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

Αλγόριθμοι ανταγωνιστικών επιθέσεων σε βαθιά νευρωνικά δίκτυα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΩΝΣΤΑΝΤΙΝΟΣ ΠΡΟΥΣΑΛΙΔΗΣ

Επιβλέπων : Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2022

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

Αλγόριθμοι ανταγωνιστικών επιθέσεων σε βαθιά νευρωνικά δίκτυα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΩΝΣΤΑΝΤΙΝΟΣ ΠΡΟΥΣΑΛΙΔΗΣ

Επιβλέπων : Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 19η Οκτωβρίου 2022.

Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.	Αθανάσιος Βουλόδημος Επ. Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2022

Κωνσταντίνος Προυσαλίδης

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

Περίληψη

Η βαθιά μηχανική μάθηση έχει καταφέρει εντυπωσιακή πρόοδο καταφέρνοντας καλύτερα αποτελέσματα από προηγούμενες μεθόδους μηχανικής μάθησης, σε ορισμένες περιπτώσεις μάλιστα έχει καταφέρει να ξεπεράσει ακόμα και τις ανθρώπινες επιδώσεις. Για τον λόγο αυτό πλέον χρησιμοποιείται ευρέως για μια σειρά από εφαρμογές, αρκετές από τις οποίες μπορούν να χαρακτηριστούν και ως κρίσιμες (π.χ. για λήψη ιατρικών, δικαστικών αποφάσεων κλπ). Όμως παρά τις επιτυχίες που σημειώνει σε τέτοιες εφαρμογές, έχει αποδειχθεί ότι είναι ευάλωτη σε επιθέσεις, γεγονός που την καθιστά αναξιόπιστη για πολλούς και επικίνδυνη σε κάποιες περιπτώσεις. Αυτό το αρκετά σημαντικό μειονέκτημα λοιπόν, έχει δημιουργήσει ανησυχία στην ερευνητική, και όχι μόνο, κοινότητα σχετικά με το αν μπορούμε να εμπιστευτούμε τα μοντέλα βαθιάς μάθησης για τέτοιου είδους εφαρμογές. Σε μια προσπάθεια να γίνει κατανοητό το πως επηρεάζονται τα μοντέλα από τις ανταγωνιστικές επιθέσεις, τα τελευταία χρόνια έχουν χρησιμοποιηθεί διάφορες μέθοδοι της εξηγήσιμης τεχνητής νοημοσύνης. Τέτοιες μέθοδοι μας δίνουν πληροφορίες σχετικά με το πώς ένα μοντέλο καταλήγει σε μια συγκεκριμένη πρόβλεψη, βοηθώντας μας έτσι να κατανοήσουμε καλύτερα τον τρόπο με τον οποίο επηρεάζεται από τις επιθέσεις. Η εργασία μας επικεντρώθηκε στο να χρησιμοποιήσουμε την γνώση που αποκτούμε από τέτοιες μεθόδους εξηγήσιμης τεχνητής νοημοσύνης για να τροποποιήσουμε τα ανταγωνιστικά παραδείγματα. Τα ανταγωνιστικά παραδείγματα είναι δεδομένα τα οποία έχουν τροποποιηθεί ελαφρώς ούτως ώστε να αναγκάζουν το μοντέλο να καταλήγει σε λάθος απόφαση. Στην παρούσα διπλωματική εργασία χρησιμοποιήσαμε μεθόδους επίθεσης για να δημιουργήσουμε, από τις αρχικές εικόνες του συνόλου δεδομένων μας, ανταγωνιστικά παραδείγματα καθώς επίσης και μια μέθοδο εξηγήσιμης τεχνητής νοημοσύνης για την δημιουργία saliency map για κάθε εικόνα, το οποίο μας έδειχνε σε ποια σημεία κάθε εικόνας εστιάζει το μοντέλο για να καταλήξει στην τελική απόφαση. Σκοπός μας ήτανε να εξετάσουμε αν τα saliency maps μπορούν να συνδυαστούν με τις ανταγωνιστικές επιθέσεις και αν η αλλαγή των επιθέσεων σε συγκεκριμένα σημαντικά σημεία επηρεάζουν την επιτυχία της επίθεσης. Πιο συγκεκριμένα υλοποιήσαμε δύο αρχιτεκτονικές ταξινόμησης εικόνων, έναν αλγόριθμο εξηγήσιμης τεχνητής νοημοσύνης, πειραματιστήκαμε με τρεις αλγορίθμους ανταγωνιστικών επιθέσεων και εκτελέσαμε διάφορα πειράματα προσπαθώντας να μελετήσουμε το πως αλλαγές στις υπάρχουσες επιθέσεις με βάση την γνώση που αποκτούμε από την εξηγήσιμη τεχνητής νοημοσύνη, επηρεάζουν τις μετρικές αξιολόγησης που είχαμε θέσει.

Λέξεις κλειδιά

Βαθιά νευρωνικά δίκτυα, ανταγωνιστικές επιθέσεις, ανταγωνιστικά παραδείγματα, εξηγήσιμη τεχνητή νοημοσύνη

Abstract

Deep learning has made impressive progress, outperforming previous machine learning methods, in some cases even surpassing human performance. For this reason, it is now widely used for a number of applications, several of which can also be characterized as critical (e.g. for making medical, judicial decisions, etc.). But despite its successes in such applications, it has proven to be vulnerable to attacks, which makes it unreliable and in some cases dangerous. This rather significant drawback has created concerns in the research community, among others, about whether we can trust deep learning models for these kinds of applications. In an effort to understand how models are affected by adversarial attacks, various methods of explainable artificial intelligence have been used in recent years. Such methods give us information about how a model arrives at a particular prediction, thus helping us better understand how it is affected by attacks. Our work focused on using the knowledge gained from explainable AI methods to modify adversarial examples. Adversarial examples are input data that have been slightly modified to cause the model to make the wrong decision. In this thesis we used attack methods to create adversarial examples, from our input data, as well as an explainable artificial intelligence method to create saliency maps, that highlights the focus areas in each image. Our purpose was to examine whether saliency maps can be combined with adversarial attacks and whether changing attack examples at specific important areas affects the success of the attack. More specifically, we implemented two image classification architectures, an explainable artificial intelligence algorithm, we experimented with three adversarial attack algorithms, and performed various experiments trying to study how changes in existing attacks based on the knowledge we gain from explainable artificial intelligence affect the evaluation metrics that we had set.

Key words

Deep neural networks, adversarial attacks, adversarial examples, explainable artificial intelligence

Ευχαριστίες

Κωνσταντίνος Προυσαλίδης,

Αθήνα, 19η Οκτωβρίου 2022

Περιεχόμενα

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

Τα τελευταία χρόνια η βαθιά μάθηση έχει εξελιχθεί πάρα πολύ κάνοντας τεράστια βήματα προόδου πετυχαίνοντας κάποια εντυπωσιακά αποτελέσματα σε μια σειρά από εφαρμογές. Η εξέλιξη της είναι τόσο σημαντική που έχει καταφέρει σε πολλές εφαρμογές να ξεπεράσει προηγούμενες μεθόδους μηχανικής μάθησης, φτάνοντας στο σημείο να σε ορισμένες περιπτώσεις να σημειώνει καλύτερες επιδόσεις ακόμα και από τον άνθρωπο. Αυτή η πρόοδος έχει οδηγήσει σε ολοένα και περισσότερες εφαρμογές να χρησιμοποιούν μοντέλα βαθιάς μάθησης, αρκετές από τις οποίες μπορούν να χαρακτηριστούν ως κρίσιμες (πχ για λήψη ιατρικών ή δικαστικών αποφάσεων). Παρόλα αυτά όμως, έχει βρεθεί ότι τα μοντέλα βαθιάς μάθησης είναι ιδιαίτερα ευάλωτα σε επιθέσεις, γεγονός που σύμφωνα με αρκετούς τα καθιστά αναξιόπιστα ή ακόμα και επικίνδυνα σε ορισμένες περιπτώσεις πιο κρίσιμων εφαρμογών. Σε μια προσπάθεια να καταλάβουμε καλύτερα το γιατί να μοντέλα βαθιάς μάθησης είναι τόσο ευάλωτα σε επιθέσεις, στην ερευνητική κοινότητα έχουν χρησιμοποιηθεί μέθοδοι εξηγήσιμης τεχνητής νοημοσύνης για να μας δώσουν περισσότερες πληροφορίες για το πως καταλήγει ένα μοντέλο σε μια απόφαση.

1.1 Κίνητρο

Δύο από τους πιο ανερχόμενους κλάδους της μηχανικής μάθησης είναι οι ανταγωνιστικές επιθέσεις και η εξηγήσιμη τεχνητή νοημοσύνη. Όσο βελτιώνονται οι επιδόσεις των μοντέλων βαθιάς μάθησης τόσο τέτοια μοντέλα ενσωματώνονται σε όλο και περισσότερες εφαρμογές και συστήματα. Από την μια οι ανταγωνιστικές επιθέσεις έχουν απασχολήσει την επιστημονική κοινότητα, ειδικά όσον αφορά κρίσιμες εφαρμογές καθώς επηρεάζουν άμεσα την επίδοση και την αξιοπιστία των μοντέλων. Από την άλλη η εξηγήσιμη τεχνητή νοημοσύνη προσπαθεί να μας δώσει κάποιες παραπάνω πληροφορίες σχετικά με το πως καταλήγει σε μια απόφαση ένα μοντέλο. Πρόκειται για δύο κλάδους λοιπόν που έχουν απασχολήσει πολύ έντονα την ερευνητική και όχι μόνο κοινότητα τα τελευταία χρόνια έχοντας ακόμα πολύ μεγάλο ερευνητικό ενδιαφέρον, ειδικά όσον αφορά τον συνδυασμό τους, για τους παραπάνω λόγους επιλέξαμε να ασχοληθούμε και με τους δύο αυτούς τομείς. Στα πλαίσια αυτής της διπλωματικής προσπαθήσαμε να συνδυάσουμε αυτές τις δύο ερευνητικές περιοχές, μελετώντας το πως κάποιες μέθοδοι εξηγήσιμης τεχνητής νοημοσύνης επηρεάζουν τις επιδόσεις των ανταγωνιστικών επιθέσεων.

1.2 Συνεισφορά Διπλωματικής Εργασίας

Στην παρούσα διπλωματική εργασία πραγματευόμαστε τις ανταγωνιστικές επιθέσεις σε Βαθιά Νευρωνικά Δίκτυα και μελετήσαμε την σύνδεση τους με την εξηγήσιμη τεχνητή νοημοσύνη. Αρχικά χρησιμοποιήσαμε έτοιμες μεθόδους επίθεσης για να δημιουργήσουμε ανταγωνιστικά παραδείγματα με βάση τα αρχικά δεδομένα εισόδου μας. Ακόμα για τα αρχικά δεδομένα εισόδου χρησιμοποιήσαμε έναν αλγόριθμο εξηγήσιμης τεχνητής νοημοσύνης για τη δημιουργία ενός saliency map για κάθε εικόνα, το οποίο μας έδειχνε σε ποια σημεία κάθε εικόνας εστιάζει το μοντέλο για να καταλήξει στην τελική απόφαση. Στην συνέχεια συνδυάσαμε τα παραπάνω δημιουργώντας δικά μας ανταγωνιστικά παραδείγματα, χρησιμοποιώντας την γνώση από τα saliency map και κρατώντας συγκεκριμένο ποσοστό pixel από τα αρχικά ανταγωνιστικά παραδείγματα. Συγκεκριμένα για κάθε εικόνα πειραματιστήκαμε με εφτά διαφορετικές τιμές σχετικά με το ποσοστό των pixel (20,40,60,80,90,95,100) κάθε μια από τις οποίες υποδεικνύει το ποσοστό pixel της αρχικής εικόνας που θα αντικατασταθεί από τα αντίστοιχα pixel του αρχικού ανταγωνιστικού παραδείγματος. Η διαδικασία αυτή για κάθε εικόνα έγινε δυο φορές, μια φορά αντικαθιστώντας τα pixel με τυχαίο τρόπο και μια φορά αντικαθιστώντας τα pixel που έχουν τις μεγαλύτερες αντίστοιχες τιμές στο saliency map. Την παραπάνω διαδικασία την υλοποιήσαμε για τρεις διαφορετικές επιθέσεις και πραγματοποιήσαμε τα πειράματα σε δυο διαφορετικές αρχιτεκτονικές κατηγοριοποίησης εικόνων. Παράλληλα πραγματοποιήσαμε κάποια πειράματα για το πως επηρεάζεται το ποσοστό πρόβλεψης της πραγματικής και της κυρίαρχης κλάσης σε κάποιες περιπτώσεις καθώς επίσης και το πως μεταβάλλονται τα ποσοστά όλων των κλάσεων για κάποια συγκεκριμένα παραδείγματα.

1.3 Δομή Διπλωματικής Εργασίας

Στο Κεφάλαιο 2 καλύπτεται το θεωρητικό υπόβαθρο όσων κρίθηκαν απαραίτητα για την κατανόηση αυτής της διπλωματικής εργασίας. Αρχικά γίνεται μια γενική αναφορά στο τομέα της βαθιάς μάθησης και σε κάποιες βασικές ορολογίες. Στη συνέχεια γίνεται μια παρουσίαση των ανταγωνιστικών επιθέσεων σε βαθιά νευρωνικά δίκτυα και πιο συγκεκριμένα παρουσιάζουμε έναν ορισμό των ανταγωνιστικών επιθέσεων, αναλύουμε τις κατηγορίες και τα είδη τους και παρουσιάζουμε την απαραίτητη θεωρία για τις συγκεκριμένες μεθόδους που χρησιμοποιήσαμε. Η τρίτη υποενότητα του κεφαλαίου αφορά την απαραίτητη θεωρία σχετικά με τον τομέα της εξηγήσιμης τεχνητής νοημοσύνης. Από τον ορισμό και μια ιστορική αναδρομή, μέχρι τις βασικές μεθόδους και την σύνδεση εξηγήσιμης τεχνητής νοημοσύνης με τις ανταγωνιστικές επιθέσεις, όλα όσα θα χρειαστούν για να μπορεί ο αναγνώστης να ακολουθήσει τη ροή της εργασίας βρίσκονται εκεί.

Στο Κεφάλαιο 3 αναλύονται οι σχεδιαστικές επιλογές που κάναμε κατά την ανάπτυξη των μοντέλων μας και τα βήματα υλοποίησης που ακολουθήσαμε. Παρουσιάζονται τα δεδομένα με τα οποία εργαστήκαμε και τα χαρακτηριστικά τους. Αναπτύσσονται διλήμματα που προέκυψαν στην πορεία και η λογική επίλυσής τους, και τέλος γίνεται μια αναφορά στο περιβάλλον εκτέλεσης των πειραμάτων.

Στο Κεφάλαιο 4 παρουσιάζονται τα αποτελέσματα των πειραμάτων κι εξάγονται τα συμπεράσματα. Ξεκινάμε με μια συγκριτική μελέτη των διαφορετικών μοντέλων που δημιουργήσαμε, αναλύοντας ξεχωριστά τα διάφορα χαρακτηριστικά και τον τρόπο με τον οποίο αυτά επηρεάζουν την εκπαίδευση και τα αποτελέσματα του μοντέλου. Στη συνέχεια αναλύουμε προσωπικές πα-

ρατηρήσεις του συγγραφέα και κάνουμε μια ποιοτική αξιολόγηση των αποτελεσμάτων έχοντας απομακρυνθεί από τα αριθμητικά αποτελέσματα και κοιτώντας τη γενικότερη εικόνα. Τέλος συγκρίνουμε τα αποτελέσματά μας με όσες εργασίες βρέθηκαν να έχουν δημοσιευθεί μέχρι τη στιγμή συγγραφής της διπλωματικής και είναι σχετικές και συγκρίσιμες με το πρόβλημα που καλούμαστε να λύσουμε.

