αξιολόγησης των αποτελεσμάτων της αρχικής μελέτης. Επιπλέον, εξετάσαμε πόσο αποτελεσματικός είναι ο ανιχνευτής PCA-whitening στον εντοπισμό αντιφατικών παραδειγμάτων που έχουν προκύψει από τη διπλή επίθεση που εφαρμόσαμε και συγχρόνως εάν η διπλή επίθεση είναι περισσότερο ανθεκτική απέναντι σε αμυντικούς μηχανισμούς συγκριτικά με τις μεμονωμένες επιθέσεις FGSM και one-pixel.



Σχήμα 5.14: «Μη ισχυρά» αντιφατικά παραδείγματα πριν (δεύτερη σειρά) και μετά (πρώτη σειρά) από την εφαρμογή της PCA-whitening μεθόδου. Σημειώνεται ότι τα δείγματα αυτά έχουν προκύψει από επίθεση one-pixel.



Σχήμα 5.15: «Ισχυρά» αντιφατικά παραδείγματα πριν (δεύτερη σειρά) και μετά (πρώτη σειρά) από την εφαρμογή της PCA-whitening μεθόδου. Σημειώνεται ότι τα δείγματα αυτά έχουν προκύψει από επίθεση FGSM.

Εφαρμόζοντας τα βήματα που περιγράφονται στην αντίστοιχη ενότητα στο Κεφάλαιο 4, από τα αντιφατικά παραδείγματα εξάγαμε αντιφατικές εικόνες που έχουν υποστεί την τεχνική PCA-whitening (Σχήματα 5.14,5.15,5.16). Για να αποφανθούμε εάν ένα δείγμα κρίνεται αντιφατικό ή όχι μέσω της τεχνικής αυτής χρειάζεται να ορίσουμε ένα όριο κύριων συνιστωσών πάνω από το οποίο οι συντελεστές διακύμανσης των αντιφατικών παραδειγμάτων διαφέρουν κατα πολύ από αυτούς των κανονικών

Σχήμα 5.16: «Μη ισχυρά» αντιφατικά παραδείγματα πριν (δεύτερη σειρά) και μετά (πρώτη σειρά) από την εφαρμογή της PCA-whitening μεθόδου. Σημειώνεται ότι τα δείγματα αυτά έχουν προκύψει από διπλή επίθεση.

δειγμάτων. Έπειτα από οπτική παρατήρηση των Σχημάτων 5.17,5.18,5.19 και από ορισμένες δοκιμές ορίων, καταλήξαμε ότι στην περίπτωση του συνόλου δεδομένων MNIST, από την 900^{στή} κύρια συνιστώσα και μετά ο διαχωρισμός των αντιφατικών δειγμάτων από τις καλοήθεις εικόνες είναι διακριτός σε μεγάλο ποσοστό και για τα 3 σενάρια επιθέσεων που εκτελέσαμε. Στην περίπτωση του συνόλου δεδομένων kaggle CIFAR10, το όριο αυτό παρουσιάζει μια μικρή διαφορά μεταξύ των «ισχυρών» και «μη ισχυρών» αντιφατικών παραδειγμάτων. Πιο συγκεκριμένα, για τα «ισχυρά» αντιφατικά παραδείγματα που προέκυψαν από τις επιθέσεις FGSM και one-pixel το όριο τέθηκε στην 2700^{στή} κύρια συνιστώσα ενώ για τα αντίστοιχα «μη ισχυρά» το όριο τέθηκε στην 2500^{στή} κύρια συνιστώσα. Στην περίπτωση των αντιφατικών παραδειγμάτων που προέκυψαν μέσω της διπλής επίθεσης, το όριο αυτό είναι κοινό για τα «ισχυρά» και «μη» δείγματα και ορίζεται από την 2700^{στή} κύρια συνιστώσα.

$$\begin{aligned}
\text{Precision} &= \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \\
\text{Recall} &= \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \\
\text{Specificity} &= \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \\
\text{False Positive Rate} &= \frac{\text{False Positive}}{\text{True Negative} + \text{False Positive}} = 1 - \text{Specificity} \\
\text{Accuracy} &= \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Negative} + \text{True Negative} + \text{False Positive}}
\end{aligned} \tag{5.2}$$

Για την αξιολόγηση της τεχνικής αυτής χρησιμοποιήθηκαν η *καμπύλη λειτουργικών χαρακτηριστικών* (receiver operating characteristic - ROC) καθώς και η *εκτίμηση περιοχής κάτω από την καμπύλη* (area under the curve - AUC). Η ROC είναι μια μέτρηση αξιολόγησης για δυαδικά προβλήματα ταξινόμησης. Πρόκειται για μια καμπύλη πιθανότητας που σχεδιάζει την *εξειδίκευση* (specificity) ή αλλιώς *αληθώς θετική αναλογία* (true positive rate - TPR) έναντι της *ψευδούς θετικής αναλογίας* (false positive rate -FPR), όπως ορίζονται στην Εξίσωση 5.2, σε διάφορες τιμές κατωφλίου και ουσιαστικά διαχωρίζει το «σήμα» από το «θόρυβο». Η AUC είναι το μέτρο της ικανότητας ενός ταξινομητή να διακρίνει μεταξύ των κλάσεων και χρησιμοποιείται ως σύνοψη της καμπύλης ROC. Όσο υψηλότερη είναι η AUC, τόσο καλύτερη είναι η απόδοση του μοντέλου στη διάκριση μεταξύ θετικών και αρνητικών κλάσεων. Ο *μέσος όρος της πιστότητας* (average precision score - AP) υπολογίζεται ως η μέση τιμή της *πιστότητας* (precision) σε όλες τις κλάσεις και της συνολικής επικάλυψης που παρουσιάζουν οι κλάσεις στην ένωσή τους. Επιπλέον, για σύγκριση των αποτελεσμάτων με τις δύο επόμενες τεχνικές ανίχνευσης που εφαρμόσαμε και αναλύονται στην πορεία, κρίθηκε χρήσιμος ο υπολογισμός της

ακρίβειας (accuracy) του ανιχνευτή, αλλά και των επιμέρους ποσοστών επιτυχίας πάνω στα αντιφατικά παραδείγματα (TNR), και πάνω στα καλοήθη δείγματα ξεχωριστά (TPR/Recall), μέσα από τις Εξισώσεις 5.2. Τα αποτελέσματα από αυτές τις μετρικές παρουσιάζονται στον Πίνακα 5.2.

			AUROC	AUPR	TNR	TPR	Accuracy
MNIST	One-pixel attack	”Strong”	100%	100%	100%	97,06%	98,53%
		”Not so strong”	100%	100%	100%	91,80%	95,90%
	FGSM attack	”Strong”	100%	100%	100%	96,23%	98,11%
		”Not so strong”	100%	100%	100%	93,15%	96,58
	Double attack	”Strong”	100%	100%	100%	100%	**100**%
		”Not so strong”	100%	100%	100%	93,41%	**96**, **71**%
Kaggle CIFAR10	One-pixel attack	”Strong”	87,08%	87,30%	94,12%	50,42%	72,27%
		”Not so strong”	77,44%	80,17%	83,06%	46,77%	64,92%
	FGSM attack	”Strong”	79,34%	67,69%	100%	47,83%	73,91%
		”Not so strong”	79,78%	69,34%	100%	62,96%	81,48%
	Double attack	”Strong”	95,76%	95,38%	100%	55,92%	**77**, **96**%
		”Not so strong”	95,53%	95,22%	100%	61,54%	**80**, **77**%

Πίνακας 5.2: Μετρικές αξιολόγησης του ανιχνευτή αντιφατικών παραδειγμάτων PCA-whitening,