Στο Κεφάλαιο 5 γίνεται μια σύνοψη της διπλωματικής και στη συνέχεια εξάγονται και συγκεντρώνονται τα τελικά συμπεράσματα. Τέλος αναφέρονται πιθανές μελλοντικές κατευθύνσεις της επιστημονικής μελέτης, περισσότερες από τις οποίες προκύπτουν ως λογική συνέχεια της δουλειάς που κάναμε για αυτή την εργασία.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Στο παρόν κεφάλαιο θα καλυφθεί το θεωρητικό υπόβαθρο που κρίνεται απαραίτητο για την κατανόηση της διπλωματικής εργασίας. Χωρίζεται σε τρία βασικά μέρη, το πρώτο που αφορά σε κάποια γενικά στοιχεία για την βαθιά μάθηση, το δεύτερο για τις ανταγωνιστικές επιθέσεις σε βαθιά νευρωνικά δίκτυα, τους τρόπους λειτουργίας και τα βασικά είδη τους. Τέλος, το τρίτο κομμάτι ασχολείται με το explainable AI, τις βασικές μεθόδους του και τις συσχετίσεις που έχουν γίνει μέχρι τώρα από την ερευνητική κοινότητα, με τις ανταγωνιστικές επιθέσεις.

2.1 Βαθιά Μάθηση

Η βαθιά μάθηση είναι ένα υποσύνολο της μηχανικής μάθησης, στην ουσία αφορά σε νευρωνικά δίκτυα με τρία ή περισσότερα επίπεδα. Αυτά τα νευρωνικά δίκτυα προσπαθούν να προσομοιώσουν τη συμπεριφορά του ανθρώπινου εγκεφάλου, επιτρέποντας του να «μάθει» από μεγάλες ποσότητες δεδομένων. Ενώ ένα νευρωνικό δίκτυο με ένα μόνο στρώμα μπορεί να κάνει κατά προσέγγιση προβλέψεις, πρόσθετα κρυφά στρώματα μπορούν να βοηθήσουν στη βελτιστοποίηση και τη τελειοποίηση της ακρίβειας της πρόβλεψης.

Τα νευρωνικά δίκτυα βαθιάς μάθησης ή τεχνητά νευρωνικά δίκτυα επιχειρούν να μιμηθούν τον ανθρώπινο εγκέφαλο μέσω ενός συνδυασμού εισόδου δεδομένων, βαρών και bias. Αυτά τα στοιχεία συνεργάζονται για την ακριβή αναγνώριση, ταξινόμηση και περιγραφή αντικειμένων μέσα στα δεδομένα.

Τα βαθιά νευρωνικά δίκτυα αποτελούνται από πολλαπλά στρώματα διασυνδεδεμένων κόμβων, το καθένα βασίζεται στο προηγούμενο επίπεδο για να βελτιώσει και να βελτιστοποιήσει την πρόβλεψη ή την κατηγοριοποίηση. Αυτή η εξέλιξη των υπολογισμών μέσω του δικτύου ονομάζεται διάδοση προς τα εμπρός (forward propagation). Τα στρώματα εισόδου και εξόδου ενός βαθιού νευρωνικού δικτύου ονομάζονται ορατά στρώματα. Το επίπεδο εισόδου είναι όπου το μοντέλο βαθιάς μάθησης απορροφά τα δεδομένα για επεξεργασία και το επίπεδο εξόδου είναι όπου γίνεται η τελική πρόβλεψη ή ταξινόμηση.

Μια άλλη διαδικασία που ονομάζεται αντίστροφη διάδοση (back propagation) χρησιμοποιεί αλγόριθμους, όπως η gradient descent, για τον υπολογισμό των σφαλμάτων στις προβλέψεις και στη συνέχεια προσαρμόζει τα βάρη (weights) και τις προκαταλήψεις (biases) της συνάρτησης μετακινώντας προς τα πίσω μέσα από τα επίπεδα σε μια προσπάθεια να εκπαιδεύσει το μοντέλο. Μαζί, η προς τα εμπρός διάδοση (forward propagation) και η αντίστροφη διάδοση (back propagation) επιτρέπουν σε ένα νευρωνικό δίκτυο να κάνει προβλέψεις και να διορθώνει ανάλογα τυχόν σφάλματα. Μέσω επαναλήψεων ο αλγόριθμος γίνεται σταδιακά πιο ακριβής.

Τα παραπάνω περιγράφουν τον απλούστερο τύπο βαθιού νευρωνικού δικτύου με τους απλούστερους όρους. Ωστόσο, οι αλγόριθμοι βαθιάς μάθησης είναι αρκετά περίπλοκοι και υπάρχουν διαφορετικοί τύποι νευρωνικών δικτύων για την αντιμετώπιση συγκεκριμένων προβλημάτων ή συνόλων δεδομένων. Για παράδειγμα, τα συνελικτικά νευρωνικά δίκτυα (CNN), που χρησιμοποιούνται κυρίως σε εφαρμογές όρασης υπολογιστών και ταξινόμησης εικόνων, μπορούν να ανιχνεύσουν χαρακτηριστικά και μοτίβα μέσα σε μια εικόνα, επιτρέποντας εργασίες, όπως η ανίχνευση ή η αναγνώριση αντικειμένων. Ενώ τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) χρησιμοποιούνται συνήθως σε εφαρμογές φυσικής γλώσσας και αναγνώρισης ομιλίας, καθώς αξιοποιεί δεδομένα διαδοχικών ή χρονικών σειρών.

2.2 Ανταγωνιστικές Επιθέσεις σε ΒΝΔ

2.2.1 Ορισμός Ανταγωνιστικών Επιθέσεων σε ΒΝΔ

Τα τελευταία χρόνια έχει σημειωθεί μια τεράστια βελτίωση στη ακρίβεια που μπορούν να πετυχαίνουν τα βαθιά νευρωνικά δίκτυα και για αυτό τον λόγο χρησιμοποιούνται σε όλο και περισσότερες κρίσιμες εφαρμογές, όπως μεταξύ άλλων είναι συστήματα κυβερνοασφάλειας, ιατρικές διαγνώσεις και αυτοοδηγούμενα οχήματα. Όπως είναι λογικό σε τέτοιου είδους εφαρμογές τα περιθώρια λάθους που μπορούν να γίνουν ανεκτά είναι ελάχιστα. Όμως εδώ και κάποια χρόνια έχει αποδειχθεί ότι παρά τις κορυφαίες επιδόσεις που πετυχαίνουν τα βαθιά νευρωνικά δίκτυα, είναι ευάλωτα σε ανταγωνιστικά παραδείγματα (adversarial examples) Goodfellow (2015) δηλαδή σε ελαφρώς τροποποιημένα δεδομένα εισόδου τα οποία κατηγοριοποιούνται λάθος από το δίκτυο, παρότι σε πολύ μεγάλο βαθμό είναι παρόμοια με τα αρχικά σωστά κατηγοριοποιημένα δεδομένα. Αυτό το γεγονός σε συνδυασμό με την χρήση μοντέλων βαθιάς μάθησης σε όλο και αυξανόμενο εύρος εφαρμογών έχει κεντρίσει το ενδιαφέρον της ερευνητικής κοινότητας γύρω από τις ανταγωνιστικές επιθέσεις. Σε μια προσπάθεια τόσο για δημιουργία πιο ανθεκτικών μοντέλων που θα παρέχουν μεγαλύτερη ασφάλεια απέναντι σε ανταγωνιστικές επιθέσεις, όσο και σε δημιουργία πιο σύνθετων και πολύπλοκων επιθέσεων.

Μερικά παραδείγματα περιλαμβάνουν επιθέσεις στο φιλτράρισμα ανεπιθύμητων μηνυμάτων, όπου τα ανεπιθύμητα μηνύματα συγκαλύπτονται μέσω της ορθογραφίας των "κακών" λέξεων ή της εισαγωγής "καλών" λέξεων B.Biggio (2010) Bruckner (2012) επιθέσεις στην ασφάλεια του υπολογιστή, όπως η συσκότιση κώδικα κακόβουλου λογισμικού σε πακέτα δικτύου ή η τροποποίηση των χαρακτηριστικών μιας ροής δικτύου για την παραπλανητική ανίχνευση εισβολής Apruzzese (2021) Vitorino (2022) επιθέσεις στη βιομετρική αναγνώριση όπου μπορούν να εκμεταλλευτούν πλαστά βιομετρικά χαρακτηριστικά για να πλαστογραφήσουν έναν νόμιμο χρήστη Rodrigues (2009) ή για να θέσουν σε κίνδυνο τις γκαλερί προτύπων των χρηστών που προσαρμόζονται σε ενημερωμένα χαρακτηριστικά με την πάροδο του χρόνου.

Οι ερευνητές έδειξαν ότι αλλάζοντας μόνο ένα pixel ήταν δυνατό να εξαπατηθούν οι αλγόριθμοι βαθιάς μάθησης Su (2009). Άλλοι εκτύπωσαν μια τρισδιάστατη χελώνα-παιχνίδι με υφή σχεδιασμένη για να κάνει την τεχνητή νοημοσύνη ανίχνευσης αντικειμένων της Google να την ταξινομήσει ως τουφέκι ανεξάρτητα από τη γωνία από την οποία παρατηρήθηκε η χελώνα Athalye. Η δημιουργία της χελώνας απαιτούσε μόνο εμπορικά διαθέσιμη τεχνολογία τρισδιάστατης εκτύπωσης χαμηλού κόστους.

2.2.2 Κατηγορίες Ανταγωνιστικών Επιθέσεων

Οι επιθέσεις εναντίον (επιβλεπόμενων) αλγορίθμων μηχανικής μάθησης έχουν κατηγοριοποιηθεί σε τρεις κύριους άξονες Barreno (2010) οι οποίοι είναι η επιρροή στον ταξινομητή, η παραβίαση ασφαλείας και η ειδικότητά τους.

- Επιρροή ταξινομητή:
  - Οι αιτιολογικές επιθέσεις επηρεάζουν τη μάθηση με έλεγχο των δεδομένων εκπαίδευσης. Ο εισβολέας έχει την ικανότητα να επηρεάζει τα δεδομένα εκπαίδευσης που χρησιμοποιούνται για την κατασκευή του ταξινομητή
  - Οι διερευνητικές επιθέσεις εκμεταλλεύονται εσφαλμένες ταξινομήσεις αλλά δεν επηρεάζουν την εκπαίδευση. Ο εισβολέας δεν επηρεάζει τον εκπαιδευμένο ταξινομητή, αλλά μπορεί να στείλει νέα στιγμιότυπα στον ταξινομητή και πιθανώς να παρατηρήσει τις αποφάσεις του για κάποιες περιπτώσεις
- Παραβίαση ασφαλείας:
  - Οι επιθέσεις ακεραιότητας προκαλούν κίνδυνο μέσω ψευδών αρνητικών στοιχείων. Επιτρέπουν σε επιβλαβή στοιχεία να περάσουν μέσα από το φίλτρο ως ψευδώς αρνητικά στοιχεία.
  - Οι επιθέσεις διαθεσιμότητας προκαλούν άρνηση υπηρεσίας, συνήθως μέσω ψευδών θετικών στοιχείων. Δημιουργούν μια άρνηση υπηρεσίας στην οποία καλοήθεις στοιχεία φιλτράρονται εσφαλμένα ως ψευδείς.
- Ειδικότητα:
  - Οι στοχευμένες επιθέσεις επικεντρώνονται σε μια συγκεκριμένη περίπτωση. Η επίθεση είναι εξαιρετικά στοχευμένη στην υποβάθμιση της απόδοσης του ταξινομητή σε μια συγκεκριμένη περίπτωση
  - Οι αδιάκριτες επιθέσεις περιλαμβάνουν μια ευρεία κατηγορία περιπτώσεων. Η επίθεση στοχεύει να προκαλέσει την αποτυχία του ταξινομητή με τρόπο αδιάκριτο σε μια ευρεία κατηγορία περιπτώσεων.

Αυτή η ταξινόμηση έχει επεκταθεί σε ένα πιο ολοκληρωμένο μοντέλο απειλής που επιτρέπει σαφείς υποθέσεις σχετικά με τον στόχο του αντιπάλου, τη γνώση του συστήματος που δέχεται επίθεση, την ικανότητα χειρισμού των δεδομένων εισόδου/συστήματος και τη στρατηγική επίθεσης (Biggio et al., 2017b) Biggio et al. (2014). Αυτή η ταξινόμηση έχει επεκταθεί περαιτέρω για να συμπεριλάβει διαστάσεις για στρατηγικές άμυνας ενάντια σε αντίπαλες επιθέσεις Heinrich (2020).

2.2.2.1 Δηλητηρίαση δεδομένων

Η δηλητηρίαση συνίσταται στη μόλυνση του συνόλου δεδομένων εκπαίδευσης. Δεδομένου ότι οι αλγόριθμοι μάθησης διαμορφώνονται από τα σύνολα δεδομένων εκπαίδευσης, η δηλητηρίαση μπορεί να προγραμματίσει από την αρχή αποτελεσματικά τους αλγόριθμους. Έχουν δημιουργηθεί σοβαρές ανησυχίες ειδικά για τα δεδομένα εκπαίδευσης που δημιουργούνται από τους χρήστες,

π.χ. για σύσταση περιεχομένου ή μοντέλα φυσικής γλώσσας, ειδικά δεδομένης της πανταχού παρουσίας ψεύτικων λογαριασμών. Για να μετρηθεί η κλίμακα του κινδύνου, αρκεί να σημειωθεί ότι το Facebook φέρεται να αφαιρεί περίπου 7 δισεκατομμύρια ψεύτικους λογαριασμούς ετησίως Price nyp. Στην πραγματικότητα, η δηλητηρίαση δεδομένων έχει αναφερθεί ως η κύρια ανησυχία για βιομηχανικές εφαρμογές Siva Kumar et al. (2020).

Στα μέσα κοινωνικής δικτύωσης, οι καμπάνιες παραπληροφόρησης είναι γνωστό ότι παράγουν τεράστιες ποσότητες κατασκευασμένων δραστηριοτήτων για την προκατάληψη (bias) αλγορίθμων συστάσεων και εποπτείας, για να προωθήσουν συγκεκριμένο περιεχόμενο έναντι άλλων.

Μια συγκεκριμένη περίπτωση δηλητηρίασης δεδομένων ονομάζεται επίθεση με κερκόπορτα (backdoor attack), Schwarzschild et al. (2021), η οποία στοχεύει να διδάξει μια συγκεκριμένη συμπεριφορά για εισόδους με δεδομένο έναυσμα, π.χ. ένα μικρό ελάττωμα σε εικόνες, ήχους, βίντεο ή κείμενα. Για παράδειγμα, τα συστήματα ανίχνευσης εισβολής (IDS) συχνά επανεκπαιδεύονται χρησιμοποιώντας δεδομένα που συλλέγονται. Ένας εισβολέας μπορεί να δηλητηριάσει αυτά τα δεδομένα εισάγοντας κακόβουλα δείγματα κατά τη λειτουργία που στη συνέχεια διακόπτουν την επανεκπαίδευση.

2.2.2.2 Βυζαντινές επιθέσεις

Καθώς η μηχανική μάθηση είναι κλιμακούμενη, συχνά βασίζεται σε πολλαπλές υπολογιστικές μηχανές. Στην ομοσπονδιακή μάθηση, για παράδειγμα, οι συσκευές ακμών συνεργάζονται με έναν κεντρικό διακομιστή, συνήθως στέλνοντας διαβαθμίσεις ή παραμέτρους μοντέλου. Ωστόσο, ορισμένες από αυτές τις συσκευές ενδέχεται να αποκλίνουν από την αναμενόμενη συμπεριφορά τους, π.χ. να βλάπτουν το μοντέλο του κεντρικού διακομιστή Baruch et al. (2019) ή να προκαταλάβουν αλγόριθμους προς ορισμένες συμπεριφορές (π.χ. ενίσχυση της σύστασης περιεχομένου παραπληροφόρησης). Από την άλλη πλευρά, εάν η εκπαίδευση εκτελείται σε ένα μόνο μηχάνημα, τότε το μοντέλο είναι πολύ ευάλωτο σε αστοχία του μηχανήματος ή επίθεση στο μηχάνημα, το μηχάνημα είναι ένα μόνο σημείο αστοχίας (single point of failure) Mhamdi (2022). Στην πραγματικότητα, ο ιδιοκτήτης του μηχανήματος μπορεί ο ίδιος να εισάγει αποδεδειγμένα μη ανιχνεύσιμες κερκόπορτες Goldwasser et al. (2022).