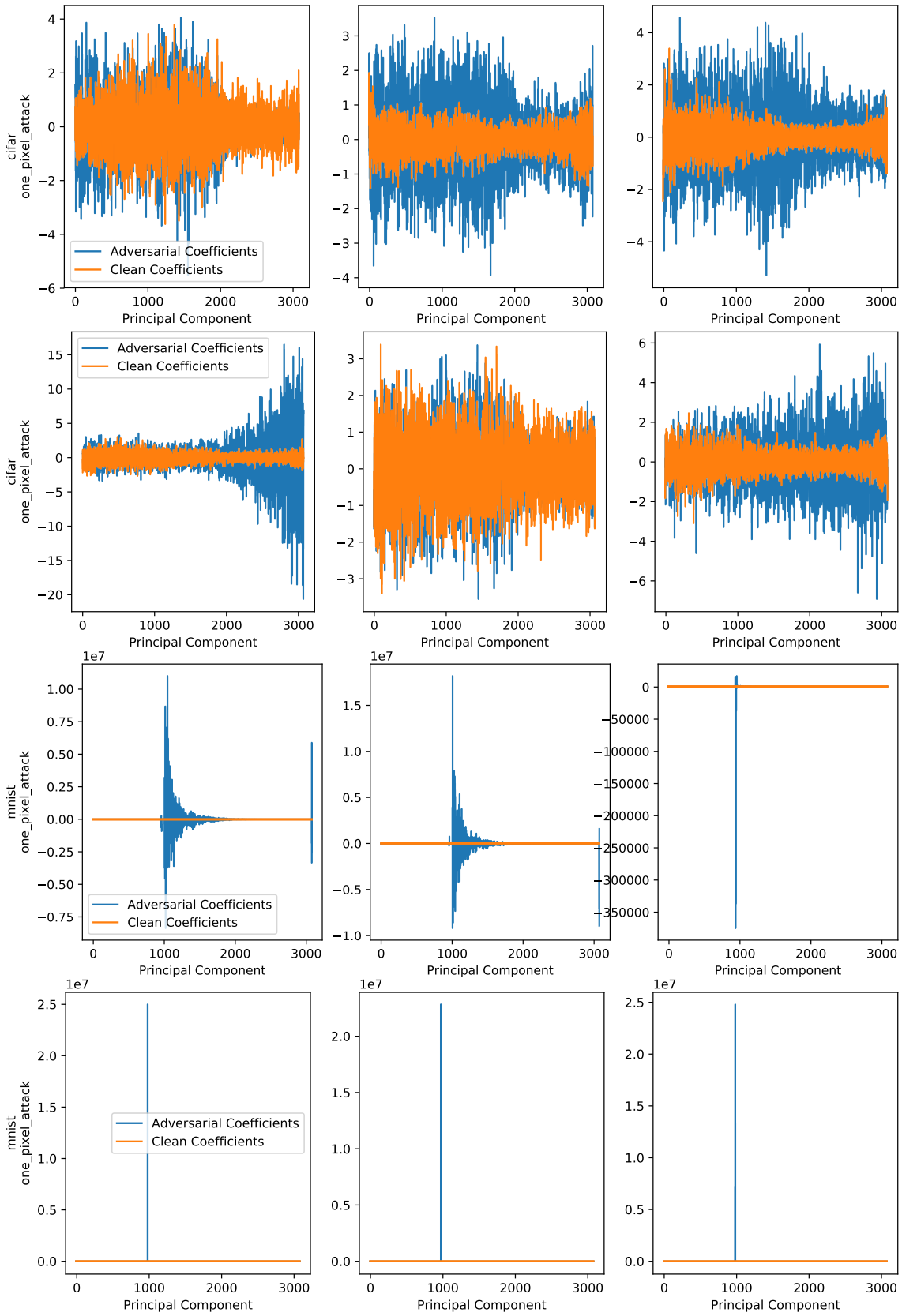
Παρατηρώντας τις τιμές της ακρίβειας του ανιχνευτή στον Πίνακα 5.2, καταλήγουμε στα εξής συμπεράσματα. Αρχικά και οι δύο επιθέσεις φαίνεται ότι μπορούν να εντοπιστούν επιτυχώς σε μεγάλο ποσοστό από τον συγκεκριμένο ανιχνευτή, με την επίθεση FGSM να τείνει σε ορισμένες περιπτώσεις να διευκολύνει ακόμα περισσότερο το έργο της PCA-whitening. Μέσω της παραπάνω παρατήρησης διαπιστώνουμε ότι η επίθεση one-pixel, αν και εκ πρώτης πρόκειται για μια απλή επίθεση, αποτελεί μια βάσιμη απειλή για τα ΒΝΔ, συγκρίσιμη με την επίθεση FGSM (που αποτελεί σημείο αναφοράς). Επιπλέον, μέσω των τιμών του Πίνακα 5.2, συμπεραίνουμε ότι η διπλή επίθεση δημιουργεί αντιφατικά παραδείγματα τα οποία εντοπίζονται με μεγαλύτερη ευκολία από τον ανιχνευτή PCA-whitening από ότι τα παραδείγματα που προκύπτουν από τις αντίστοιχες μεμονωμένες επιθέσεις.

5.5.2 Τεχνική ανίχνευσης Squeezing color bits

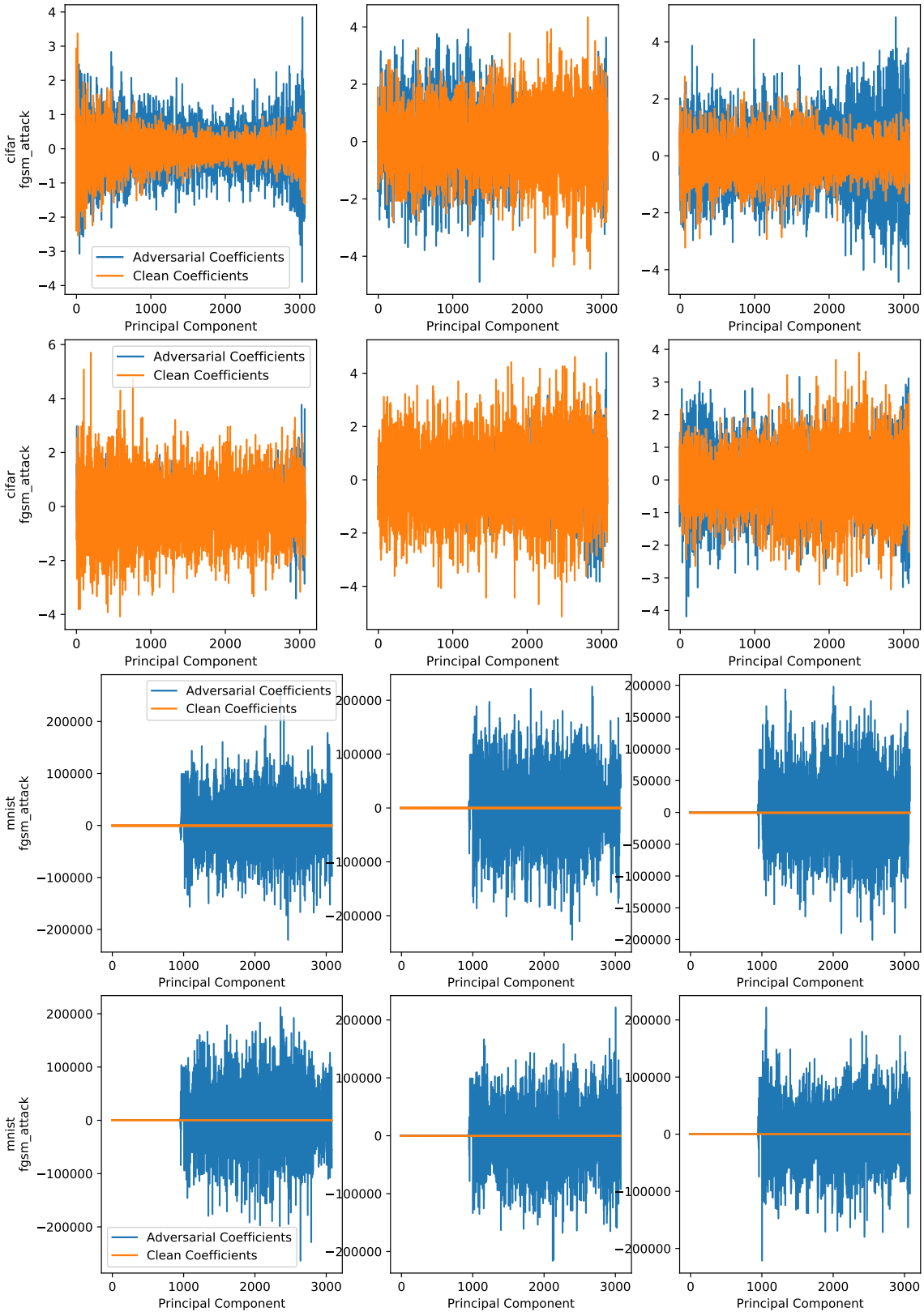
Μια άλλη τεχνική ανίχνευσης που είδαμε στο τέλος του Κεφαλαίου 4 είναι η squeezing color bits. Εφαρμόσαμε τη μέθοδο αυτή για να εξετάσουμε τη σχέση που αναπτύσσεται μεταξύ του ανιχνευτή αυτού και των 3 επιθέσεων που εκτελέσαμε σε προηγούμενη φάση, διεξάγοντας συμπεράσματα τόσο για την αποτελεσματικότητα του ανιχνευτή όσο και για την ισχύ των επιθέσεων. Σε πρώτο στάδιο, κληθήκαμε να επιλέξουμε τον κατάλληλο αριθμό των bits, σύμφωνα με τον οποίο αν συρρικνώσουμε το βάθος χρώματος κάθε καναλιού της RGB εικόνας σε αυτό το όριο, ο ανιχνευτής θα μπορέσει να διακρίνει με μεγαλύτερη ακρίβεια τα κακόβουλα από τα καλοήθη παραδείγματα. Για το σκοπό αυτό δοκιμάσαμε όλα τα πιθανά βάθη χρώματος, από 1 έως 8 bits, για τα αντιφατικά παραδείγματα κάθε επίθεσης και υπολογίσαμε την πιστότητα, το TNR και την *ανάκληση* (recall) για κάθε όριο συρρίκνωσης. Από τα αποτελέσματα που πήραμε, οι τιμές των οποίων παραθέτονται στον Πίνακα 5.4, καταλήξαμε στα εξής όρια συρρίκνωσης των εικόνων:

1. Στην περίπτωση της συλλογής δεδομένων kaggle CIFAR10, για τα αντιφατικά παραδείγματα που προέκυψαν από την επίθεση one-pixel, το όριο επιλέχθηκε στα 3 bits, ενώ για τα αντιφατικά παραδείγματα που προέκυψαν από τις άλλες 2 επιθέσεις θεωρήθηκε καταλληλότερο το όριο των 2 bits.

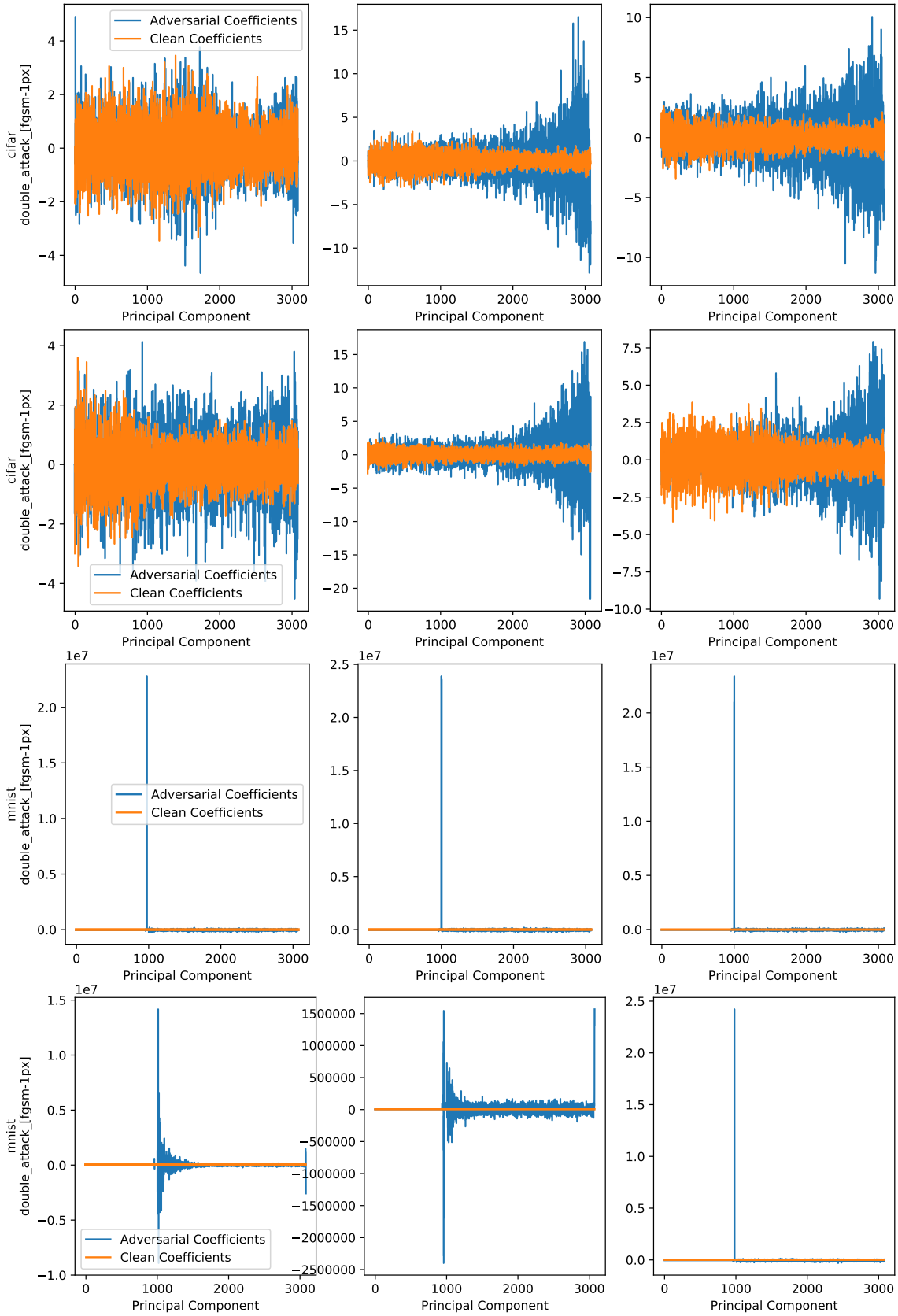
2. Στην περίπτωση της συλλογής δεδομένων MNIST, για τα αντιφατικά παραδείγματα που προέκυψαν από την επίθεση FGSM, το όριο επιλέχθηκε στα 2 bits, ενώ για τα αντιφατικά παραδείγματα που προέκυψαν από τις άλλες 2 επιθέσεις θεωρήθηκε καταλληλότερο το όριο του 1 bit.



Σχήμα 5.17: Διαγράμματα συντελεστών διακύμανσης «ισχυρών» (γραμμές 2,4) και «μη» (γραμμές 1,3) αντιφατικών παραδειγμάτων από την επίθεση one-pixel.



Σχήμα 5.18: Διαγράμματα συντελεστών διακύμανσης «ισχυρών» (γραμμές 2,4) και «μη» (γραμμές 1,3) αντιφατικών παραδειγμάτων από την επίθεση FGSM.



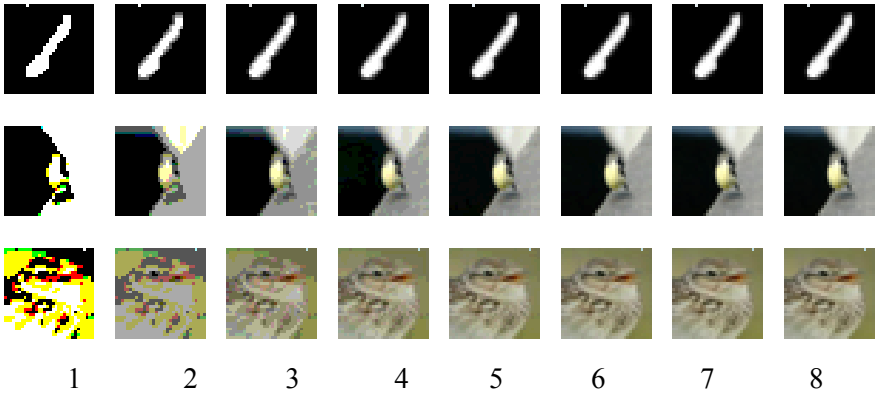
Σχήμα 5.19: Διαγράμματα συντελεστών διακύμανσης «ισχυρών» (γραμμές 2,4) και «μη» (γραμμές 1,3) αντιφατικών παραδειγμάτων από τη διπλή επίθεση.

Στην υλοποίηση μας, ο ανιχνευτής κατηγοριοποιεί μια εικόνα ως αντιφατική ή όχι με τον εξής τρόπο:

1. Αρχικά, ο ανιχνευτής εισάγει την υπό διερεύνηση εικόνα στον ταξινομητή VGG16 και αποθηκεύει το αποτέλεσμα της πρόβλεψής του.

2. Στη συνέχεια,, φιλτράρει την εικόνα αυτή συρρικνώνοντας το βάθος χρώματος κάθε καναλιού της, σύμφωνα με το όριο που τέθηκε παραπάνω.

3. Τέλος, η φιλτραρισμένη εικόνα διέρχεται ξανά από τον ταξινομητή του μοντέλου μας. Εάν η πρόβλεψη του ταξινομητή είναι η ίδια με την αρχική, τότε το δείγμα θεωρείται καλοήθες. Σε κάθε άλλη περίπτωση η εικόνα κατατάσσεται στα αντιφατικά παραδείγματα.



Σχήμα 5.20: Ενδεικτικά παραδείγματα έπειτα από την εφαρμογή του φίλτρου squeezing color bits σε δεδομένα από το MNIST και το kaggle CIFAR10. Στη δεξιά στήλη, χρησιμοποιούμε 8 bits για τη συρρίκνωση του βάθους χρώματος κάθε καναλιού της RGB εικόνας και κινούμενοι προς τα αριστερά κατεβαίνουμε κατά 1 bit. Στην ουσία η δεξιά στήλη παρουσιάζει τις αρχικές εικόνες καθώς το βάθος του χρώματός τους δεν έχει υποστεί κάποια αλλαγή. Τα αντιφατικά δείγματα στην πρώτη σειρά έχουν προκύψει από εφαρμογή της επίθεσης one-pixel, στη δεύτερη σειρά από εφαρμογή της επίθεσης FGSM και στην τρίτη από εκτέλεση της διπλής επίθεσης.