Οι τρέχουσες κορυφαίες λύσεις για να γίνουν οι (κατανεμημένοι) αλγόριθμοι μάθησης αποδεδειγμένα ανθεκτικοί σε μια μειοψηφία κακόβουλων (γνωστών και ως βυζαντινών) επιθέσεων, βασίζονται σε ισχυρούς κανόνες συγκέντρωσης διαβάθμισης Goldwasser et al. (2022) Blanchard et al. (2017) Chen et al. (2018) El Mhamdi et al. (2018) Mhamdi et al. (2021) Data and Diggavi (2021). Ωστόσο, στο πλαίσιο ετερογενών ειλικρινών συμμετεχόντων, όπως χρήστες με διαφορετικές καταναλωτικές συνήθειες για αλγόριθμους συστάσεων ή στιλ γραφής για μοντέλα γλώσσας, υπάρχουν αποδεδειγμένα θεωρήματα αδυναμίας για το τι μπορεί να εγγυηθεί οποιοσδήποτε ισχυρός αλγόριθμος μάθησης Karimireddy et al. (2022) Karimireddy et al. (2022).

2.2.2.3 Υπεκφυγή

Οι επιθέσεις υπεκφυγής Biggio et al. (2017a) Biggio et al. (2017b) Biggio et al. (2014) Nelson et al. (2010) συνίστανται στην εκμετάλλευση της ατέλειας ενός εκπαιδευμένου μοντέλου. Για παράδειγμα, οι αποστολείς ανεπιθύμητης αλληλογραφίας (spam) και οι χάκερ συχνά προσπαθούν να αποφύγουν τον εντοπισμό θολώνοντας το περιεχόμενο ανεπιθύμητων μηνυμάτων ηλεκτρονικού

ταχυδρομείου και κακόβουλου λογισμικού. Τα δείγματα τροποποιούνται για να αποφεύγουν την ανίχνευση, δηλαδή να χαρακτηριστούν ως νόμιμα. Αυτό δεν συνεπάγεται επιρροή στα δεδομένα εκπαίδευσης. Ένα σαφές παράδειγμα υπεκφυφυγής είναι τα ανεπιθύμητα μηνύματα που βασίζονται σε εικόνες, στα οποία το περιεχόμενο ανεπιθύμητης αλληλογραφίας ενσωματώνεται σε μια συνημμένη εικόνα για να αποφευχθεί η ανάλυση κειμένου από φίλτρα κατά της ανεπιθύμητης αλληλογραφίας. Ένα άλλο παράδειγμα υπεκφυγής δίνεται από επιθέσεις πλαστογράφησης κατά βιομετρικών συστημάτων επαλήθευσης Rodrigues (2009).

Οι επιθέσεις υπεκφυγής μπορούν γενικά να χωριστούν σε δύο διαφορετικές με βάση την γνώση του επιτιθέμενου για το μοντέλο: επιθέσεις μαύρου κουτιού (Black-Box) [1] και επιθέσεις λευκού κουτιού (White-Box). Θα αναφερθούμε αναλυτικότερα στην υποενότητα 2.2.3.

2.2.2.4 Εξαγωγή μοντέλου

Η εξαγωγή μοντέλου περιλαμβάνει έναν αντίπαλο που διερευνά ένα σύστημα μηχανικής εκμάθησης μαύρου κουτιού προκειμένου να εξαγάγει τα δεδομένα στα οποία εκπαιδεύτηκε Kalpesh Krishna. Αυτό μπορεί να προκαλέσει προβλήματα όταν είτε τα δεδομένα εκπαίδευσης είτε το ίδιο το μοντέλο είναι ευαίσθητα και εμπιστευτικά. Για παράδειγμα, η εξαγωγή μοντέλου θα μπορούσε να χρησιμοποιηθεί για την εξαγωγή ενός ιδιόκτητου μοντέλου διαπραγμάτευσης μετοχών το οποίο ο αντίπαλος θα μπορούσε στη συνέχεια να χρησιμοποιήσει για δικό του οικονομικό όφελος.

Στην ακραία περίπτωση, η εξαγωγή μοντέλου μπορεί να οδηγήσει σε κλοπή μοντέλου, η οποία αντιστοιχεί στην εξαγωγή επαρκούς όγκου δεδομένων από το μοντέλο για να καταστεί δυνατή η πλήρης ανακατασκευή του μοντέλου.

Από την άλλη πλευρά, το συμπέρασμα μέλους είναι μια στοχευμένη επίθεση εξαγωγής μοντέλου, η οποία συνάγει τον κάτοχο ενός σημείου δεδομένων, συχνά αξιοποιώντας την υπερπροσαρμογή (overfitting) που προκύπτει από κακές πρακτικές μηχανικής εκμάθησης Dickson. Προκαλεί ανησυχία ότι, μερικές φορές αυτό μπορεί να επιτευχθεί ακόμη και χωρίς γνώση ή πρόσβαση στις παραμέτρους ενός μοντέλου στόχου, εγείροντας ανησυχίες για την ασφάλεια για μοντέλα που έχουν εκπαιδευτεί σε ευαίσθητα δεδομένα, συμπεριλαμβανομένων, ενδεικτικά, των ιατρικών αρχείων ή/και των προσωπικών στοιχείων ταυτοποίησης. Με την εμφάνιση της μεταφοράς μάθησης (transfer learning) και της δημόσιας προσβασιμότητας πολλών μοντέλων μηχανικής μάθησης τελευταίας τεχνολογίας, οι εταιρείες τεχνολογίας στρέφονται όλο και περισσότερο στη δημιουργία μοντέλων που βασίζονται σε άλλα δημόσια μοντέλα, δίνοντας στους επιτιθέμενους ελεύθερα προσβάσιμες πληροφορίες για τη δομή και τον τύπο του μοντέλου που χρησιμοποιείται Dickson.

2.2.3 Συγκεκριμένα είδη Ανταγωνιστικών Επιθέσεων σε ΒΝΔ

Υπάρχει μια μεγάλη ποικιλία διαφορετικών ανταγωνιστικών επιθέσεων που μπορούν να χρησιμοποιηθούν εναντίον συστημάτων μηχανικής μάθησης. Πολλά από αυτά λειτουργούν τόσο σε συστήματα βαθιάς μάθησης όσο και σε παραδοσιακά μοντέλα μηχανικής μάθησης όπως τα SVM Biggio et al. (2012) και η γραμμική παλινδρόμηση Jagielski et al. (2018). Ένα γενικό δείγμα αυτών των τύπων επιθέσεων περιλαμβάνει:

- Ανταγωνιστικά παραδείγματα (Adversarial Examples)

[1] https://github.com/Trusted-AI/adversarial-robustness-toolbox

- Επιθέσεις Δούρειου ίππου/κερκόπορτας (Trojan Attacks / Backdoor Attacks)
- Αναστροφή μοντέλου (Model Inversion)
- Συμπέρασμα μέλους (Membership Inference)

Από τους παραπάνω τύπους επιθέσεων εμείς επικεντρωθήκαμε στα ανταγωνιστικά παραδείγματα. Ένα ανταγωνιστικό παράδειγμα αναφέρεται σε ειδικά κατασκευασμένα στοιχεία που έχουν σχεδιαστεί για να φαίνονται «φυσιολογικά» στους ανθρώπους, αλλά προκαλούν εσφαλμένη ταξινόμηση σε ένα μοντέλο μηχανικής μάθησης. Συχνά, μια μορφή ειδικά σχεδιασμένου "θορύβου" χρησιμοποιείται για να προκαλέσει τις εσφαλμένες ταξινομήσεις.

Επίσης όπως αναφέρθηκε και στην υποενότητα 2.2.2.3. οι επιθέσεις μπορούν να χωριστούν σε δύο κατηγορίες με βάση την γνώση που έχει ο επιτιθέμενος για το μοντέλο.

- Επιθέσεις μαύρου κουτιού (Black box): Από την μία οι επιθέσεις μαύρου κουτιού στην ανταγωνιστική μηχανική μάθηση προϋποθέτουν ότι ο αντίπαλος μπορεί να λάβει εξόδους μόνο για τις παρεχόμενες εισόδους και δεν γνωρίζει τη δομή ή τις παραμέτρους του μοντέλου Sensen Guo (2021). Σε αυτήν την περίπτωση, το ανταγωνιστικό παράδειγμα δημιουργείται είτε χρησιμοποιώντας ένα μοντέλο που δημιουργήθηκε από την αρχή, είτε χωρίς κανένα μοντέλο (εξαιρουμένης της δυνατότητας αναζήτησης στο αρχικό μοντέλο). Σε κάθε περίπτωση, ο στόχος αυτών των επιθέσεων είναι να δημιουργήσουν ανταγωνιστικά παραδείγματα που μπορούν να μεταφερθούν στο εν λόγω μοντέλο μαύρου κουτιού Gomes.
- Επιθέσεις λευκού κουτιού (White box): Από την άλλη οι επιθέσεις λευκού κουτιού προϋποθέτουν ότι ο αντίπαλος έχει πρόσβαση στις παραμέτρους του μοντέλου εκτός από τη δυνατότητα λήψης ετικετών για τις παρεχόμενες εισόδους Gomes.

Μια ακόμα κατηγοριοποίηση που μπορεί να γίνει στις ανταγωνιστικές επιθέσεις είναι με βάση το αν στοχεύουν σε κάποιο συγκεκριμένο αποτέλεσμα ή απλώς να μπερδέψουν το μοντέλο. Έτσι χωρίζονται σε δύο κατηγορίες τις στοχευμένες και τις μη-στοχευμένες ανταγωνιστικές επιθέσεις.

- Στοχευμένες επιθέσεις: Οι στοχευμένες επιθέσεις έχουν μια κλάση στόχο, Y, που θέλουν το μοντέλο-στόχος, M, να ταξινομήσει την εικόνα I της κατηγορίας X ως τέτοια. Ως εκ τούτου, ο στόχος της στοχευμένης επίθεσης είναι να κάνει το M να ταξινομήσει εσφαλμένα προβλέποντας το ανταγωνιστικό παράδειγμα, I, ως την επιδιωκόμενη κατηγορία στόχου Y αντί για την πραγματική κατηγορίας X.
- Μη-Στοχευμένες επιθέσεις: Από την άλλη πλευρά, η μη στοχευμένες επιθέσεις δεν έχουν μια κατηγορία στόχο στην οποία θέλουν το μοντέλο να ταξινομήσει την εικόνα. Αντίθετα, ο στόχος είναι απλώς να κάνουν το μοντέλο-στόχο να ταξινομήσει εσφαλμένα προβλέποντας το ανταγωνιστικό παράδειγμα, I, ως μια κατηγορία, εκτός από την αρχική κατηγορία, X.

Στη συνέχεια θα αναλύσουμε το θεωρητικό υπόβαθρο για ορισμένες σύγχρονες τεχνικές για τη δημιουργία ανταγωνιστικών παραδειγμάτων τις οποίες υλοποιήσαμε στα πλαίσια αυτής της διπλωματικής εργασίας

2.2.3.1 Fast Gradient Sign Method (FGSM)

Μία από τις πρώτες προτεινόμενες επιθέσεις για τη δημιουργία ανταγωνιστικών παραδειγμάτων προτάθηκε από τους ερευνητές της Google, Ian J. Goodfellow, Jonathon Shlens και Christian Szegedy Goodfellow (2015). Η επίθεση ονομάστηκε μέθοδος πρόσημου γρήγορης κλίσης και αποτελείται από την προσθήκη μιας γραμμικής ποσότητας μη αντιληπτού θορύβου στην εικόνα και την πρόκληση εσφαλμένης ταξινόμησης ενός μοντέλου. Αυτός ο θόρυβος υπολογίζεται πολλαπλασιάζοντας το πρόσημο της κλίσης σε σχέση με την εικόνα που θέλουμε να διαταράξουμε με ένα μικρό σταθερό έψιλον. Καθώς το έψιλον αυξάνεται, το μοντέλο είναι πιο πιθανό να εξαπατηθεί, αλλά οι διαταραχές γίνονται επίσης πιο εύκολο να εντοπιστούν. Παρακάτω φαίνεται η εξίσωση για τη δημιουργία ενός ανταγωνιστικού παραδείγματος όπου x είναι η αρχική εικόνα, ϵ είναι ένας πολύ μικρός αριθμός, Δχ είναι η συνάρτηση κλίσης, J είναι η συνάρτηση απώλειας, θ είναι τα βάρη του μοντέλου και y είναι η αληθινή ετικέτα Tsui.

$$adv_x = x + \epsilon \cdot sign(\Delta_x J(\theta, x, y))$$

Μια σημαντική ιδιότητα αυτής της εξίσωσης είναι ότι η κλίση (gradient) υπολογίζεται σε σχέση με την εικόνα εισόδου, καθώς ο στόχος είναι να δημιουργηθεί μια εικόνα που μεγιστοποιεί την απώλεια για την αρχική εικόνα της πραγματικής ετικέτας y. Στην παραδοσιακή μείωση κλίσης (για εκπαίδευση μοντέλων), η κλίση χρησιμοποιείται για την ενημέρωση των βαρών του μοντέλου, καθώς ο στόχος είναι να ελαχιστοποιηθεί η απώλεια για το μοντέλο σε ένα συγκεκριμένο σύνολο δεδομένων. Η μέθοδος Fast Gradient Sign Method προτάθηκε ως ένας γρήγορος τρόπος δημιουργίας ανταγωνιστικών παραδειγμάτων για την αποφυγή του μοντέλου, με βάση την υπόθεση ότι τα νευρωνικά δίκτυα δεν μπορούν να αντισταθούν ακόμη και σε γραμμικά ποσά διαταραχών στην είσοδο Tsui [2]Carlini (2016).

2.2.3.2 Basic Iterative Method (BIM)

Η βασική επαναληπτική μέθοδος (BIM) είναι μια απλή επέκταση της μεθόδου Fast Gradient Sign Method, όπου αντί να κάνει ένα μεγάλο βήμα, ακολουθεί μια επαναληπτική προσέγγιση εφαρμόζοντας FGSM πολλές φορές σε μια εικόνα με μέγεθος βήματος α, την αλλαγή στην τιμή pixel ανά επανάληψη. Το ανταγωνιστικό παράδειγμα που προκύπτει μπορεί στη συνέχεια να κοπεί για να περιοριστεί η μέγιστη διαταραχή για κάθε pixel.

Οι επαναληπτικές μέθοδοι όπως η BIM είναι πιο αργές, αλλά γενικά παράγουν πιο επιτυχημένες και ανεπαίσθητες διαταραχές στις εικόνες επομένως δημιουργούν καλύτερα ανταγωνιστικά παραδείγματα.

Πρώτον, μια καθαρή εικόνα X χρησιμοποιείται για προετοιμασία στην επανάληψη N=0

$$\bar{X}0 = x$$

Χρησιμοποιώντας αυτήν την εικόνα εκτελείται ένα βήμα παρόμοιο με την εξίσωση (1.2) από τη FGSM:

$$X'1 = \bar{X}0 + \alpha \cdot sign(\nabla x J(\bar{X}0, Ytrue))$$

[2] https://www.tensorflow.org/tutorials/generative/adversarial$_fgsm$

Στη συνέχεια, το ανταγωνιστικό παράδειγμα αποκόπτεται για να διασφαλιστεί ότι όλες οι τιμές εικονοστοιχείων βρίσκονται εντός των ορίων του epsilon και της μέγιστης και ελάχιστης έντασης εικονοστοιχείων:

$$\bar{X}'1 = min(255, x + \epsilon, max(0, -\epsilon, X'1))$$

Επαναλαμβάνοντας αυτά τα βήματα για N επαναλήψεις βρίσκουμε τον τελικό ανταγωνιστικό παράδειγμα.