Στη βιβλιογραφία σημειώνεται ότι ορισμένες φορές, η πρόβλεψη του ταξινομητή πάνω στη φιλτραρισμένη εικόνα μπορεί να επιστρέψει την πραγματική κλάση του αντιφατικού παραδείγματος. Για το λόγο αυτό, υπολογίσαμε και το ποσοστό επιτυχίας εύρεσης της σωστής κλάσης πάνω στα αντιφατικά παραδείγματα και πάνω στις κανονικές εικόνες που παρουσιάζει ο ανιχνευτής για τις διάφορες τιμές ορίων των bits, πληροφορία η οποία παρατίθεται αναλυτικά στο Παράρτημα Α.2. Στο Σχήμα 5.20 παρουσιάζονται ενδεικτικά ορισμένες εικόνες που προέκυψαν από την εφαρμογή της μεθόδου αυτής. Σημειώνεται ότι στα πειράματα της μεθόδου squeezing color bits, όπως και στα μετέπειτα του διπλού ανιχνευτή, χρησιμοποιήθηκαν από κάθε επίθεση μόνο τα «ισχυρά» αντιφατικά παραδείγματα.

Συγκρίνοντας τους Πίνακες 5.2 και 5.4 παρατηρούμε ότι η μέθοδος PCA-whitening φαίνεται πιο αποτελεσματική στην ανίχνευση αντιφατικών παραδειγμάτων από ότι η squeezing color bits, όσον αφορά τις επιθέσεις FGSM και one-pixel, καθώς η ακρίβεια είναι αισθητά υψηλότερα στην πρώτη περίπτωση. Απομονώνοντας τον Πίνακα 5.4, παρατηρούμε ότι τα αντιφατικά παραδείγματα που έχουν δημιουργηθεί από τις επιθέσεις που δέχθηκαν δείγματα της συλλογής δεδομένων MNIST μπορούν να εντοπιστούν επιτυχώς και στις δύο επιθέσεις σε ίσο και σχετικά υψηλό ποσοστό από τον ανιχνευτή squeezing color bits. Στην περίπτωση της διπλής επίθεσης, επαναλαμβάνεται η συμπεριφορά αύξησης της ακρίβειας του ανιχνευτή που παρατηρήσαμε με την μέθοδο PCA-whitening. Οι παραπάνω παρατηρήσεις ωστόσο δεν ισχύουν και στην περίπτωση που τα αντιφατικά παραδείγματα προέρχονται από δείγματα της συλλογής δεδομένων kaggle CIFAR10. Εφαρμόζοντας την επίθεση one-pixel σε δείγματα από το kaggle CIFAR10, φαίνεται να διευκολύνεται ελαφρώς η ανίχνευση των αντιφατικών παραδειγμάτων μέσω της μεθόδου squeezing color bits σε σύγκριση πάντα με την επίθεση FGSM,

ενώ στην περίπτωση της διπλής επίθεσης το ποσοστό επιτυχίας του δεδομένου ανιχνευτή συγκλίνει σε αυτό της μεμονωμένης επίθεσης FGSM.

5.5.3 Εφαρμογή διπλής ανίχνευσης

Ολοκληρώνοντας τη μελέτη μας θελήσαμε - και για λόγους συμμετρίας - να υλοποιήσουμε μια διπλή στρατηγική ανίχνευσης αντιφατικών παραδειγμάτων συνδυάζοντας τις τεχνικές PCA-whitening και squeezing color bits. Τα «ισχυρά» αντιφατικά παραδείγματα που δημιουργήσαμε μέσω των 3 επιθέσεων υποβλήθηκαν σε αυτή τη συνδυαστική μέθοδο. Πιο συγκεκριμένα, η διπλή μέθοδος ανίχνευσης αντιφατικών παραδειγμάτων εφαρμόστηκε με δύο προσεγγίσεις. Κάθε δείγμα προς διερεύνηση εισάγεται μεμονωμένα και στους δύο ανιχνευτές. Στη συνέχεια, για να αποφανθούμε εάν το δείγμα είναι αντιφατικό ή όχι:

1. Ελέγχουμε αν έστω και ένας από τους δύο ανιχνευτές (OR) έχει ταξινομήσει την εικόνα ως αντιφατική. Εάν η εικόνα θεωρείται και από τους δύο ανιχνευτές καλοήθης, τότε χαρακτηρίζεται καλοήθης και από τη συνδυαστική μέθοδο. Σε αντίθετη περίπτωση το δείγμα κατατάσσεται στα αντιφατικά παραδείγματα.

2. Ελέγχουμε αν και οι δύο ανιχνευτές (AND) έχουν ταξινομήσει την εικόνα ως αντιφατική. Εάν έστω και ένας ανιχνευτής θεωρεί ότι η εικόνα δεν είναι αντιφατική, τότε το δείγμα κατατάσσεται στα καλοήθη παραδείγματα.

Τα αποτελέσματα των μετρικών που χρησιμοποιήθηκαν για την αξιολόγηση της μεθόδου παρουσιάζονται αναλυτικά στον Πίνακα 5.3. Παρατηρούμε ότι στη γενική περίπτωση η δεύτερη προσέγγιση παρουσιάζει υψηλότερα ποσοστά ακρίβειας, πλην της επίθεσης FGSM πάνω σε δείγματα της συλλογής δεδομένων MNIST, όπου η πρώτη προσέγγιση αυξάνει σημαντικά το ποσοστό επιτυχίας συγκριτικά με τη δεύτερη. Σε κάθε περίπτωση τα ποσοστά επιτυχίας της διπλής ανίχνευσης είναι μεγαλύτερα σε σχέση με αυτά της μεμονωμένης ανίχνευσης squeezing color bits, αλλά αισθητά μικρότερα σε σχέση με αυτά του μεμονωμένου ανιχνευτή PCA-whitening.

			TNR	TPR	Accuracy
MNIST	One-pixel attack	AND	75%	94,12%	**84**, **56**%
		OR	100%	61,76%	80,88%
	FGSM attack	AND	58,49%	100%	79,25%
		OR	100%	86,79%	**93**, **40**%
	Double attack	AND	100%	100%	**100**%
		OR	100%	100%	**100**%
Kaggle CIFAR10	One-pixel attack	AND	45,38%	95,80%	**70**, **59**%
		OR	97,06%	39,92%	68,49%
	FGSM attack	AND	67,93%	76,63%	**72**, **28**%
		OR	99,46%	32,61%	66,03%
	Double attack	AND	65,13%	80,92%	**73**, **03**%
		OR	100%	25,66%	62,83%

Πίνακας 5.3: Μετρικές αξιολόγησης της διπλής συνδυαστικής μεθόδου ανίχνευσης αντιφατικών παραδειγμάτων,

			TNR	TPR	Accuracy
MNIST	One-pixel attack	**using 1 bit**	**75%**	**73, 53%**	**74, 26%**
		using 2 bits	41,18%	97,06%	69,12%
		using 3 bits	2,94%	98,53%	50,74%
		using 4 bits	0,0%	98,53%	49,26%
		using 5 bits	0,0%	100%	50%
		using 6 bits	0,0%	100%	50%
		using 7 bits	0,0%	100%	50%
		using 8 bits	0,0%	100%	50%
	FGSM attack	using 1 bit	83,02%	60,38%	71,70%
		using 2 bits	**58, 49%**	**92, 45%**	**75, 47%**
		using 3 bits	50,94%	98,11%	74,53%
		using 4 bits	20,75%	98,11%	59,43%
		using 5 bits	16,98%	100%	58,49%
		using 6 bits	9,43%	100%	54,72%
		using 7 bits	9,43%	100%	54,72%
		using 8 bits	0,0%	100%	50%
	Double attack	**using 1 bit**	**100%**	**70%**	**85%**
		using 2 bits	70%	90%	80%
		using 3 bits	30%	100%	65%
		using 4 bits	30%	100%	65%
		using 5 bits	20%	100%	60%
		using 6 bits	20%	100%	60%
		using 7 bits	10%	100%	55%
		using 8 bits	0,0%	100%	50%
Kaggle CIFAR10	One-pixel attack	using 1 bit	81,93%	34,87%	58,40%
		using 2 bits	67,23%	52,10%	59,67%
		using 3 bits	**50%**	**82, 77%**	**66, 39%**
		using 4 bits	34,03%	95,38%	64,71%
		using 5 bits	11,34%	100%	55,67%
		using 6 bits	2,52%	100%	51,26%
		using 7 bits	0,42%	99,58%	50%
		using 8 bits	0,0%	100%	50%
	FGSM attack	using 1 bit	85,87%	30,43%	58,15%
		using 2 bits	**68, 48%**	**51, 63%**	**60, 05%**
		using 3 bits	41,85%	78,80%	60,33%
		using 4 bits	16,85%	92,93%	54,89%
		using 5 bits	1,63%	96,74%	49,18%
		using 6 bits	1,09%	98,91%	50%
		using 7 bits	0,54%	100%	50,27%
		using 8 bits	0,0%	100%	50%
	Double attack	using 1 bit	90,13%	36,84%	63,49%
		using 2 bits	**65, 13%**	**54, 61%**	**59, 87%**
		using 3 bits	33,55%	86,18%	59,87%
		using 4 bits	11,18%	96,05%	53,62%
		using 5 bits	97,37%	2,63%	50%
		using 6 bits	1,32%	98,03%	49,67%
		using 7 bits	0,0%	100%	50%
		using 8 bits	0,0%	100%	50%

Πίνακας 5.4: Μετρικές αξιολόγησης του ανιχνευτή αντιφατικών παραδειγμάτων squeezing color bits.

Κεφάλαιο 6

Επίλογος

6.1 Ανακεφαλαίωση

Κατά την εκπόνηση της παρούσας διπλωματικής εργασίας, η έρευνα και η μελέτη μας επικεντρώθηκε σε πρώτη φάση στη διερεύνηση της επίθεσης one-pixel καθώς και στο πόσο ανθεκτική είναι απέναντι στις τεχνικές ανίχνευσης αντιφατικών παραδειγμάτων PCA-whitening και squeezing color bits. Σε δεύτερη φάση μελετήσαμε την ισχύ και την αποτελεσματικότητα μιας συνδυαστικής μεθόδου επίθεσης χρησιμοποιώντας τις μεθόδους επίθεσης one-pixel και FGSM. Τέλος, αποφασίσαμε να επεκτείνουμε την τελευταία σκέψη και στον τομέα ανίχνευσης των αντιφατικών παραδειγμάτων εφαρμόζοντας έναν νέο ανιχνευτή που είχε προκύψει από τον συνδυασμό των τεχνικών PCA-whitening και squeezing color bits. Από την παραπάνω έρευνα διεξάγαμε ορισμένα συμπεράσματα τα οποία αναφέρονται σε προηγούμενες ενότητες. Στη συνέχεια θα παρουσιάσουμε συγκεντρωτικά τα βασικά αποτελέσματα στα οποία καταλήξαμε έπειτα από ενδελεχή μελέτη και ανάλυση των πειραμάτων μας.

6.2 Βασικά συμπεράσματα

Βασικός στόχος της παρούσας διπλωματικής υπήρξε η μελέτη και περαιτέρω διερεύνηση της επίθεσης one-pixel, μιας ιδιαίτερα περιορισμένης διαταραχής που αφορά ένα μόνο εικονοστοιχείο της αρχικής εικόνας. Τα αποτελέσματα της μεθόδου αυτής ξεπέρασαν τις προσδοκίες μας καθώς παρατηρήσαμε τα εξής:

- Εφαρμόζοντας την επίθεση one-pixel δημιουργήσαμε «ισχυρά και μη» αντιφατικά παραδείγματα με την ίδια ευκολία που δημιουργήσαμε τα αντίστοιχα μέσω της επίθεση FGSM, καθώς διαπιστώσαμε ότι στο ίδιο πλήθος αρχικών εικόνων καταφέραμε να δημιουργήσουμε ελαφρώς περισσότερα αντιφατικά παραδείγματα με την πρώτη μέθοδο από ότι με τη δεύτερη. Φυσικά, θα πρέπει να λάβουμε υπόψιν μας ότι με την εύρεση βέλτιστων παραμέτρων και για τις δύο μεθόδους πιθανότατα αυτή η αναλογία στο πλήθος των επιτυχημένων αντιφατικών παραδειγμάτων να μεταβληθεί, ωστόσο πρόκειται για δύο επιθέσεις με συγκρίσιμα αποτελέσματα και επιβεβαιωμένη επιτυχία.

- Επιπλέον, συγκρίνοντας τα αποτελέσματα της επίθεση one-pixel στα δύο σύνολα δεδομένων που χρησιμοποιήσαμε, διαπιστώσαμε ότι οι πιο περίπλοκες εικόνες, όπως αυτές του kaggle CIFAR10, είναι πιο αποτελεσματικοίστόχοι. Καταλήξαμε σε αυτό το συμπέρασμα παρατηρώντας ότι για τη δημιουργία ίδιου αριθμού αντιφατικών παραδειγμάτων χρειάστηκαν περίπου 1.000 αρχικές εικόνες από το σύνολο δεδομένων του kaggle CIFAR10 και 10.000 αρχικές εικόνες από το σύνολο δεδομένων του MNIST.

- Οι δυσκολίες που αντιμετωπίσαμε στα πλαίσια της διπλής επίθεσης μας οδηγούν στο συμπέρασμα ότι η από κοινού εφαρμογή δύο επιτυχημένων μεθόδων δημιουργίας αντιφατικών παραδειγμάτων δεν συνεπάγεται κατ' ανάγκη ισχυρότερα αντιφατικά παραδείγματα, καθώς οι δύο μέθοδοι δεν δρουν απαραίτητα αθροιστικά. Χαρακτηριστικά στη δική μας πειραματική διαδικασία, οι μεμονωμένα επιτυχημένες επιθέσεις one-pixel και FGSM φαίνεται ότι αλληλοαναιρούνται όταν εφαρμόζονται παράλληλα και έτσι η διαδικασία δημιουργίας αντιφατικών

παραδειγμάτων δυσκολεύει σε μεγάλο βαθμό. Πιο συγκεκριμένα, στην περίπτωση που εφαρμόσαμε στα δείγματα-στόχο πρώτα την επίθεση one-pixel και έπειτα την FGSM, τα επιτυχημένα αντιφατικά παραδείγματα που λάβαμε ήταν ελάχιστα σε μια συνολικά πολύ χρονοβόρα σε επίπεδο υπολογιστικών πόρων διαδικασία, γεγονός που κατέστησε απαγορευτικό τον παραπάνω συνδυασμό των επιθέσεων. Όταν αντιστρέψαμε τον συνδυασμό των μεθόδων επιθέσεων και εφαρμόσαμε πρώτα την επίθεση FGSM και στη συνέχεια την one-pixel, τα αποτελέσματα ήταν πιο ενθαρρυντικά καθώς λάβαμε περισσότερα επιτυχημένα αντιφατικά παραδείγματα από ότι στην πρώτη περίπτωση. Ωστόσο, σε κάθε περίπτωση το υπολογιστικό κόστος για την εφαρμογή μιας συνδυαστικής επίθεσης με τις δεδομένες μεθόδους είναι υψηλό και ο χρόνος που απαιτείται πολύς. Κατά συνέπεια μια διπλή επίθεση τέτοιου τύπου κρίνεται μάλλον ασύμφορη από την πλευρά του εισβολέα, αν αναλογιστούμε κιόλας την ευκολία ανίχνευσης των αντιφατικών παραδειγμάτων - που δημιουργήθηκαν μέσω της διπλής επίθεσης - που παρουσίασαν κατά την εφαρμογή των μεθόδων PCA-whitening, squeezing color bits και του συνδυασμού τους.

Ένας άλλος κύριος πυρήνας της δεδομένης διπλωματικής εργασίας υπήρξε η ανάπτυξη και μελέτη δύο τεχνικών ανίχνευσης αντιφατικών παραδειγμάτων καθώς και του συνδυασμού τους. Βασικό κίνητρο της περαιτέρω αυτής διερεύνησης αποτέλεσε το ερώτημα του πόσο εύκολα μπορεί να ανιχνευτεί μια επίθεση one-pixel. Από τα αποτελέσματα που λάβαμε καταλήξαμε στα εξής συμπεράσματα:

- Η τεχνική PCA-whitening μπορεί με ευκολία να ανιχνεύει επιτυχώς αντιφατικά παραδείγματα που έχουν δημιουργηθεί μέσω της επίθεσης one-pixel, FGSM ή του συνδυασμού τους. Παρατηρούμε ότι στις πιο απλές εικόνες όπως αυτές από το σύνολο δεδομένων του MNIST , η ανίχνευση των αντιφατικών εικόνων ξεπερνά το ποσοστό επιτυχίας του 95%. Ο ανιχνευτής PCA-whitening παρουσιάζει τα υψηλότερα ποσοστά επιτυχίας σε κάθε μία από τις τρεις περιπτώσεις επίθεσης, σε σύγκριση με την squeezing color bits και την συνδυαστική μέθοδο ανίχνευσης που εφαρμόστηκε. Πιο συγκεκριμένα, τα αντιφατικά παραδείγματα που έχουν δημιουργηθεί μέσω της διπλής επίθεσης φαίνεται ότι είναι πιο εύκολο να εντοπιστούν από τον ανιχνευτή PCA-whitening, σε σύγκριση με αυτά των άλλων δύο επιθέσεων, τα οποία παρουσιάζουν παρόμοια ποσοστά εντοπισμού από το δεδομένο ανιχνευτή. Σημειώνεται ότι στην περίπτωση των αντιφατικών παραδειγμάτων που έχουν δημιουργηθεί από την επίθεση FGSM στο σύνολο δεδομένων του kaggle CIFAR10, ο ανιχνευτής φαίνεται να εντοπίζει με μεγαλύτερη επιτυχία τα αντιφατικά παραδείγματα σε σχέση με την περίπτωση που έχει εφαρμοστεί η επίθεση one-pixel.