2.2.3.3 Carlini and Wagner (CW) attack

Σε μια προσπάθεια να αναλύσουν τις υπάρχουσες ανταγωνιστικές επιθέσεις και άμυνες, οι ερευνητές στο Πανεπιστήμιο της Καλιφόρνια στο Μπέρκλεϋ, Nicholas Carlini και David Wagner το 2016 πρότεινα μια ταχύτερη και πιο ισχυρή μέθοδο για τη δημιουργία ανταγωνιστικών παραδειγμάτων Carlini (2016). Η επίθεση που πρότειναν ξεκινά με την προσπάθεια επίλυσης μιας δύσκολης μη γραμμικής εξίσωσης βελτιστοποίησης:

$$\text{min}(||\delta||_p) \text{ subject to } C(x + \delta) = t, x + \delta \in [0, 1]^n$$

Εδώ ο στόχος είναι να ελαχιστοποιηθεί ο θόρυβος (δ), που προστέθηκε στην αρχική είσοδο x, έτσι ώστε ο αλγόριθμος μηχανικής μάθησης (C) να προβλέπει την αρχική είσοδο με δέλτα (ή x + δ) ως κάποια άλλη κατηγορία t. Ωστόσο, αντί για απευθείας την παραπάνω εξίσωση, οι Carlini και Wagner προτείνουν τη χρήση μιας νέας συνάρτησης f έτσι ώστε:

$$C(x + \delta) = t \iff f(x + \delta) \leq 0$$

Αυτό συμπυκνώνει την πρώτη εξίσωση στο παρακάτω πρόβλημα:

$$\text{min}(||\delta||_p) \text{ subject to } f(x + \delta) \leq 0, x + \delta \in [0, 1]^n$$

και ακόμη περισσότερο στην εξίσωση παρακάτω:

$$\text{min}(||\delta||_p + c \cdot f(x + \delta)), x + \delta \in [0, 1]^n$$

Οι Carlini και Wagner προτείνουν στη συνέχεια τη χρήση της παρακάτω συνάρτησης στη θέση της f χρησιμοποιώντας το Z, μια συνάρτηση που καθορίζει τις πιθανότητες κλάσης για δεδομένη είσοδο x. Όταν αντικατασταθεί, αυτή η εξίσωση μπορεί να θεωρηθεί ότι βρίσκει μια κλάση στόχο που είναι πιο σίγουρη από την επόμενη πιο πιθανή κατηγορία κατά κάποιο σταθερό ποσό:

$$f(x) = ([\max_{i \neq t} Z(x)_i] - Z(x)_t)^+$$

Όταν επιλύεται χρησιμοποιώντας μείωση κλήσης, αυτή η εξίσωση είναι σε θέση να παράγει ισχυρότερα ανταγωνιστικά παραδείγματα σε σύγκριση με τη μέθοδο του πρόσημου γρήγορης κλίσης που είναι επίσης ικανή να παρακάμψει την αμυντική απόσταξη, μια άμυνα που κάποτε προτάθηκε να είναι αποτελεσματική έναντι των αντίθετων παραδειγμάτων.

2.3 Εξηγήσιμη τεχνητή νοημοσύνη

2.3.1 Ορισμός εξηγήσιμης τεχνητής νοημοσύνης και Ιστορική Αναδρομή

Τα μοντέλα μηχανικής μάθησης σημειώνουν πολύ σημαντικές επιδόσεις και για αυτό το λόγο χρησιμοποιούνται σε όλο και περισσότερες εφαρμογές. Ωστόσο, πολλά από αυτά τα μοντέλα δεν γίνονται εύκολα κατανοητά από τους ανθρώπους που αλληλεπιδρούν μαζί τους. Αυτή η κατανόηση, που αναφέρεται ως «επεξηγησιμότητα» ή «ερμηνευσιμότητα», επιτρέπει στους χρήστες να αποκτήσουν πληροφορίες σχετικά με τη διαδικασία λήψης αποφάσεων του μηχανήματος. Η κατανόηση του τρόπου λειτουργίας των πραγμάτων είναι απαραίτητη για τον τρόπο με τον οποίο πλοηγούμαστε στον κόσμο γύρω μας και είναι απαραίτητη για την ενίσχυση της εμπιστοσύνης και της αυτοπεποίθησης στα συστήματα τεχνητής νοημοσύνης.

Η εξηγήσιμη τεχνητή νοημοσύνη (XAI) είναι ένα σύνολο διαδικασιών και μεθόδων που επιτρέπει στους ανθρώπινους χρήστες να κατανοούν και να εμπιστεύονται τις εξόδους και τα αποτελέσματα που δημιουργούνται από αλγόριθμους μηχανικής μάθησης. Η εξηγήσιμη τεχνητή νοημοσύνη χρησιμοποιείται για να περιγράψει ένα μοντέλο τεχνητής νοημοσύνης, τον αναμενόμενο αντίκτυπό του και τις πιθανές προκαταλήψεις (bias). Βοηθά στον χαρακτηρισμό της ακρίβειας, της δικαιοσύνης, της διαφάνειας και των αποτελεσμάτων του μοντέλου στη λήψη αποφάσεων που βασίζονται σε τεχνητή νοημοσύνη. Το εξηγήσιμο AI είναι ζωτικής σημασίας για την οικοδόμηση εμπιστοσύνης και αυτοπεποίθησης όταν αναθέτονται σε μοντέλα τεχνητής νοημοσύνης κρίσιμες εφαρμογές. Η επεξήγηση της τεχνητής νοημοσύνης βοηθά επίσης στην υιοθέτηση μιας υπεύθυνης προσέγγισης στην ανάπτυξη της τεχνητής νοημοσύνης.

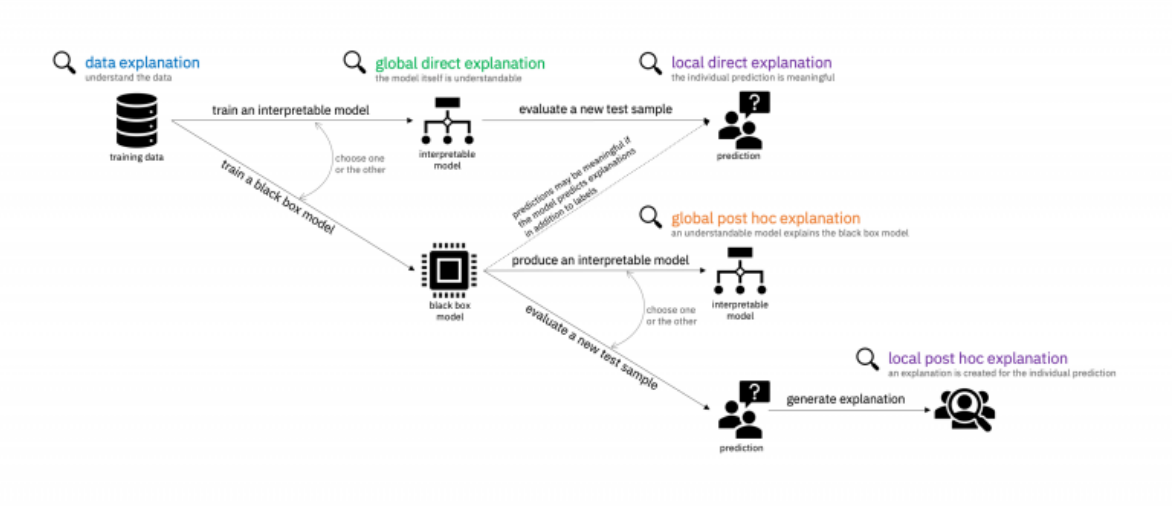
Καθώς η τεχνητή νοημοσύνη γίνεται πιο προηγμένη, οι άνθρωποι καλούνται να κατανοήσουν και να ξαναβρούν πώς ο αλγόριθμος κατέληξε σε ένα αποτέλεσμα. Ολόκληρη η διαδικασία υπολογισμού μετατρέπεται σε αυτό που συνήθως αναφέρεται ως "μαύρο κουτί" που είναι αδύνατο να ερμηνευτεί. Αυτά τα μοντέλα μαύρου κουτιού δημιουργούνται απευθείας από τα δεδομένα. Και, ακόμη και οι μηχανικοί ή οι επιστήμονες δεδομένων που δημιουργούν τον αλγόριθμο δεν μπορούν να καταλάβουν ή να εξηγήσουν τι ακριβώς συμβαίνει μέσα τους ή πώς ο αλγόριθμος τεχνητής νοημοσύνης έφτασε σε ένα συγκεκριμένο αποτέλεσμα.

Υπάρχουν πολλά πλεονεκτήματα στην κατανόηση του πώς ένα σύστημα με δυνατότητα τεχνητής νοημοσύνης έχει οδηγήσει σε μια συγκεκριμένη έξοδο. Η επεξήγηση μπορεί να βοηθήσει τους προγραμματιστές να διασφαλίσουν ότι το σύστημα λειτουργεί όπως αναμένεται, μπορεί να είναι απαραίτητο να πληρούνται τα ρυθμιστικά πρότυπα ή μπορεί να είναι σημαντικό να επιτραπεί σε όσους επηρεάζονται από μια απόφαση να αμφισβητήσουν ή να αλλάξουν αυτό το αποτέλεσμα [3].

Είναι σημαντικό να έχουμε πλήρη κατανόηση των διαδικασιών λήψης αποφάσεων της τεχνητής νοημοσύνης με την παρακολούθηση μοντέλων και τη λογοδοσία της τεχνητής νοημοσύνης και να μην την εμπιστευόμαστε τυφλά. Το εξηγήσιμο AI μπορεί να βοηθήσει τους ανθρώπους να κατανοήσουν και να εξηγήσουν τους αλγόριθμους μηχανικής μάθησης (ML), τη βαθιά μάθηση και τα νευρωνικά δίκτυα.

Τα νευρωνικά δίκτυα που χρησιμοποιούνται στη βαθιά μάθηση είναι μερικά από τα πιο δύσκολα για να τα κατανοήσει ο άνθρωπος. Η μεροληψία (bias), συχνά με βάση τη φυλή, το φύλο, την ηλικία ή την τοποθεσία, είναι ένας μακροχρόνιος κίνδυνος στην εκπαίδευση μοντέλων τεχνη-

[3] https://royalsociety.org/topics-policy/projects/explainable-ai/



Σχήμα 2.1: Explainable AI

τής νοημοσύνης. Επιπλέον, η απόδοση του μοντέλου τεχνητής νοημοσύνης μπορεί να αποκλίνει ή να υποβαθμιστεί επειδή τα πραγματικά δεδομένα διαφέρουν από τα δεδομένα στην εκπαίδευση.

Για να δώσουμε εξηγήσεις στην καθημερινή μας ζωή, βασιζόμαστε σε ένα πλούσιο και εκφραστικό λεξιλόγιο: χρησιμοποιούμε παραδείγματα και αντιπαραδείγματα, δημιουργούμε κανόνες και πρωτότυπα και επισημαίνουμε σημαντικά χαρακτηριστικά που υπάρχουν και απουσιάζουν.

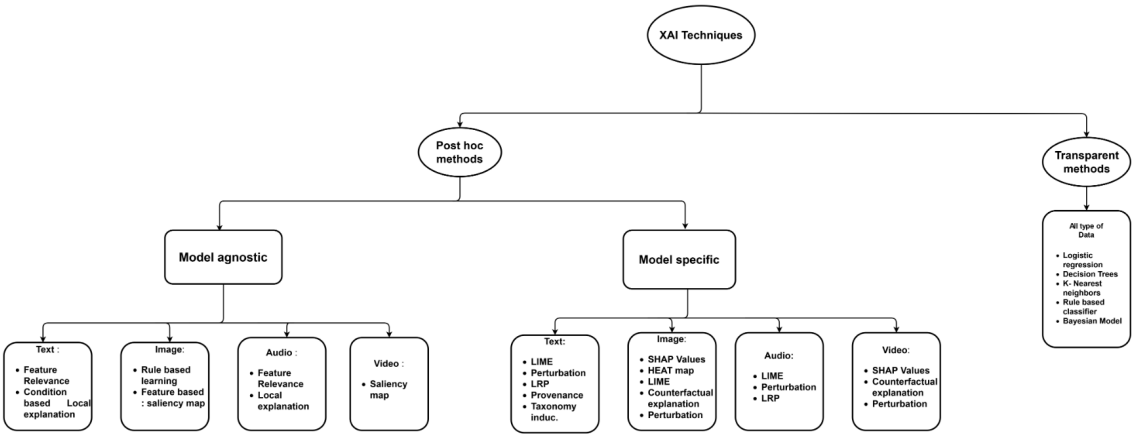
Όταν αλληλεπιδρούν με αλγοριθμικές αποφάσεις, οι χρήστες θα περιμένουν και θα απαιτούν το ίδιο επίπεδο εκφραστικότητας από την τεχνητή νοημοσύνη. Ένας γιατρός που διέγνωσε έναν ασθενή μπορεί να ωφεληθεί βλέποντας περιπτώσεις που είναι πολύ παρόμοιες ή πολύ διαφορετικές. Ένας αιτών του οποίου το δάνειο απορρίφθηκε θα θέλει να κατανοήσει τους κύριους λόγους της απόρριψης και τι μπορεί να κάνει για να ανατρέψει την απόφαση. Μια ρυθμιστική αρχή, από την άλλη πλευρά, δεν θα διερευνήσει μόνο ένα σημείο δεδομένων και μια απόφαση, θα θέλει να κατανοήσει τη συμπεριφορά του συστήματος στο σύνολό του για να διασφαλίσει ότι συμμορφώνεται με τους κανονισμούς. Ένας προγραμματιστής μπορεί να θέλει να καταλάβει πού το μοντέλο είναι περισσότερο ή λιγότερο σίγουρο ως μέσο βελτίωσης της απόδοσής του. Επομένως, όταν πρόκειται για την εξήγηση των αποφάσεων που λαμβάνονται από αλγόριθμους, δεν υπάρχει καμία προσέγγιση που να λειτουργεί καλύτερα. Υπάρχουν πολλοί τρόποι εξήγησης. Η κατάλληλη επιλογή εξαρτάται από τις συνθήκες και τις απαιτήσεις της κάθε εφαρμογής.

2.3.2 Βασικές μέθοδοι εξηγήσιμης τεχνητής νοημοσύνης

2.3.2.1 Posthoc μέθοδοι

Όταν υπάρχει μια μη γραμμική σχέση ή υπάρχει μεγαλύτερη πολυπλοκότητα δεδομένων, οι μέθοδοι posthoc είναι χρήσιμες για την ερμηνεία της πολυπλοκότητας του μοντέλου. Σε αυτήν την περίπτωση, η posthoc προσέγγιση είναι χρήσιμο εργαλείο για να εξηγήσει τι έχει μάθει το μοντέλο όταν δεν ακολουθεί μια απλή σχέση μεταξύ δεδομένων και χαρακτηριστικών.

Οι μέθοδοι ερμηνείας προσανατολισμένες στα αποτελέσματα βασίζονται στη στατιστική παρουσίαση της σύνοψης χαρακτηριστικών και στην παρουσίαση βάσει οπτικοποίησης. Η στατι-



Σχήμα 2.2: Explainable AI Classification with respect to type of data Gohel et al. (2021)

στική παρουσίαση υποδηλώνει στατιστικά στοιχεία για κάθε χαρακτηριστικό όπου η σημασία του χαρακτηριστικού ποσοτικοποιείται βάσει του βάρος στην πρόβλεψη.