- Σε αντίθεση με τον ανιχνευτή PCA-whitening, ο squeezing color bits φαίνεται να εντοπίζει με μεγαλύτερη επιτυχία τα αντιφατικά παραδείγματα που έχουν δημιουργηθεί από την επίθεση one-pixel στο σύνολο δεδομένων του kaggle CIFAR10, σε σύγκριση με αυτά που έχουν δημιουργηθεί από την επίθεση FGSM. Αυτή η αντίφαση παρατηρείται και στην περίπτωση της διπλής επίθεσης σε δεδομένα του kaggle CIFAR10, όπου ο ανιχνευτής φαίνεται να δυσκολεύεται αρκετά να εντοπίσει τα συγκεκριμένα αντιφατικά παραδείγματα. Η κατάσταση αυτή αντιστρέφεται στην περίπτωση της συλλογής δεδομένων MNIST, όπου ο ανιχνευτής φαίνεται να εντοπίζει με μεγαλύτερη επιτυχία τα δείγματα που έχουν δημιουργηθεί από τη διπλή επίθεση, ενώ οι άλλες δύο επιθέσεις παρουσιάζουν παρόμοια ποσοστά επιτυχούς ανίχνευσης.

- Τέλος, εφαρμόζοντας τις συνδυαστικές μεθόδους ανίχνευσης παρατηρήσαμε ότι στη γενική περίπτωση η AND προσέγγιση παρουσιάζει υψηλότερα ποσοστά επιτυχίας, με εξαίρεση την επίθεση FGSM πάνω σε δείγματα της συλλογής δεδομένων MNIST, όπου η OR προσέγγιση σημείωσε σημαντικά αυξημένη ακρίβεια συγκριτικά με την πρώτη. Ωστόσο, σε όλες τις περιπτώσεις η ακρίβεια της διπλής ανίχνευσης είναι μεγαλύτερη σε σχέση με αυτά της μεμονωμένης ανίχνευσης squeezing color bits, αλλά αισθητά μικρότερη σε σχέση με αυτά του μεμονωμένου ανίχνευσης PCA-whitening. Από την τελευταία παρατήρηση συμπεραίνουμε ότι η συνδυαστική μέθοδος ανίχνευσης που προέκυψε από τις δεδομένες τεχνικές ανίχνευσης δεν μπορεί να αυξήσει την αποτελεσματικότητα του βέλτιστου ανιχνευτή, καθώς όπως φαίνεται η τεχνική διπλής

ανίχνευσης υπολογίζει περίπου έναν μέσο όρο των ποσοστών επιτυχίας των μεμονωμένων τεχνικών ανίχνευσης.

6.3 Μελλοντικές κατευθύνσεις έρευνας

Στη συνέχεια παραθέτουμε ορισμένες ιδέες για μελλοντικές εργασίες και επεκτάσεις γύρω από το αντικείμενο που αναλύσαμε στην παρούσα διπλωματική.

1. Στο πλαίσιο της διπλής επίθεσης, η εφαρμογή της δεύτερης επίθεσης στο σύνολο των επιτυχημένων και αποτυχημένων αντιφατικών παραδειγμάτων που έχουν παραχθεί από την πρώτη επίθεση.

2. Επέκταση της μεθόδου squeezing color bits και της διπλής τεχνικής ανίχνευσης χρησιμοποιώντας το σύνολο των επιτυχημένων αντιφατικών παραδειγμάτων των επιθέσεων και όχι μόνο των «ισχυρών».

3. Εκτέλεση της παραπάνω έρευνας πάνω σε στοχευμένες επιθέσεις.

4. Εφαρμογή διαφορετικού ζεύγους μεθόδων στη διπλή επίθεση και ομοίως στο διπλό ανιχνευτή.

5. Εφαρμογή πολλαπλών συνδυαστικών μεθόδων επιθέσεων και ομοίως πολλαπλών συνδυαστικών τεχνικών ανίχνευσης αντιφατικών παραδειγμάτων.

6. Χρήση συνόλων δεδομένων με μεγαλύτερες εικόνες όπως αυτές του ImageNet για τη δημιουργία αντιφατικών παραδειγμάτων.

Βιβλιογραφία

[Alza19] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh and Mani Srivastava, "GenAttack: Practical Black-box Attacks with Gradient-Free Optimization", 2019.

[Atha18a] Anish Athalye, Nicholas Carlini and David Wagner, "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples", 2018.

[Atha18b] Anish Athalye, Logan Engstrom, Andrew Ilyas and Kevin Kwok, "Synthesizing Robust Adversarial Examples", 2018.

[Ayub20] Md. Ahsan Ayub, William A. Johnson, Douglas A. Talbert and Ambareen Siraj, "Model Evasion Attack on Intrusion Detection Systems using Adversarial Machine Learning", in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, 2020.

[Bail01] C. Bailer-Jones, Reeta Gupta and Harinder Singh, "An introduction to artificial neural networks", 03 2001.

[Barr10] Marco Barreno, Blaine Nelson, Anthony D. Joseph and J. D. Tygar, "The security of machine learning", *Machine Learning*, vol. 81, no. 2, p. 121–148, Nov 2010.

[Bigg13a] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto and Fabio Roli, "Evasion Attacks against Machine Learning at Test Time", *Lecture Notes in Computer Science*, p. 387–402, 2013.

[Bigg13b] Battista Biggio, Blaine Nelson and Pavel Laskov, "Poisoning Attacks against Support Vector Machines", 2013.

[Buck18] Jacob Buckman, Aurko Roy, Colin Raffel and Ian Goodfellow, "Thermometer Encoding: One Hot Way To Resist Adversarial Examples", in *International Conference on Learning Representations*, 2018.

[Carl17a] Nicholas Carlini and David Wagner, "Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods", 2017.

[Carl17b] Nicholas Carlini and David Wagner, "Towards Evaluating the Robustness of Neural Networks", pp. 39–57, 05 2017.

[Carl18] Nicholas Carlini, Guy Katz, Clark Barrett and David L. Dill, "Provably Minimally-Distorted Adversarial Examples", 2018.

[Chen17] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi and Cho-Jui Hsieh, "ZOO", *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, Nov 2017.

[Chen18] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi and Cho-Jui Hsieh, "EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples", 2018.

[Cire12] Dan Cireşan, Ueli Meier and Juergen Schmidhuber, "Multi-column Deep Neural Networks for Image Classification", 2012.

[Ciss17] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin and Nicolas Usunier, "Parseval Networks: Improving Robustness to Adversarial Examples", 2017.

[Dalv04] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai and Deepak Verma, "Adversarial classification", in *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, p. 99, ACM Press, 2004.

[Deb19] Debayan Deb, Jianbang Zhang and Anil K. Jain, "AdvFaces: Adversarial Face Synthesis", 2019.

[Dhil18] Guneet S. Dhillon, Kamyar Azizzadenesheli, Zachary C. Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna and Anima Anandkumar, "Stochastic Activation Pruning for Robust Adversarial Defense", 2018.

[Dong18] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu and Jianguo Li, "Boosting Adversarial Attacks with Momentum", 2018.

[Eykh18] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno and Dawn Song, "Robust Physical-World Attacks on Deep Learning Models", 2018.

[Fein17] Reuben Feinman, Ryan R. Curtin, Saurabh Shintre and Andrew B. Gardner, "Detecting Adversarial Samples from Artifacts", 2017.

[Gong17] Zhitao Gong, Wenlu Wang and Wei-Shinn Ku, "Adversarial and Clean Data Are Not Twins", 2017.

[Good14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Y. Bengio, "Generative Adversarial Networks", *Advances in Neural Information Processing Systems*, vol. 3, 06 2014.

[Good15] Ian J. Goodfellow, Jonathon Shlens and Christian Szegedy, "Explaining and Harnessing Adversarial Examples", 2015.

[Gret12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf and Alexander Smola, "A kernel two-sample test", *The Journal of Machine Learning Research*, vol. 13, no. null, p. 723–773, Mar 2012.

[Gros17] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes and Patrick McDaniel, "On the (Statistical) Detection of Adversarial Examples", 2017.

[Gu15] Shixiang Gu and Luca Rigazio, "Towards Deep Neural Network Architectures Robust to Adversarial Examples", 2015.

[Guo18] Chuan Guo, Mayank Rana, Moustapha Cisse and Laurens van der Maaten, "Countering Adversarial Images using Input Transformations", 2018.

[He16] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, "Deep Residual Learning for Image Recognition", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[Hein17] Matthias Hein and Maksym Andriushchenko, "Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation", 2017.

[Hend17] Dan Hendrycks and Kevin Gimpel, "Early Methods for Detecting Adversarial Images", 2017.

[Hint15] Geoffrey Hinton, Oriol Vinyals and Jeff Dean, "Distilling the Knowledge in a Neural Network", 2015.

[Hoch97] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory", *Neural Computation*, vol. 9, no. 8, p. 1735–1780, Nov 1997.

[Ilya18] Andrew Ilyas, Logan Engstrom, Anish Athalye and Jessy Lin, "Black-box Adversarial Attacks with Limited Queries and Information", 2018.

[Ioff15] Sergey Ioffe and Christian Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", 2015.

[Jawa19] C. V. Jawahar, Hongdong Li, Greg Mori and Konrad Schindler, *Computer Vision – ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV*, Springer, May 2019. Google-Books-ID: NRaaDwAAQBAJ.

[Katz17] Guy Katz, Clark Barrett, David Dill, Kyle Julian and Mykel Kochenderfer, "Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks", 2017.

[Kear93] Michael Kearns and Ming Li, "Learning in the Presence of Malicious Errors", *SIAM Journal on Computing*, vol. 22, no. 4, p. 807–837, Aug 1993.

[Kipf17] Thomas N. Kipf and Max Welling, "Semi-Supervised Classification with Graph Convolutional Networks", 2017.

[Koh20] Pang Wei Koh and Percy Liang, "Understanding Black-box Predictions via Influence Functions", 2020.

[Kriz12a] Alex Krizhevsky, "Learning Multiple Layers of Features from Tiny Images", *University of Toronto*, 05 2012.

[Kriz12b] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks", in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, p. 1097–1105, Curran Associates Inc., Dec 2012.

[Kura17a] Alexey Kurakin, Ian Goodfellow and Samy Bengio, "Adversarial examples in the physical world", 2017.

[Kura17b] Alexey Kurakin, Ian Goodfellow and Samy Bengio, "Adversarial Machine Learning at Scale", 2017.

[LeCu] Yann LeCun, Corinna Cortes and Chris Burges, "MNIST handwritten digit database".

[LeCu89] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition", *Neural Comput.*, vol. 1, no. 4, p. 541–551, December 1989.

[Liu89] Dong C. Liu and Jorge Nocedal, "On the limited memory BFGS method for large scale optimization", *Mathematical Programming*, vol. 45, no. 1–3, p. 503–528, Aug 1989.

[Madr19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras and Adrian Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks", 2019.

[MCCU43] WARREN S. MCCULLOCHAND and WALTER PITTS, "A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY", *BULLETIN OF MATHEMATICAL BIOPHYSICS*, vol. 5, 1943.

[Metz17] Jan Hendrik Metzen, Tim Genewein, Volker Fischer and Bastian Bischoff, “On Detecting Adversarial Perturbations”, 2017.

[Miya16] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, Ken Nakae and Shin Ishii, “Distributional Smoothing with Virtual Adversarial Training”, 2016.

[Moos16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi and Pascal Frossard, “DeepFool: a simple and accurate method to fool deep neural networks”, 2016.

[Moos17] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi and Pascal Frossard, “Universal adversarial perturbations”, 2017.

[Moza15] Mehran Mozaffari-Kermani, Susmita Sur-Kolay, Anand Raghunathan and Niraj K. Jha, “Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare”, *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1893–1905, 2015.

[Nels08] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I. P. Rubinstein, Udam Saini, Charles Sutton, J. D. Tygar and Kai Xia, “Exploiting Machine Learning to Subvert Your Spam Filter”, in *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, LEET’08, USA, 2008, USENIX Association.

[Niel15] Michael A. Nielsen, “Neural Networks and Deep Learning”, 2015.

[Oden17] Augustus Odena, Christopher Olah and Jonathon Shlens, “Conditional Image Synthesis With Auxiliary Classifier GANs”, 2017.

[onlia] “Big Data Made Simple - One source. Many perspectives.”.

[onlib] “CIFAR-10 - Object Recognition in Images”.

[onlic] “Derivative-based Optimization”.

[onlid] “Imagenet”.

[onlie] “What are Neural Networks?”.

[onli18] “Class 1: Intro to Adversarial Machine Learning”, Jan 2018.

[Pape15] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik and Ananthram Swami, “The Limitations of Deep Learning in Adversarial Settings”, 2015.

[Pape16] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha and Ananthram Swami, “Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks”, 2016.

[Pape17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik and Ananthram Swami, “Practical Black-Box Attacks against Machine Learning”, 2017.

[Ragh18] Aditi Raghunathan, Jacob Steinhardt and Percy Liang, “Semidefinite relaxations for certifying robustness to adversarial examples”, 2018.

[Ragh20] Aditi Raghunathan, Jacob Steinhardt and Percy Liang, “Certified Defenses against Adversarial Examples”, 2020.

[Refa] David Refaeli, “Sigmoid, Softmax and their derivatives”.

[ROSE58] F. ROSENBLATT, "THE PERCEPTRON: A PROBABILISTIC MODEL FORINFORMATION STORAGE AND ORGANIZATIONIN THE BRAIN", *Psychological Review*, vol. 65, no. 6, 1958.

[Russ14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, vol. 115, 09 2014.

[Russ15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[Sama18] Pouya Samangouei, Maya Kabkab and Rama Chellappa, "Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models", 2018.

[Shaf18] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras and Tom Goldstein, "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks", 2018.

[Shaf19] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor and Tom Goldstein, "Adversarial Training for Free!", 2019.

[Shar18a] Yash Sharma and Pin-Yu Chen, "Attacking the Madry Defense Model with L_1-based Adversarial Examples", 2018.

[Shar18b] Yash Sharma and Pin-Yu Chen, "Bypassing Feature Squeezing by Increasing Adversary Strength", 2018.

[Shir16] Kuldeep Shiruru, "AN INTRODUCTION TO ARTIFICIAL NEURAL NETWORK", *International Journal of Advance Research and Innovative Ideas in Education*, vol. 1, pp. 27–30, 09 2016.

[Simo15] Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", in Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Sinh20] Aman Sinha, Hongseok Namkoong, Riccardo Volpi and John Duchi, "Certifying Some Distributional Robustness with Principled Adversarial Training", 2020.

[Soft19] C. H. I. Software, "Supervised vs. Unsupervised Machine Learning", May 2019.

[Song18a] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon and Nate Kushman, "PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples", 2018.

[Song18b] Yang Song, Rui Shu, Nate Kushman and Stefano Ermon, "Constructing Unrestricted Adversarial Examples with Generative Models", 2018.

[Sriv14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting", *The Journal of Machine Learning Research*, vol. 15, no. 1, p. 1929–1958, Jan 2014.

[Stut19] David Stutz, Matthias Hein and Bernt Schiele, "Disentangling Adversarial Robustness and Generalization", 2019.

[Su19a] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen and Yupeng Gao, "Is Robustness the Cost of Accuracy? – A Comprehensive Study on the Robustness of 18 Deep Image Classification Models", 2019.

[Su19b] Jiawei Su, Danilo Vasconcellos Vargas and Kouichi Sakurai, "One Pixel Attack for Fooling Deep Neural Networks", *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, p. 828–841, Oct 2019.

[Szeg14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow and Rob Fergus, "Intriguing properties of neural networks", 2014.

[Tjen19] Vincent Tjeng, Kai Xiao and Russ Tedrake, "Evaluating Robustness of Neural Networks with Mixed Integer Programming", 2019.

[Tram20] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh and Patrick McDaniel, "Ensemble Adversarial Training: Attacks and Defenses", 2020.

[Tsip19] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner and Aleksander Madry, "Robustness May Be at Odds with Accuracy", 2019.

[Tvei03] Amund Tveit, Magnus Lie Hetland and Håavard Engum, *Incremental and Decremental Proximal Support Vector Classification using Decay Coefficients*, vol. 2737, p. 422–429, Springer Berlin Heidelberg, 2003.

[Vand96] Lieven Vandenberghe and Stephen Boyd, "Semidefinite Programming", *SIAM Review*, vol. 38, no. 1, p. 49–95, Mar 1996.

[Werb74] Paul Werbos and Paul John, "Beyond regression : new tools for prediction and analysis in the behavioral sciences /", 01 1974.

[Wier11] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun and Jürgen Schmidhuber, "Natural Evolution Strategies", 2011.

[Wong18a] Eric Wong and J. Zico Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope", 2018.

[Wong18b] Eric Wong, Frank Schmidt, Jan Hendrik Metzen and J. Zico Kolter, "Scaling provable adversarial defenses", in S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, editors, *Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018.

[Xiao18] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu and Dawn Song, "Spatially Transformed Adversarial Examples", 2018.

[Xiao19a] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu and Dawn Song, "Generating Adversarial Examples with Adversarial Networks", 2019.

[Xiao19b] Kai Y. Xiao, Vincent Tjeng, Nur Muhammad Shafiullah and Aleksander Madry, "Training for Faster Adversarial Robustness Verification via Inducing ReLU Stability", 2019.

[Xie18] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren and Alan Yuille, "Mitigating Adversarial Effects Through Randomization", 2018.

[Xu18] Weilin Xu, David Evans and Yanjun Qi, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks", *Proceedings 2018 Network and Distributed System Security Symposium*, 2018.

[Yang17] Chaofei Yang, Qing Wu, Hai Li and Yiran Chen, “Generative Poisoning Attack Method Against Neural Networks”, 2017.