Μια post-hoc μέθοδος XAI λαμβάνει εκπαιδευμένο και/ή δοκιμασμένο το μοντέλο AI ως είσοδο, στη συνέχεια δημιουργεί χρήσιμες προσεγγίσεις της εσωτερικής λογικής λειτουργίας και απόφασης του μοντέλου παράγοντας κατανοητές αναπαραστάσεις με τη μορφή βαθμολογιών σπουδαιότητας χαρακτηριστικών, συνόλων κανόνων, θερμικών χαρτών ή φυσικής γλώσσας. Πολλές posthoc μέθοδοι προσπαθούν να αποκαλύψουν σχέσεις μεταξύ των τιμών των χαρακτηριστικών και των αποτελεσμάτων ενός μοντέλου πρόβλεψης, ανεξάρτητα από τα εσωτερικά του στοιχεία. Αυτό βοηθά τους χρήστες να αναγνωρίζουν τα πιο σημαντικά χαρακτηριστικά σε μια εργασία ML, να ποσοτικοποιούν τη σημασία των χαρακτηριστικών, να αναπαράγουν αποφάσεις που λαμβάνονται από το μοντέλο του μαύρου κουτιού και να εντοπίζουν προκαταλήψεις στο μοντέλο ή τα δεδομένα

Οι posthoc μέθοδοι ταξινομούνται περαιτέρω σε αγνωστικιστικά μοντέλα και σε συγκεκριμένα μοντέλα. Οι τεχνικές συγκεκριμένου μοντέλου υποστηρίζουν περιορισμούς επεξήγησης σε σχέση με τον αλγόριθμο μάθησης και την εσωτερική δομή του δεδομένου μοντέλου βαθιάς μάθησης. Οι τεχνικές αγνωστικιστικού μοντέλου εφαρμόζουν ανάλυση των εισόδων και προβλέψεων του μοντέλου κατά ζεύγη για την κατανόηση του μηχανισμού μάθησης και για τη δημιουργία εξηγήσεων

2.3.2.2 Διαφανείς μέθοδοι

Οι διαφανείς μέθοδοι όπως η λογιστική παλινδρόμηση, τα διάνυσμα υποστήριξης μηχανής, ο Bayesian ταξινομητής, ο K πλησιέστερος γείτονας παρέχουν αιτιολόγηση με τοπικά βάρη χαρακτηριστικών. Τα μοντέλα που εμπίπτουν σε αυτήν την κατηγορία ικανοποιούν τρεις ιδιότητες που ονομάζονται αλγοριθμική διαφάνεια, δυνατότητα αποσύνθεσης και προσομοιότητα.

- Η προσομοιότητα σημαίνει τη προσομοίωση μοντέλου που πρέπει να εκτελείται από άνθρωπο. Για την προσομοίωση που ενεργοποιείται από τον άνθρωπο, η πολυπλοκότητα του μοντέλου παίζει σημαντικό ρόλο. Για παράδειγμα το μοντέλο αραιής μήτρας είναι εύκολο να ερμηνευτεί σε σύγκριση με το μοντέλο πυκνής μήτρας επειδή το μοντέλο αραιής μήτρας είναι εύκολο να δικαιολογηθεί και οπτικοποιηθεί από τον άνθρωπο.

- Η δυνατότητα αποσύνθεσης σημαίνει επεξήγηση κάθε πτυχής του μοντέλου από την εισαγωγή δεδομένων σε υπερπαραμέτρους καθώς και εγγενείς υπολογισμούς. Αυτό το χαρακτηριστικό καθορίζει τη συμπεριφορά του μοντέλου και τους περιορισμούς απόδοσης του. Τα σύνθετα χαρακτηριστικά εισόδου δεν είναι εύκολα ερμηνεύσιμα. Εξαιτίας αυτού του περιορισμού τέτοια μοντέλα δεν ανήκουν στην κατηγορία των διαφανών μεθόδων.
- Η αλγοριθμική διαφάνεια ορίζει την ερμηνευτικότητα σε επίπεδο αλγορίθμου από την εισαγωγή δεδομένων έως την τελική απόφαση ή ταξινόμηση. Η διαδικασία λήψης αποφάσεων πρέπει να γίνει κατανοητή από χρήστες με διαφάνεια. Για παράδειγμα, το γραμμικό μοντέλο θεωρείται διαφανές επειδή η γραφική παράσταση σφάλματος είναι εύκολο να απεικονιστεί και να ερμηνευτεί. Με τη βοήθεια της οπτικοποίησης ο χρήστης μπορεί να καταλάβει πώς το μοντέλο αντιδρά σε διαφορετικές καταστάσεις.

Το διαφανές μοντέλο υλοποιείται με τις ακόλουθες τεχνικές εξηγήσιμης τεχνητής νοημοσύνης:

1. Γραμμική/Λογιστική Παλινδρόμηση: Η Logistic Regression (LR) είναι ένα διαφανές μοντέλο για την πρόβλεψη μιας εξαρτημένης μεταβλητής που ακολουθεί την ιδιότητα της δυαδικής μεταβλητής. Αυτή η μέθοδος προϋποθέτει ότι υπάρχει μια ευέλικτη προσαρμογή μεταξύ του μοντέλου και και των προβλεπόμενων μεταβλητών. Για την κατανόηση της λογιστικής παλινδρόμησης μοντέλου, απαιτείται το κοινό να έχει γνώση των τεχνικών παλινδρόμησης και της μεθοδολογίας εργασίας του. Λόγω αυτών των περιορισμών, ανάλογα με τον τύπο του κοινού, η λογιστική παλινδρόμηση εμπίπτει είτε σε διαφανείς είτε σε posthoc μεθόδους. Παρόλο που η λογιστική παλινδρόμηση είναι η απλούστερη μορφή εποπτευόμενων τεχνικών ταξινόμησης, πρέπει να ληφθούν υπόψη οι μαθηματικές και στατιστικές έννοιές της.

2. Δέντρα απόφασης: Τα δέντρα απόφασης είναι ένα διαφανές εργαλείο που ικανοποιεί τη διαφάνεια σε ένα μεγάλο πλαίσιο. Είναι ένα εργαλείο ιεραρχικής λήψης αποφάσεων . Τα δέντρα αποφάσεων μικρότερης κλίμακας μπορούν εύκολα να προσομοιωθούν. Η αύξηση του αριθμού των επιπέδων στα δέντρα τα καθιστά αλγοριθμικά πιο διαφανή αλλά λιγότερο προσομοιώσιμα. Για να αντιμετωπιστεί η κακή του ιδιότητα γενίκευσης, χρησιμοποιείται ένα σύνολο εκπαιδευμένων δέντρων απόφασης. Αυτή η τροποποίηση καθιστά τα δέντρα απόφασης λιγότερο διαφανή

3. Κ-Κοντινότεροι Γείτονες: Το KNN (K-Nearest Neighbors) είναι ένα εργαλείο που βασίζεται στην ψηφοφορία που προβλέπει την κατηγορία του δείγματος δοκιμής με τη βοήθεια ψηφοφορίας των κλάσεων των Κ πλησιέστερων γειτόνων του. Η ψηφοφορία στο KNN εξαρτάται από την απόσταση και ομοιότητα μεταξύ παραδειγμάτων. Απλά KNN υποστηρίζουν διαφάνεια, αλγοριθμική διαφάνεια και ανθρωποκεντρική προσομοίωση. Η διαφάνεια του KNN εξαρτάται από τα χαρακτηριστικά, την παράμετρο Ν και τη συνάρτηση απόστασης που χρησιμοποιείται για τη μέτρηση της ομοιότητας. Υψηλότερη τιμή του Κ επηρεάζει την προσομοίωση του μοντέλου από τον άνθρωπο-χρήστη. Η συνάρτηση σύνθετης απόστασης περιορίζει τη δυνατότητα αποσύνθεσης του μοντέλου και τη διαφάνεια της αλγοριθμικής λειτουργίας.

4. Μάθηση με βάση κανόνες: Το μοντέλο που βασίζεται σε κανόνες ορίζει κανόνες για την εκπαίδευση μοντέλου. Ο κανόνας μπορεί να οριστεί στην απλή υπό όρους μορφή αν-αλλίως

ή πρώτης τάξης προγνωστική λογική. Η μορφή των κανόνων εξαρτάται από τον τύπο της βάσης γνώσεων. Οι κανόνες παρέχουν δύο πλεονεκτήματα σε αυτόν τον τύπο του μοντέλου. Πρώτον, δεδομένου ότι η μορφή των κανόνων είναι σε γλωσσικούς όρους είναι διαφανείς για να τους καταλάβει ο χρήστης. Δεύτερον, μπορεί να χειριστεί την αβεβαιότητα καλύτερα από το κλασικό μοντέλο που βασίζεται σε κανόνες Pröllochs et al. (2019). Ο αριθμός των κανόνων στο μοντέλο βελτιώνει την απόδοση του μοντέλο συμβιβάζοντας όμως την ερμηνευτικότητα και τη διαφάνεια του μοντέλου. Μοντέλο με λιγότερο αριθμό κανόνων μπορεί εύκολα να προσωμοιωθεί από τον άνθρωπο

5. Μπεϋζιανό μοντέλο: Το Μπεϋζιανό μοντέλο είναι πιθανοτικό μοντέλο με έννοια εξαρτήσεων υπό όρους μεταξύ συνόλου εξαρτημένων και ανεξάρτητων μεταβλητών. Το μοντέλο Bayesian είναι αρκετά διαφανές για τελικούς χρήστες που γνωρίζουν τις υπό όρους πιθανότητες. Τα Μπεϋζιανά μοντέλα είναι αρκετά κατάλληλα και για τις τρεις ιδιότητες δυνατότητα αποσύνθεσης, αλγοριθμική διαφάνεια και ανθρώπινη προσομοίωση. Η εξάρτηση σύνθετης μεταβλητής μπορεί να επηρεάσει τη διαφάνεια και την ανθρώπινη προσομοίωση για τα Μπεϋζιανά μοντέλο.

2.3.2.3 Τεχνικές εξηγήσιμης τεχνητής νοημοσύνης

Τα συγκεκριμένα μοντέλα εξηγήσιμης τεχνητής νοημοσύνης υλοποιούνται χρησιμοποιώντας τις ακόλουθες τεχνικές.

1. Συνάφεια χαρακτηριστικών: Είναι πάντα σημαντικό να ανακαλύπτουμε τα πια επιδραστικά χαρακτηριστικά που είναι κρίσιμα για τη λήψη αποφάσεων. Για αυτά τα χαρακτηριστικά, εισάγεται η έννοια της σημασίας. Η σημασία των χαρακτηριστικών δείχνει το συντελεστή επιρροής κάθε χαρακτηριστικού σε παραγόμενες αποφάσεις Rajani et al. (2019). Μαζί με τη σημασία των χαρακτηριστικών, η συσχέτιση μεταξύ των χαρακτηριστικών είναι επίσης χρήσιμο για επεξήγηση. Στην ιατρική διάγνωση που βασίζεται σε μοντέλα τεχνητής νοημοσύνης, η συσχέτιση χαρακτηριστικών στα δεδομένα εκπαίδευσης είναι μια από τις κινητήριες δυνάμεις για τη διάγνωση

2. Εξήγηση βάσει συνθηκών: Απαιτείται εξήγηση βάσει συνθηκών όταν χρειάζεται να απαντήσουμε ερωτήματα του τύπου «γιατί», «γιατί παρότι» και «γιατί ενώ». Ορισμένες συγκεκριμένες παρατηρούμενες εισροές διαδραματίζουν βασικό ρόλο στην αιτιολόγηση της πρόβλεψης. Ρωτώντας ερωτήσεις που σχετίζονται με το ”Γιατί?”, το μοντέλο θα παρέχει όλες τις δυνατές εξηγήσεις με ένα σύνολο συνθηκών. Αυτό το σύνολο συνθηκών δημιουργείται με φαινόμενα πληρότητας. Το ”Και αν” παρέχει υποθετική αιτιολογία για αντιπραγματική αιτιολόγηση. Ένα απλό λογικό μοντέλο μετατρέπει τις εισόδους των χρηστών στη μορφή εισόδων που βασίζονται σε περιορισμούς και παρέχει αιτιολόγηση εάν οι περιορισμοί ικανοποιούνται με τη μορφή συνθηκών.

3. Μάθηση με βάση κανόνες: Απαιτείται επεξήγηση επειδή η έξοδος του μοντέλου ML είναι αριθμητική και το νευρωνικό δίκτυο είναι πολύ περίπλοκο σε σημείο που ο κανονικός χρήστης δεν μπορεί να κατανοήσει την πολυπλοκότητα των υπερπαραμέτρων και την επίδρασή

τους στην τελική πρόβλεψη. Αφού κανείς αποκτήσει κάποια εσωτερική κατανόηση των εκπαιδευμένων μοντέλο και ερμηνευσιμότητα των αποτελεσμάτων, κατάλληλη προσέγγιση είναι να εξηγήσει τα παράγωγα αποτελέσματα σε πελάτες και αφελείς χρήστες, είναι η μετάφραση αυτών των γνώσεων σε κανόνες τέτοιους που μπορεί να παρέχει πλήρης διαφάνεια για την εξηγήσιμη τεχνητή νοημοσύνη. Pröllochs et al. (2019). Μόλις πλαισιωθούν οι κανόνες για όλες τις πιθανές προβλέψεις, γίνεται ακόμη και το πιο πολύπλοκο νευρικό δίκτυο, διαφανές μοντέλο.

4. Χάρτης προεξοχής βάσει χαρακτηριστικών: Οι χάρτες προεξοχής χρησιμοποιούνται γενικά κομμάτια εφαρμογών με επεξεργασία εικόνας για να δείξουν ποια μέρη των καρέ από βίντεο ή των εικόνων είναι οι πιο σημαντικές για την απόφαση του CNN. Ο χάρτης προεξοχής εξηγήσιμης τεχνητής νοημοσύνης είναι ένα εργαλείο που είναι χρήσιμο για την διερεύνηση της εσωτερικής λειτουργίας των DNN. Ο Υπολογισμός της κλίσης με χρήση αλγόριθμων αντίστροφης διάδοσης χρησιμοποιούνται ως ποσοτικά μέτρα για την προβολή της έντασης χρωμάτων στο επίπεδο.

2.3.2.4 Τεχνικές αγνωστικών μοντέλων

Τα αγνωστικά μοντέλα εξηγήσιμης τεχνητής νοημοσύνης υλοποιούνται χρησιμοποιώντας τις ακόλουθες τεχνικές.

1. Τοπικό ερμηνευτικό μοντέλο-αγνωστικές επεξηγήσεις - LIME: Ο αγνωστικισμός του μοντέλου καθορίζει την ιδιότητα που είναι ικανή το LIME να παρέχει αιτιολόγηση για κάθε τύπο πρόβλεψης μοντέλου εποπτευόμενης μάθησης. Αυτή η τεχνική είναι εφαρμόσιμη για κάθε είδους δεδομένων όπως εικόνα, κείμενο και βίντεο. Αυτό σημαίνει ότι το LIME είναι σε θέση να χειριστεί οποιοδήποτε εποπτευόμενο μοντέλο μάθησης και να παρέχει αιτιολόγηση.

   Το LIME παρέχει τοπικές βέλτιστες επεξηγήσεις που υπολογίζουν σημαντικά χαρακτηριστικά γύρω από τη συγκεκριμένη περίπτωση που πρέπει να εξηγηθεί. Από προεπιλογή παράγει 5000 δείγματα του διανύσματος χαρακτηριστικών που ακολουθούν κανονική κατανομή. Μετά την παραγωγή δειγμάτων κανονικής κατανομής βρίσκει τις μεταβλητές-στόχους για δείγματα των οποίων οι αποφάσεις επεξηγούνται από το LIME.

   Μετά τη λήψη του τοπικού δημιουργημένου συνόλου δεδομένων και των προβλέψεών τους, εκχωρεί βάρη σε κάθε γραμμή ανάλογα με το πόσο κοντά είναι από τα πρωτότυπα δείγματα. Στη συνέχεια, χρησιμοποιεί μια τεχνική επιλογής χαρακτηριστικών όπως h lasso ή το PCA (Principle Component Analysis) για να αποκτήσετε τα σημαντικά χαρακτηριστικά

2. Διατάραξη: Η διατάραξη βοηθά στη δημιουργία επιθυμητών οδηγών επεξήγησης και στην ανάλυσης της επίδρασης των διαταραγμένων χαρακτηριστικών στον δεδομένο στόχο. Παρέχει σύνοψη όλων των χαρακτηριστικών για τα δοσμένα διαταραγμένα αποτελέσματα. Η μέθοδος διατάραξης είναι εύκολη στην εφαρμογή και δεν εφαρμόζεται σε συγκεκριμένη αρχιτεκτονική μοντέλου. Αυτή η μέθοδος μπορεί να εφαρμοστεί σε οποιοδήποτε τύπο μοντέλου AI/ML. Μειονέκτημα της συγκεκριμένης μεθόδου είναι, ότι είναι υπολογιστικά ακριβή αν ο αριθμός των χαρακτηριστικών είναι σχετικά μεγαλύτερος από τον κανονικό μέσο όρο. Καθώς ο αριθμός των χαρακτηριστικών είναι μεγαλύτερος, χρειάζεται περισσότερος χρόνος για την αξιολόγηση του συνδυασμού όλων των χαρακτηριστικών.

3. Διάδοση συνάφειας σε επίπεδο στρώματος - LRP: Το LRP είναι χρήσιμο για την αποσυσκευασία πολύπλοκων νευρωνικών δικτύων. Προβάλλει προβλέψεις προς τα πίσω στο νευρωνικό δίκτυο. Για προς τα πίσω προτάσεις σχεδιάζονται συγκεκριμένοι κανόνες

4. Επαγωγή προέλευσης και ταξινόμησης: Η επαγωγή προέλευσης και ταξινόμησης είναι τεχνικές που βασίζονται σε λογικά τεκμήρια, για την αιτιολόγηση του αποτελέσματος που βασίζεται σε μερικώς συμπληρωματικά αποτελέσματα.

2.3.3 Σύνδεση εξηγήσιμης τεχνητής νοημοσύνης με Ανταγωνιστικές Επιθέσεις

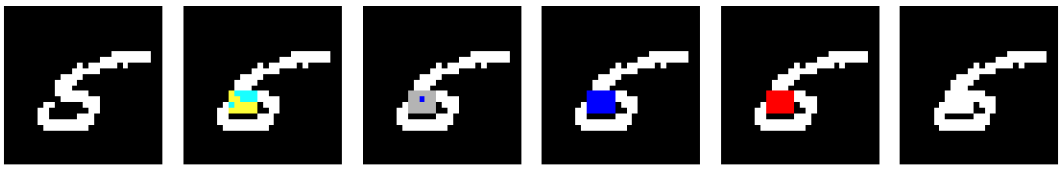
Μέχρι στιγμής έχουμε περιγράψει ένα γενικό θεωρητικό υπόβαθρο όσον αφορά τις ανταγωνιστικές επιθέσεις σε ΒΝΔ και την εξηγήσιμη τεχνητή νοημοσύνη. Όπως έχει φανεί από τις μέχρι τώρα αναφορές και τα δύο αυτά αντικείμενα αποτελούν αντικείμενα αιχμής στον τομέα της μηχανικής μάθησης. Λόγω του τεράστιου ενδιαφέροντος που παρουσιάζουν έχουν κεντρίσει σε μεγάλο βαθμό το ενδιαφέρον της ερευνητικής κοινότητας. Τα τελευταία χρόνια όμως έχει αρχίσει να κεντρίζει το ενδιαφέρον και ο συνδυασμός των δύο αυτών αντικειμένων. Το να καταφέρουμε να εξηγήσουμε γιατί ένα μοντέλο καταλήγει σε μια συγκεκριμένη πρόβλεψη μπορεί να μας βοηθήσει τόσο στο να δημιουργήσουμε ανταγωνιστικές επιθέσεις που θα δημιουργούν πιο αποτελεσματικά ανταγωνιστικά παραδείγματα όσο και να δημιουργήσουμε μοντέλα που θα είναι πολύ πιο ανθεκτικά στις επιθέσεις. Παρότι αφορά ένα αρκετά καινούργιο ερευνητικό πεδίο τα τελευταία χρόνια έχουν υπάρξει διάφορες προσεγγίσεις γύρω από το συγκεκριμένο ζήτημα.

Το 2019 οι (Dombrowski et al., 2019) δείξανε ότι παρότι οι μέθοδοι εξήγησης σκοπεύουν να κάνουν τα νευρωνικά δίκτυα πιο έμπιστα και ερμηνεύσιμα, υπάρχει μια ιδιότητα των μεθόδων εξήγησης που είναι ανησυχητική και για τους δύο παραπάνω σκοπούς. Δείξανε ότι οι επεξηγήσεις (explanations) μπορούν να χειραγωγηθούν αυθαίρετα, εισάγοντας στα δεδομένα εισόδου διαταραχές που είναι σχεδόν απαρατήρητες οπτικά και κρατάνε την έξοδο του δικτύου σχετικά σταθερή. Αποδείξανε ότι το φαινόμενο αυτό συνδέεται με μια συγκεκριμένη γεωμετρική ιδιότητα των νευρωνικών δικτύων. Οι Ignatiev (2019) πρότειναν μια ισχυρή θεωρητική σχέση μεταξύ των επεξηγήσεων και των ανταγωνιστικών παραδειγμάτων δείχνοντας ότι σχετίζονται με μια γενικευμένη μορφή δυαδικότητας συνόλων. Επίσης πρότειναν αλγορίθμους που επιτρέπουν τον υπολογισμό των ανταγωνιστικών παραδειγμάτων από επεξηγήσεις και το αντίστροφο.

Ακόμα μια πολύ ενδιαφέρουσα προσέγγιση έγινε από τους Cantareira et al. (2021) που πρότειναν ένα οπτικό πλαίσιο για την διερεύνηση της αντίληψης ενός μοντέλου σχετικά με τις ανταγωνιστικές επιθέσεις και τα πραγματικά δεδομένα και σχετικά με τις κλάσεις τους. Δείξανε ότι παρατηρώντας αυτά τα δεδομένα μπορούσαν να εντοπίσουν τις περιοχές εκμετάλλευσης σε ένα μοντέλο, επιτρέποντας περαιτέρω μελέτη των ευάλωτων χαρακτηριστικών στα δεδομένα εισόδου και την βελτίωση της εκπαίδευσης και της αρχιτεκτονικής του μοντέλου. Στην ουσία προτείναν έναν οπτικό τρόπο για τον εντοπισμό σε ένα μοντέλου του επιπέδου (layer) εκείνου που αποτελεί το σημείο καμπής και παίζει σημαντικό ρόλο στην διαφοροποίηση της κατηγοριοποίησης μεταξύ ανταγωνιστικού παραδείγματος και πραγματικής εικόνας.

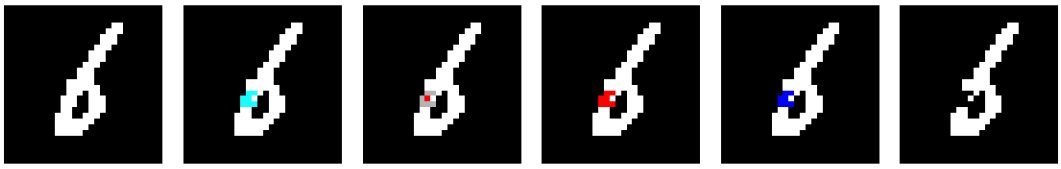
Τέλος αξίζει να αναφερθούμε στην εργασία των Chakraborty et al. (2022) οι οποίοι πρότειναν μια νέα μέθοδο που επεκτείνει το Grad-CAM από επεξηγήσεις που βασίζονται σε παραδείγματα σε μια μέθοδο για την εξήγηση της συνολικής συμπεριφοράς του μοντέλου. Για να το πετύχουν αυτό, πρότειναν δύο νέες μετρικές, (i)την Μέση παρατηρούμενη ανομοιότητα (MOD) και (ii) τη

Σχήμα 2.3: Explanation manipulation



(a) digit “5” (b) patch area (c) an XP (d) all XP’s (e) a CE (f) an AE (“6”)

Figure 1: An example of digit five.



(a) digit “6” (b) patch area (c) an XP (d) all XP’s (e) a CE (f) an AE (“5”)

Figure 2: An example of digit six.

Σχήμα 2.4: Explanation adn Adversarial examples

Διακύμανση στην ανομοιότητα (VID), για γενίκευση του μοντέλου. Αυτές οι μετρικές υπολογίζονται συγκρίνοντας μια μέτρηση Normalized Inverted Structural Similarity Index (NISSIM) του χάρτη θερμότητας που δημιουργήθηκε από Grad-CAM για δείγματα από το αρχικό σύνολο δοκιμών και δείγματα από το ανταγωνιστικό σύνολο δοκιμών. Χρησιμοποιώντας αυτές τις μετρικές παρατηρήσαν μια σταθερή μετατόπιση στη περιοχή που επισημαίνεται στον χάρτη θερμότητας Grad-CAM, αντικατοπτρίζοντας την συμμετοχή στη λήψη αποφάσεων, σε όλα τα μοντέλα ανταγωνιστικών επιθέσεων. Ισχυρίζονται ότι η προτεινόμενη μέθοδος μπορεί να χρησιμοποιηθεί για να κατανοήσουν τις αντίπαλες επιθέσεις και να εξηγήσουν τη συμπεριφορά μοντέλων CNN μαύρου κουτιού για ανάλυση εικόνας.

Κεφάλαιο 3

Σχεδιασμός και Υλοποίηση

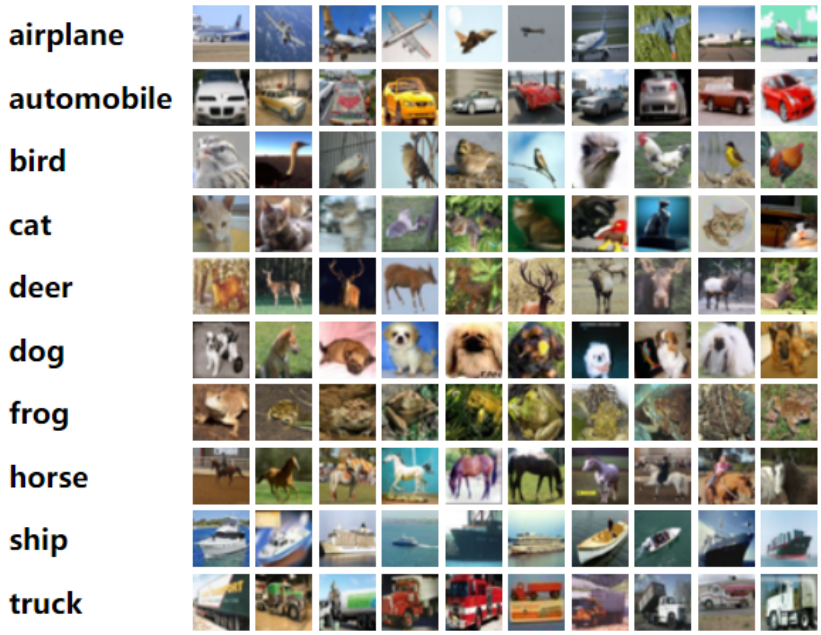
Στο παρών κεφάλαιο αναλύονται οι σχεδιαστικές επιλογές και τα βήματα υλοποίησης που εκτελέστηκαν κατά την ανάπτυξη της διπλωματικής εργασίας. Ο σχεδιασμός κι η υλοποίηση της διπλωματικής εργασίας εξελίχθηκε σε 4 βασικά στάδια: α) την εύρεση κι επεξεργασία δεδομένων, β) τη σχεδίαση του μοντέλου, γ) την υλοποίηση των ανταγωνιστικών επιθέσεων δ) την υλοποίηση του αλγορίθμου εξηγήσιμης τεχνητής νοημοσύνης ε)την ανάπτυξη και υλοποίηση του μοντέλου. Παρακάτω παρουσιάζονται αναλυτικότερα τα στάδια αυτά παραθέτοντας μαζί τα στάδια β) και ε) τα οποία αφορούν το σχεδιασμό και την υλοποίηση των πειραμάτων.

3.1 Δεδομένα

3.1.1 Σετ Δεδομένων - Datasets

Για κάθε μοντέλο βαθιάς μάθησης αφετηρία αποτελεί το κομμάτι των δεδομένων καθώς σε μεγάλο βαθμό κρίνουν την επιτυχία ή όχι του μοντέλου. Σε πολλές περιπτώσεις η εύρεση και η συλλογή των κατάλληλων δεδομένων αποτελεί μεγάλο πρόβλημα στο χώρο της βαθιάς μάθησης. Ορισμένοι από τους παράγοντες που παίζουν ρόλο για την επιλογή του κατάλληλου συνόλου δεδομένων είναι οι παρακάτω. Αρχικά είναι πάρα πολύ σημαντικό τα δεδομένα να είναι έγκυρα και με σωστό τρόπο κατηγοριοποιημένα διότι στην περίπτωση που βασίσουμε το μοντέλο μας σε ψεύτικα ή λάθος δεδομένα τα αποτελέσματα που θα προκύψουν δεν θα έχουν καμία αξία χρήσης. Ακόμα σημαντικό παράγοντα αποτελεί το μέγεθος του συνόλου δεδομένων καθώς όσο μεγαλύτερο είναι τόσο πιο στιβαρό θα είναι μοντέλο και τόσο μεγαλύτερη θα είναι η δυνατότητα που θα έχουμε για να γενικεύσουμε τα αποτελέσματα μας. Τέλος η κατανομή των δεδομένων στις διάφορες κλάσεις του συνόλου μας, επηρεάζει σε μεγάλο βαθμό τα τελικά αποτελέσματα. Ένας συνδυασμός των παραπάνω, αλλά και άλλων, παραγόντων είναι αυτός που τελικά μας βοηθάει στο να καταλήξουμε στο κατάλληλο σύνολο δεδομένων για το εκάστοτε πρόβλημα.

Στα πλαίσια της παρούσας διπλωματικής επιλέξαμε αρχικά να χρησιμοποιήσουμε για δεδομένα μας εικόνες. Ο λόγος που πήραμε αυτή την επιλογή και δεν χρησιμοποιήσαμε κάποια άλλη μορφή δεδομένων όπως για παράδειγμα δεδομένα ήχου, βίντεο ή φυσικής γλώσσας ήταν γιατί παρατηρήσαμε ότι στην βιβλιογραφία η συντριπτική πλειοψηφία των μελετών που έχουν γίνει γύρω από τις ανταγωνιστικές επιθέσεις, στις οποίες επικεντρώνεται η εργασία μας, αφορούσαν εικόνες. Επομένως μη έχοντας προηγούμενη εμπειρία πάνω στο αντικείμενο των ανταγωνιστικών επιθέσεων κρίναμε ότι θα ήταν πιο εύκολο να ξεκινήσουμε από επιθέσεις σε εικόνες που θα μπορούσαμε να βρούμε αρκετή βοήθεια στην υπάρχουσα βιβλιογραφία. Συγκεκριμένα το σύνολο δεδομένων που χρησιμοποιήσαμε ήταν το CIFAR-10 Krizhevsky (2009) το οποίο αποτελείται από 60000 εικό-



Σχήμα 3.1: Cifar 10

νες διαστάσεων 32*32 η καθεμία. Αποτελείται από 10 κλάσεις (Αεροπλάνο, Αυτοκίνητο, Πουλί, Γάτα, Ελάφι, Σκύλος, Βάτραχος, Άλογο, Πλοίο, Φορτηγό) με 6000 εικόνες να αντιστοιχούν στην κάθε κλάση. Το CIFAR-10 είναι χωρισμένο σε 50000 εικόνες εκπαίδευσης και 10000 εικόνες δοκιμής. Το συνολικό μέγεθος του CIFAR-10 είναι 163MB. Ο λόγος που επιλέχθηκε το συγκεκριμένο σύνολο δεδομένων έναντι κάποιου άλλου από τα σύνολα δεδομένων που χρησιμοποιούνται συχνά στην κατηγοριοποίηση εικόνων, όπως τα ImageNet, MNIST ή το CIFAR-100 είναι το μέγεθος και ο αριθμός των κλάσεων. Καταρχάς όσον αφορά το μέγεθος, επιλέξαμε μεν ένα αρκετά μεγάλο μέγεθος (60.000 εικόνες) ούτως ώστε να είναι στιβαρό το μοντέλο μας από την άλλη επειδή αντιμετωπίσαμε κάποιους περιορισμούς σχετικά με το υπολογιστικό περιβάλλον που αναπτύξαμε το πρόγραμμα μας, για να γλιτώσουμε χρόνους εκπαίδευσης επιλέξαμε ένα σχετικά μικρό σε μέγεθος σύνολο δεδομένων. Σχετικά με τον αριθμό των κλάσεων επιλέξαμε ένα σύνολο δεδομένων με μικρό αριθμό κλάσεων για να μπορούμε και με οπτικό τρόπο να αξιολογήσουμε τα αποτελέσματα μας.