[YLeC89] J.S. Denker D. Henderson R.E. Howard W. Hubbard L.D. Jackel Y. LeCun, B. Boser, “Backpropagation Applied to Handwritten Zip Code Recognition”, *Neural Computation*, 1989.

[Zhan19a] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu and Bin Dong, “You Only Propagate Once: Accelerating Adversarial Training via Maximal Principle”, 2019.

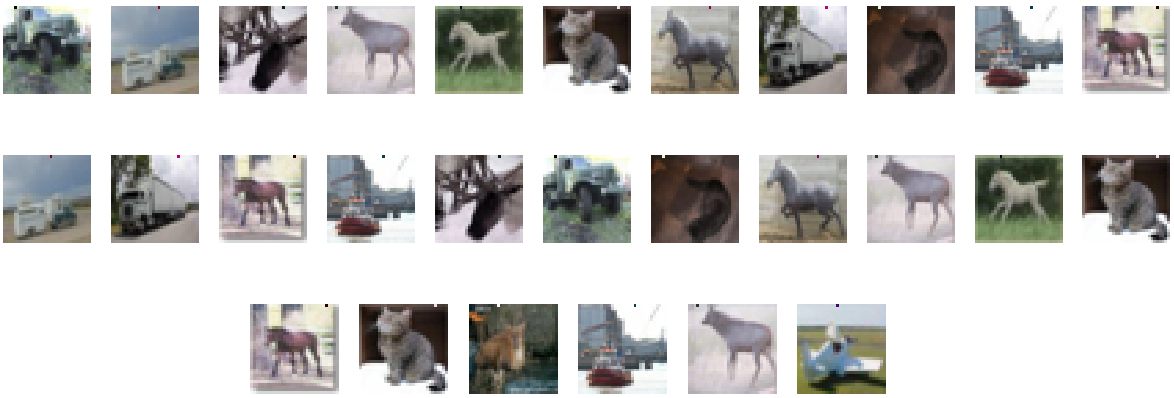
[Zhan19b] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui and Michael I. Jordan, “Theoretically Principled Trade-off between Robustness and Accuracy”, 2019.

[Züg18] Daniel Zügner, Amir Akbarnejad and Stephan Günnemann, “Adversarial Attacks on Neural Networks for Graph Data”, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul 2018.

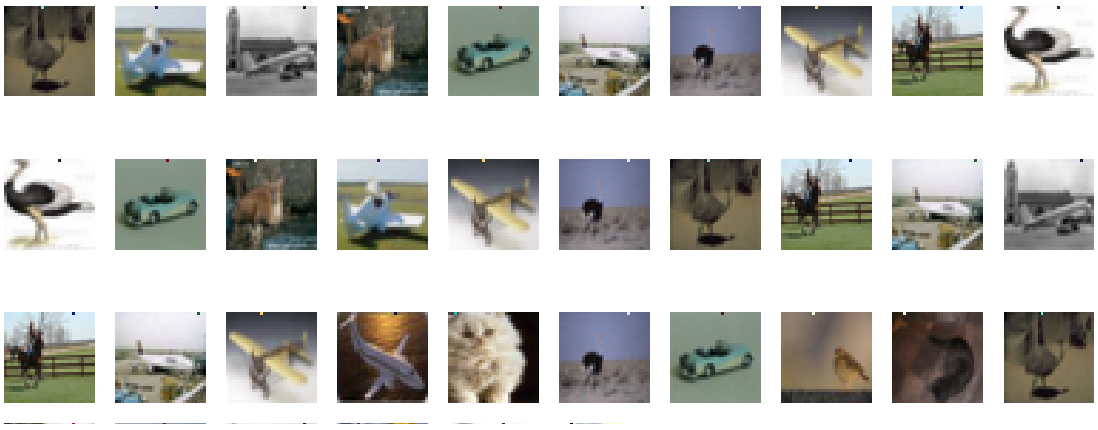
Παράρτημα Α

Επιπλέον αποτελέσματα

Α.1 Διπλή επίθεση (One pixel - FGSM)



Σχήμα Α.1: «Μη ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της διπλής επίθεσης (one-pixel-FGSM) χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων kaggle CIFAR10 για τις τιμές της παραμέτρου $\epsilon \in \{0,001, 0,5, 1,0\}$ αντίστοιχα (από πάνω προς τα κάτω).



Σχήμα Α.2: «Ισχυρά» αντιφατικά παραδείγματα που προέκυψαν μέσω της διπλής επίθεσης (one-pixel-FGSM) χρησιμοποιώντας δείγματα από τη συλλογή δεδομένων kaggle CIFAR10 για τις τιμές της παραμέτρου $\epsilon \in \{0,001, 0,5, 1,0\}$ αντίστοιχα (από πάνω προς τα κάτω).

Α.2 Squeezing color bits

			Accuracy	CL-Adv	CL-Clean
MNIST	One-pixel attack	**using 1 bit**	**74**, **26**%	**0**, **0**%	**1**, **47**%
		using 2 bits	69,12%	2,94%	5,88%
		using 3 bits	50,74%	2,94%	5,88%
		using 4 bits	49,26%	4,41%	5,88%
		using 5 bits	50%	4,41%	5,88%
		using 6 bits	50%	4,41%	5,88%
		using 7 bits	50%	4,41%	5,88%
		using 8 bits	50%	4,41%	5,88%
	FGSM attack	using 1 bit	71,70%	47,17%	11,32%
		using 2 bits	**75**, **47**%	**22**, **64**%	**3**, **77**%
		using 3 bits	74,53%	28,30%	3,77%
		using 4 bits	59,43%	22,64%	5,66%
		using 5 bits	58,49%	22,64%	5,66%
		using 6 bits	54,72%	15,09%	5,66%
		using 7 bits	54,72%	16,98%	5,66%
		using 8 bits	50%	18,87%	5,66%
	Double attack	**using 1 bit**	**85**%	**0**, **0**%	**0**, **0**%
		using 2 bits	80%	10%	10%
		using 3 bits	65%	20%	10%
		using 4 bits	65%	10%	10%
		using 5 bits	60%	20%	10%
		using 6 bits	60%	20%	10%
		using 7 bits	55%	20%	10%
		using 8 bits	50%	20%	10%
Kaggle CIFAR10	One-pixel attack	using 1 bit	58,40%	5,88%	0,0%
		using 2 bits	59,67%	9,24%	0,0%
		using 3 bits	**66**, **39**%	**7**, **98**%	**0**, **0**%
		using 4 bits	64,71%	8,82%	0,0%
		using 5 bits	55,67%	12,18%	0,0%
		using 6 bits	51,26%	12,18%	0,0%
		using 7 bits	50%	12,18%	0,0%
		using 8 bits	50%	12,18%	0,0%
	FGSM attack	using 1 bit	58,15%	5,43%	0,0%
		using 2 bits	**60**, **05**%	**4**, **89**%	**0**, **0**%
		using 3 bits	60,33%	5,43%	0,0%
		using 4 bits	54,89%	7,61%	0,0%
		using 5 bits	49,18%	7,61%	0,0%
		using 6 bits	50%	7,61%	0,0%
		using 7 bits	50,27%	7,61%	0,0%
		using 8 bits	50%	7,61%	0,0%
	Double attack	using 1 bit	63,49%	10,53%	0,0%
		using 2 bits	**59**, **87**%	**9**, **21**%	**0**, **0**%
		using 3 bits	59,87%	7,24%	0,0%
		using 4 bits	53,62%	5,92%	0,0%
		using 5 bits	50%	7,24%	0,0%
		using 6 bits	49,67%	7,89%	0,0%
		using 7 bits	50%	7,89%	0,0%
		using 8 bits	50%	7,89%	0,0%

Πίνακας Α.1: Μετρικές αξιολόγησης του ανιχνευτή squeezing color bits, όπου CL-Adv η ακρίβεια εύρεσης της σωστής κλάσης πάνω στα φιλτραρισμένα αντιφατικά παραδείγματα και CL-Clean η ακρίβεια εύρεσης της σωστής κλάσης πάνω στα φιλτραρισμένα κανονικά δείγματα.

Παράρτημα Β

Ευρετήριο ακρωνυμίων

Β.1 Ελληνικοί Όροι

ΒΝΔ: Βαθύ Νευρωνικό Δίκτυο
ΤΝΔ: Τεχνητό Νευρωνικό Δίκτυο

Β.2 Αγγλικοί Όροι

AP: Average Precision Score
AUC: Area Under the Curve
CNN: Convolutional Neural Network
DNN: Deep Neural Network
FGSM: Fast Gradient Sign Method
GAN: Generative Adversarial Network
GCN: Graph Convolutional Network
JSMA: Jacobian-based saliency map
PGD: Projected Gradient Method
ReLU: Rectified Linear Unit
ROC: Receiver Operating Characteristic
SVM: Support Vector Machines
TNR: True Negative Rate
TPR: True Positive Rate