3.2 Συνελικτικό Νευρωνικό Δίκτυο

3.2.1 Επιλογή Δεδομένων Εισόδου

Μια βασική επιλογή που κάναμε για τα δεδομένα εισόδου, όπως αναφέραμε και παραπάνω, ήταν να χρησιμοποιήσουμε σαν δεδομένα εικόνες. Η επιλογή αυτή καθορίζει σε μεγάλο βαθμό και την αρχιτεκτονική που διαμορφώσαμε καθώς θα χρειαζόταν αρκετά διαφορετική προσέγγιση αν παραδείγματος χάρη είχαμε σαν δεδομένα εισόδου δεδομένα κειμένου. Από την στιγμή που αποφασίσαμε να χρησιμοποιήσουμε σαν δεδομένα εισόδου εικόνες η επιλογή του κατάλληλου είδους νευρωνικών δικτύων ήταν αρκετά απλή. Καθώς τα συνελικτικά νευρωνικά δίκτυα αποτελούν το πιο διαδεδομένο και κορυφαίο από πλευράς επιδόσεων εργαλείο για προβλήματα που σχετίζονται με την κατηγοριοποίηση εικόνων.

Ὸπως αναφέραμε και παραπάνω το σύνολο δεδομένων που επιλέξαμε περιέχει εικόνες διαστάσεων 32*32, όμως τα προ εκπαιδευμένα CNN μοντέλα που χρησιμοποιήσαμε απαιτούν τα δεδομένα εισόδου να έχουν συγκεκριμένες διαστάσεις που ορίζονται με βάση τα χαρακτηριστικά του πρώτου τους επιπέδου (layer). Για να λύσουμε το συγκεκριμένο πρόβλημα χρησιμοποιήσαμε τα εργαλεία που μας προσφέρει η Pytorch. Συγκεκριμένα χρησιμοποιήσαμε την κλάση transforms.Compose η οποία μας επιτρέπει να κάνουμε κάποιες μετατροπές στις εικόνες εισόδου μας. Οι μετατροπές που κάναμε ήταν η προσαρμογή του μεγέθους των εικόνων στις επιθυμητές διαστάσεις μέσω του Resize και η κανονικοποίηση των εικόνων τυποποιώντας τες μέσω του Normalize. Η ιδέα της κανονικοποίησης δεδομένων είναι μια γενική έννοια που αναφέρεται στην πράξη μετατροπής των αρχικών τιμών ενός συνόλου δεδομένων σε νέες τιμές. Οι νέες τιμές τυπικά κωδικοποιούνται σε σχέση με το ίδιο το σύνολο δεδομένων και κλιμακώνονται με κάποιο τρόπο. Για αυτόν τον λόγο, ένα άλλο όνομα για την κανονικοποίηση δεδομένων που χρησιμοποιείται μερικές φορές είναι η κλιμάκωση χαρακτηριστικών. Αυτός ο όρος αναφέρεται στο γεγονός ότι κατά την κανονικοποίηση δεδομένων, συχνά μετατρέπουμε διαφορετικά χαρακτηριστικά ενός δεδομένου συνόλου δεδομένων σε παρόμοια κλίμακα. Η τυποποίηση δεδομένων είναι ένας συγκεκριμένος τύπος τεχνικής κανονικοποίησης. Μερικές φορές αναφέρεται ως κανονικοποίηση βαθμολογίας z. Η βαθμολογία z, γνωστή και ως τυπική βαθμολογία, είναι η μετασχηματισμένη τιμή για κάθε σημείο δεδομένων. Για να ομαλοποιήσουμε ένα σύνολο δεδομένων χρησιμοποιώντας τυποποίηση, παίρνουμε κάθε τιμή x μέσα στο σύνολο δεδομένων και τη μετατρέπουμε στην αντίστοιχη τιμή z χρησιμοποιώντας τον ακόλουθο τύπο:

$$z = \frac{x - mean}{std}$$

Αφού εκτελέσουμε αυτόν τον υπολογισμό σε κάθε τιμή x μέσα στο σύνολο δεδομένων μας, έχουμε ένα νέο κανονικοποιημένο σύνολο τιμών z. Ο μέσος όρος και οι τιμές τυπικής απόκλισης αφορούν το σύνολο δεδομένων ως σύνολο. Αυτή η διαδικασία τυποποίησης μετατρέπει τη μέση τιμή του συνόλου δεδομένων σε 0 και την τυπική απόκλιση σε 1. Είναι σημαντικό να σημειωθεί ότι όταν κανονικοποιούμε ένα σύνολο δεδομένων, συνήθως ομαδοποιούμε αυτές τις λειτουργίες ανά χαρακτηριστικό. Αυτό σημαίνει ότι η μέση τιμή και η τυπική απόκλιση είναι σχετικές με κάθε σύνολο χαρακτηριστικών που κανονικοποιείται. Επομένως εδώ που εργαζόμαστε με εικόνες, τα χαρακτηριστικά είναι τα έγχρωμα κανάλια RGB, επομένως κανονικοποιούμε κάθε κανάλι χρώματος σε σχέση με τις τιμές μέσης και τυπικής απόκλισης που υπολογίζονται σε όλα τα pixel σε κάθε εικόνα για το αντίστοιχο κανάλι χρώματος.

3.2.2 Επιλογή Αρχιτεκτονικής

Σχετικά με την επιλογή της κατάλληλης αρχιτεκτονικής ένα ερώτημα που κλιθήκαμε να απαντήσουμε ήταν αν θα δημιουργήσουμε εμείς μια δικιά μας αρχιτεκτονική από την αρχή ή αν θα χρησιμοποιήσουμε κάποια έτοιμη. Η απάντηση που δώσαμε χωρίς ιδιαίτερη δυσκολία ήταν να χρησιμοποιήσουμε κάποια από τις ήδη υπάρχουσες αρχιτεκτονικές για τους λόγους που αναφέρουμε παρακάτω. Αρχικά οι αρχιτεκτονικές αυτές έχουν σχεδιαστεί και δοκιμαστεί από ομάδες ερευνητών και οργανισμούς με πολυετή εμπειρία πάνω στον συγκεκριμένο τομέα μετά από χρόνια έρευνας. Αυτό έχει σαν αποτέλεσμα οι αρχιτεκτονικές αυτές να πετυχαίνουν τα βέλτιστα αποτελέσματα στον τομέα της κατηγοριοποίησης εικόνων. Επομένως θεωρήσαμε ότι δεν έχουμε ούτε τις

απαραίτητες γνώσεις ούτε την εμπειρία για να μπορέσουμε να δημιουργήσουμε κάποια αρχιτεκτονική από την αρχή που θα μπορούσε να ανταγωνιστεί τις κορυφαίες αρχιτεκτονικές που υπάρχουν ήδη.

Ακόμα χρησιμοποιώντας κάποια ήδη υπάρχουσα αρχιτεκτονική μας δινόταν η δυνατότητα να μειώσουμε σε πολύ μεγάλο βαθμό τους απαιτούμενους χρόνους εκπαίδευσης. Αυτό συνέβη διότι μπορούσαμε να χρησιμοποιήσουμε προ εκπαιδευμένες αρχιτεκτονικές που είχαν εκπαιδευτεί για πολλές εποχές σε τεράστια σύνολα δεδομένων και πετυχαίναν πολύ καλά αποτελέσματα. Σε αυτό το σημείο θα πρέπει να τονίσουμε ότι οι προ εκπαιδευμένες αυτές αρχιτεκτονικές δεν είχαν σχεδιαστεί ακριβώς για το σκοπό που θέλαμε εμείς να τις χρησιμοποιήσουμε, επομένως είχαν εκπαιδευτεί σε άλλα σύνολα δεδομένων. Παρόλα αυτά χάρη στην μεταφορά γνώσης (Transfer Learning) μπορούσαμε κάνοντας πολύ μικρές αλλαγές στο προ εκπαιδευμένο μοντέλο και εκπαιδεύοντας το στο δικό μας σύνολο δεδομένων για πολύ μικρό αριθμό εποχών να πετύχουμε πάρα πολύ καλά αποτελέσματα. Η αλλαγή που χρειάστηκε να κάνουμε ήταν να τροποποιήσουμε το τελευταίο γραμμικό επίπεδο του μοντέλου ούτως ώστε ο αριθμός χαρακτηριστικών εξόδου του να ταυτίζεται με τον αριθμό των κλάσεων του συνόλου δεδομένων μας.

Επομένως χρησιμοποιώντας κάποια ήδη υπάρχουσα αρχιτεκτονική θα μπορούσαμε να πετύχουμε καλύτερα αποτελέσματα, ελαχιστοποιώντας παράλληλα τους χρόνους εκπαίδευσης και μειώνοντας τις πιθανότητες κάποιου σφάλματος που θα επηρέαζε αρνητικά τα αποτελέσματα. Για αυτούς τους λόγους επιλέξαμε να μην σχεδιάσουμε από την αρχή κάποια δική μας αρχιτεκτονική αλλά όπως αναφέραμε να χρησιμοποιήσουμε κάποια έτοιμη. Στην συνέχεια όμως προέκυψε το ερώτημα ποιες έτοιμες αρχιτεκτονικές θα χρησιμοποιούσαμε. Μετά από διάφορες δοκιμές καταλήξαμε σε δύο αρχιτεκτονικές την VGG και συγκεκριμένα την VGG16 με την οποία πετύχαμε Accuracy 93%, και την ResNet και συγκεκριμένα την ResNet50 με την οποία πετύχαμε Accuracy 93,78%.

3.2.3 Έξοδος προγράμματος και μετρικές αξιολόγησης

Στην συνέχεια έπρεπε να αποφασίσουμε ποιες μετρικές θα χρησιμοποιούσαμε και ποια θα ήταν έξοδος του μοντέλου μας. Όπως στις περισσότερες περιπτώσεις στην βιβλιογραφία έτσι και εμείς βασίσαμε την εκπαίδευση των μοντέλου μας στο accuracy καθώς η μετρική αυτή αποτελεί το μέτρο σύγκρισης για την επιτυχία των κατηγοριοποιητών εικόνων. Βέβαια δεν αρκεί μόνο το accuracy καθώς μας δείχνει μόνο ποιο ποσοστό των προβλέψεων ήταν σωστό, μια ακόμα πολύ χρήσιμη μετρική που χρησιμοποιείται συμπληρωματικά με το accuracy είναι το loss function. Το loss function μας δείχνει πόσο απέχει μια λανθασμένη πρόβλεψη από την σωστή, καθώς μπορεί δύο προβλέψεις να είναι λανθασμένες αλλά η μία να είναι πολύ πιο κοντά στην πραγματική κατηγορία από την άλλη. Σαν έξοδο από την εκπαίδευση εξάγουμε το accuracy στα train και test δεδομένα.

3.2.4 Αποθήκευση Μοντέλου

Όπως αναφέραμε και σε προηγούμενα κεφάλαια η παρούσα διπλωματική επικεντρώθηκε στις ανταγωνιστικές επιθέσεις μέσω της δημιουργίας ανταγωνιστικών παραδειγμάτων από τα δεδομένα εισόδου, καθώς επίσης και στην προσπάθεια οπτικής επεξήγησης των αποτελεσμάτων και σύγκρισης των επεξηγήσεων των αρχικών εικόνων και των ανταγωνιστικών παραδειγμάτων. Επομένως χρειαζόμασταν να αποθηκεύσουμε τα εκπαιδευμένα μοντέλα και να μπορούμε να τα φορ-

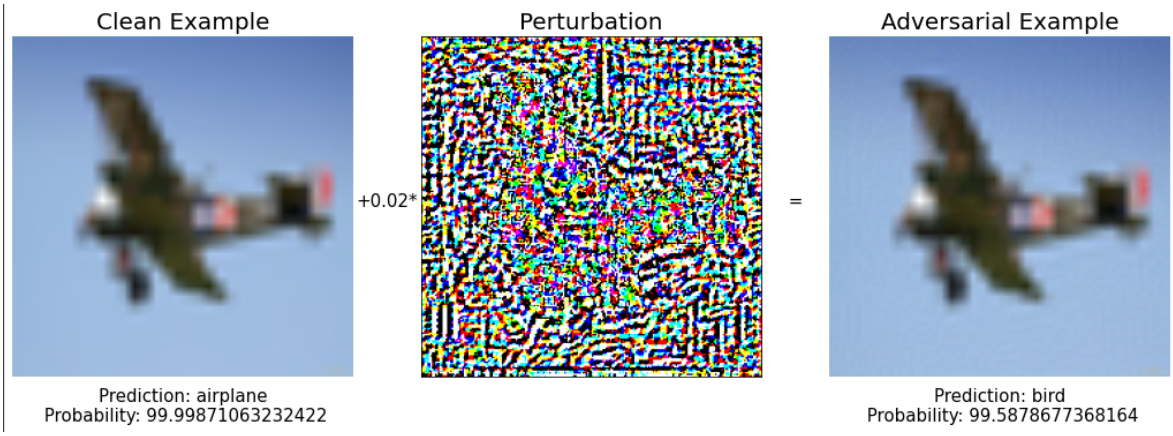
τώσουμε για να κάνουμε προβλέψεις πάνω στα δεδομένα μας. Για να το πετύχουμε αυτό μέσω του torch.save αποθηκεύαμε ολόκληρο το εκπαιδευμένο μοντέλο για να μπορούμε ανά πάσα στιγμή να ανατρέξουμε σε αυτό για να κάνουμε προβλέψεις. Αφού συνδέσαμε το google drive στο περιβάλλον διεξαγωγής των πειραμάτων (Google Colab και Kaggle) αποθηκεύσαμε σε αυτό το μοντέλο σε μορφή .pth που είναι η προεπιλεγμένη μορφή αποθήκευσης ολόκληρων μοντέλων στην Pytorch. Με αντίστοιχο τρόπο χρησιμοποιώντας το torch.load φορτώναμε το εκπαιδευμένο μοντέλο χρειαζόμασταν. Τέλος αξίζει να αναφέρουμε ότι ο λόγος που αποθηκεύαμε ολόκληρο το εκπαιδευμένο μοντέλο και όχι κάποια βάρη του ή κάποιο checkpoint ήταν ότι δεν μας ενδιέφερε να συνεχίσουμε σε κάποια φάση την εκπαίδευση του μοντέλου, ούτε σκοπεύαμε να χρησιμοποιήσουμε την γνώση που είχε αποκτήσει το μοντέλο, σε κάποια φάση της εκπαίδευσης του, για να τροφοδοτήσουμε κάποιο άλλο με μεταφορά γνώσης. Αυτό που μας ενδιέφερε ήταν να μπορούμε να έχουμε πρόσβαση στο εκπαιδευμένο μοντέλο για να κάνουμε προβλέψεις στις αρχικές εικόνες και τα ανταγωνιστικά παραδείγματα.

3.3 Αλγόριθμοι Ανταγωνιστικών Επιθέσεων

Αφού υλοποιήσαμε τα CNN μοντέλα για την κατηγοριοποίηση των εικόνων, με βάση όσα αναφέρουμε παραπάνω έπρεπε να αποφασίσουμε ποιες ανταγωνιστικές επιθέσεις θα υλοποιήσουμε και γιατί. Στην ενότητα 2.2.3. αναφέραμε κάποια θεωρητικά στοιχεία για τις ανταγωνιστικές επιθέσεις, σε συνέχεια αυτών επιλέξαμε τις επιθέσεις FGSM, BIM και CW την καθεμία για τους παρακάτω λόγους. Αρχικά η FGSM επιλέχθηκε γιατί αποτελεί ίσως την πρώτη μέθοδο δημιουργίας ανταγωνιστικών παραδειγμάτων για βαθιά νευρωνικά δίκτυα που προτάθηκε στην βιβλιογραφία, είναι πολύ απλή και εύκολη στην υλοποίηση , καθώς επίσης στις περισσότερες εργασίες που βρήκαμε στην βιβλιογραφία επιλέγεται για την διεξαγωγή πειραμάτων οπότε θεωρήσαμε ότι μπορεί να αποτελέσει και ένα μέτρο σύγκρισης. Η BIM επιλέχθηκε γιατί αποτελεί επίσης μια αρκετά απλή και εύκολη μέθοδο στην υλοποίηση, η οποία βασίζεται στην FGSM, επειδή όμως λειτουργεί επαναληπτικά, εφαρμόζοντας την FGSM πολλές φορές καταφέρνει πολύ καλύτερα αποτελέσματα. Τέλος η CW είναι μια αρκετά πιο πολύπλοκη μέθοδος που όμως είναι πιο ισχυρή σε σχέση με τις προηγούμενες και αποτελεί μια από τις κορυφαίες μεθόδους για την δημιουργία ανταγωνιστικών παραδειγμάτων για αυτό τον λόγο και επιλέχθηκε. Στην συνέχεια θα περιγράψουμε κάποια πιο συγκεκριμένα στοιχεία σχετικά με την υλοποίηση κάθε μιας από τις επιλεγμένες μεθόδους επίθεσης και θα εξηγήσουμε κάποιες σχεδιαστικές επιλογές που πήραμε για τις υπερπαραμέτρους και την υλοποίηση τους.

3.3.1 FGSM

Όπως περιγράφει και το όνομα της η συγκεκριμένη μέθοδος για να παράξει τα ανταγωνιστικά παραδείγματα χρησιμοποιεί το gradient της loss function. Για να μπορέσουμε να υπολογίσουμε το gradient χρησιμοποιήσαμε ένα πακέτο της Pytorch που ονομάζεται autograd και υπολογίζει αυτόματα το gradient. Όταν χρησιμοποιούμε το autograd, το πέρασμα προς τα εμπρός (forward pass) του δικτύου ορίζει ένα υπολογιστικό γράφημα. Οι κόμβοι στο γράφημα είναι τανυστές (Tensors) και οι ακμές είναι συναρτήσεις που παράγουν τανυστές (Tensors) εξόδου από τανυστές (Tensors) εισόδου. Στη συνέχεια, η αντίστροφη διάδοση (back propagation) μέσω αυτού του γραφήματος επι-



Σχήμα 3.2: FGSM with epsilon = 0.02

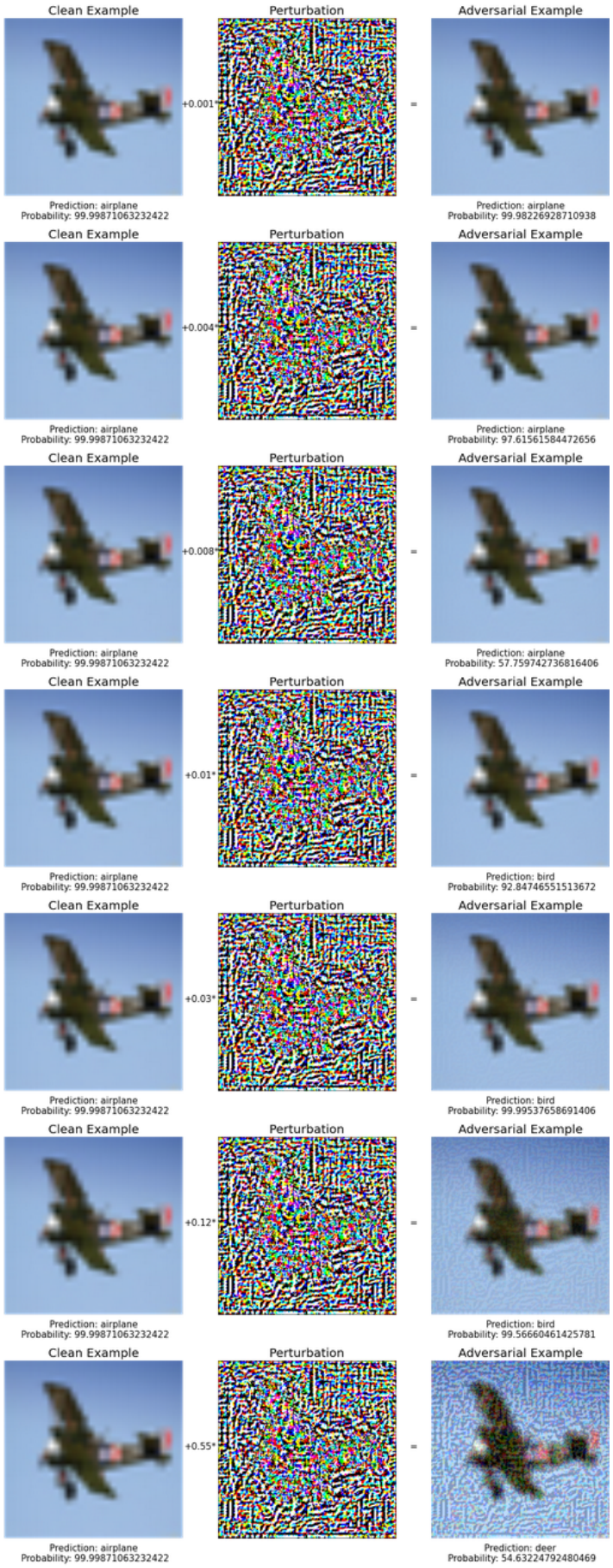
τρέπει τον εύκολο υπολογισμό των gradient. Όσο περίπλοκο και να ακούγεται, είναι αρκετά απλό στην πράξη. Στην ουσία τυλίγουμε τους Pytorch Tensor σε Variable objects. Μια Variable αντιπροσωπεύει έναν κόμβο σε ένα υπολογιστικό γράφημα. Εάν το x είναι Variable, τότε το x.data είναι Tensor και το x.grad είναι μια άλλη Variable που κρατά το gradient του x σε σχέση με κάποια κλιμακωτή τιμή. Στην συνέχεια αφού με το Variable και την αντίστροφη διάδοση (back propagation) υπολογίσαμε τα gradient, υπολογίζουμε την συνάρτηση προσήμου για το gradient και τελικά δημιουργούμε το ανταγωνιστικό παράδειγμα χρησιμοποιώντας τον παρακάτω τύπο. Όπου x είναι η αρχική εικόνα, ϵ είναι ένας πολύ μικρός αριθμός και

$$sign(\Delta_x J(\theta, x, y))$$

το gradient του loss function

$$adv_x = x + \epsilon \cdot sign(\Delta_x J(\theta, x, y))$$

Η συγκεκριμένη επίθεση καθορίζεται σε μεγάλο βαθμό από την παράμετρο epsilon, η συγκεκριμένη παράμετρος είναι ένας πολύ μικρός αριθμός που επί της ουσίας καθορίζει το ποσοστό του θορύβου που θα προσθέσουμε στην αρχική εικόνα για να δημιουργήσουμε το ανταγωνιστικό παράδειγμα. Όπως αναφέραμε και στο κεφάλαιο 2.2.3. όσο μεγαλύτερο είναι το epsilon τόσο πιο πιθανό είναι η επίθεση να είναι επιτυχημένη, όμως τόσο πιο εμφανής θα είναι και με γυμνό μάτι η επίθεση. Αυτό είναι ένα πάρα πολύ σημαντικό μειονέκτημα καθώς σκοπός της επίθεσης είναι από την μια να κατηγοριοποιηθεί με λάθος τρόπο το ανταγωνιστικό παράδειγμα αλλά από την άλλη αυτό να μην μπορεί να γίνει αντιληπτό από τον άνθρωπο. Επομένως χρειάζεται μια ισορροπία για την τιμή του epsilon ούτως ώστε να πετύχουμε τα βέλτιστα αποτελέσματα. Εμείς δοκιμάσαμε αρκετές τιμές του epsilon και θα παρουσιάσουμε στο κεφάλαιο 4 πιο αναλυτικά τα αποτελέσματα που προέκυψαν, παρόλα αυτά στο σχήμα 3.3 μπορεί να δει κανείς το πώς το epsilon επηρεάζει την επίθεση.



Σχήμα 3.3: FGSM with diferent values of epsilon

3.3.2 BIM

Η συγκεκριμένη επίθεση αποτελεί προέκταση της FGSM, αυτό που αλλάζει είναι ότι αντί να εφαρμόσουμε την FGSM σε ένα βήμα, την εφαρμόζουμε επαναληπτικά σε πολλά βήματα. Σε κάθε βήμα με τον ίδιο τρόπο που περιγράψαμε παραπάνω υπολογίζουμε το gradient και την συνάρτηση προσήμου του, η οποία είναι το petrubation που θα προσθέσουμε στην αρχική εικόνα. Στο τέλος αυτής της επαναληπτικής διαδικασίας προσθέτουμε στην αρχική εικόνα το συνολικό pertubation που υπολογίσαμε πολλαπλασιασμένο με μια παράμετρο alpha. Η παράμετρος αυτή, αντίστοιχα με την παράμετρο epsilon στην FGSM, καθορίζει το ποσοστό του θορύβου που θα προσθέσουμε στην αρχική εικόνα για να δημιουργήσουμε το ανταγωνιστικό παράδειγμα.

3.3.3 CW

Για την υλοποίηση της επίθεσης CW υλοποιήσαμε μια συνάρτηση η οποία παίρνει σαν είσοδο ένα Tensor για την εικόνα και ένα για το label της εικόνας καθώς και το μοντέλο. Στην συνέχεια υλοποιεί μια άλλη συνάρτηση f ως εξείς

$$f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -\kappa)$$

Όπου το k παρουσιάζει την εμπιστοσύνη της ίστης ετικέτας και το αρχικοποιήσαμε με 0.

Στην συνέχει ορίσαμε δύο loss funtctions ως εξής

$$loss1 = ||\frac{1}{2}\,(tanh(w) + 1) - x||_2^2$$

$$loss2 = c \cdot f(\frac{1}{2}\,(tanh(w) + 1))$$

Και επαναληπτικά προσπαθούμε να ελαχιστοποιήσουμε την Cost function

$$||\frac{1}{2}\,(tanh(w) + 1) - x||_2^2 + c \cdot f(\frac{1}{2}\,(tanh(w) + 1))$$

3.4 Αλγόριθμος Οπτικής Επεξήγησης - Visual Explanation

Όπως αναφέραμε και στην ενότητα 2.3. υπάρχουν αρκετές τεχνικές και μέθοδοι εξηγήσιμης τεχνητής νοημοσύνης. Κάθε μια από αυτές λειτουργεί με διαφορετικό τρόπο και μας δίνει διαφορετικά αποτελέσματα όλες όμως στοχεύουν στο να μας βοηθήσουμε να κατανοήσουμε και να εξηγήσουμε γιατί ένα μοντέλο βαθιάς μάθησης κάνει συγκεκριμένες προβλέψεις. Εμείς θέλοντας να συσχετίσουμε την εξηγήσιμη τεχνητή νοημοσύνη με τις ανταγωνιστικές επιθέσεις έπρεπε να επιλέξουμε σε ποια μέθοδο θα ήταν καλύτερο να εστιάσουμε. Ανατρέχοντας στην ήδη υπάρχουσα βιβλιογραφία σχετικά με την σύνδεση του Exblainable AI με τις Ανταγωνιστικές επιθέσεις παρατηρήσαμε ότι ενώ υπήρχαν έρευνες σχετικά με επιθέσεις στα explanation, και προσπάθειες να χρησιμοποιηθούν τα explanation για να κατανοήσουμε πως και σε ποιο σημείο επηρεάζεται το μοντέλο από τις ανταγωνιστικές επιθέσεις, πολύ λίγη ερευνητική δουλεία είχε γίνει γύρω από την αξιοποίηση πιο οπτικών επεξηγήσεων, όπως είναι τα heatmaps, για την διερεύνηση του πώς επηρεάζεται το μοντέλο από την επίθεση. Επομένως αποφασίσαμε να χρησιμοποιήσουμε έναν αλγόριθμο οπτι-

κής επεξήγησης για να συγκρίνουμε τις διαφορές μεταξύ των explanation στις αρχικές εικόνες και τα ανταγωνιστικά παραδείγματα.

3.4.1 GradCam

Έχοντας σαν δεδομένα τα παραπάνω επιλέξαμε να χρησιμοποιήσουμε σαν αλγόριθμο οπτικής επεξήγησης το GradCam. Το GradCam είναι ένας αλγόριθμος οπτικής επεξήγησης που χρησιμοποιεί τις διαβαθμίσεις (gradient) οποιασδήποτε κλάσης-στόχου (για παράδειγμα της κλάσης «σκύλος» σε ένα δίκτυο ταξινόμησης) που ρέουν στο τελικό συνελικτικό επίπεδο (layer) για να δημιουργήσει έναν χάρτη εντοπισμού που επισημαίνει τις σημαντικές περιοχές στην εικόνα για την πρόβλεψη της κλάσης. Στην ουσία δημιουργείται ένας χάρτης που επισημαίνει τις περιοχές της εικόνας οι οποίες έπαιξαν τον μεγαλύτερο ρόλο για την πρόβλεψη της συγκεκριμένης κλάσης. Με αυτόν τον τρόπο μπορούμε να εντοπίσουμε αν το μοντέλο που έχουμε κατασκευάσει ”κοιτάει” στα σημεία της εικόνας που θα κοιτάγαμε και εμείς για να κάνουμε την κατηγοριοποίηση.

Η τεχνική αυτή είναι μια βελτίωση σε σχέση με προηγούμενες προσεγγίσεις ως προς την ευελιξία και την ακρίβεια. Είναι πολύπλοκη, αλλά, η έξοδος που δίνει είναι διαισθητική. Περιληπτικά ο τρόπος λειτουργίας της είναι ο εξής, παίρνουμε μια εικόνα ως είσοδο και δημιουργούμε ένα μοντέλο που είναι αποκομμένο στο επίπεδο (layer) για το οποίο θέλουμε να δημιουργήσουμε έναν χάρτη θερμότητας Grad-CAM. Συνδέουμε τα πλήρως συνδεδεμένα επίπεδα (layers) για πρόβλεψη. Στη συνέχεια εκτελούμε την είσοδο μέσω του μοντέλου, παίρνουμε την έξοδο του επιπέδου και το loss. Στη συνέχεια, βρίσκουμε το gradient της εξόδου του επιθυμητού επιπέδου του μοντέλου με σεβασμό (with respect to) στο loss του μοντέλου. Από εκεί, παίρνουμε τα τμήματα του gradient που συμβάλλουν στην πρόβλεψη και κάνουμε τις απαραίτητες προσαρμογές σε μέγεθος, έτσι ώστε ο χάρτης θερμότητας να μπορεί να επικαλύπτεται με την αρχική εικόνα.

Σχήμα 3.4: GradCam

3.5 Περιβάλλον Διεξαγωγής Πειραμάτων

Για την ανάπτυξη των προγραμμάτων χρησιμοποιήσαμε το Pytorch, πρόκειται για ένα framework ανοιχτού κώδικα για μηχανική μάθηση το οποίο βασίζεται στην βιβλιοθήκη Torch. Το Pytorch χρησιμοποιείται σε διάφορες εφαρμογές βαθιάς μάθησης, όπως η όραση υπολογιστών και η επεξεργασία φυσικής γλώσσας. Είναι δωρεάν λογισμικό ανοιχτού κώδικα που κυκλοφορεί με την άδεια Modified BSD που αναπτύχθηκε κυρίως από την Meta AI. Το Pytorch παρέχει δυο δυνατότητες υψηλού επιπέδου, υπολογισμός τανυστή (όπως το NumPy) με ισχυρή επιτάχυνση μέσω μονάδων επεξεργασίας γραφικών (GPU) και βαθιά νευρωνικά δίκτυα χτισμένα σε ένα σύστημα αυτόματης διαφοροποίησης.

Τα υπολογιστικά περιβάλλοντα που χρησιμοποιήσαμε ήταν το Google Colab [1] και το Kaggle [2]. Πρόκειται για δύο πολύ εύχρηστα περιβάλλοντα που δίνουν την δυνατότητα στο χρήστη να τρέχει jupyter notebooks online. Οι δύο βασικοί λόγοι που επιλέχθηκαν τα συγκεκριμένα περιβάλλοντα ήταν α) Ότι παρέχουν δωρεάν την δυνατότητα επιτάχυνσης των εκπαιδεύσεων με την χρήση μονάδων επεξεργασίας γραφικών (GPU). Από την μια το colab παρέχει Nvidia T4 16GB 1.59GHz 8.1TFLOPS από την άλλη το Kaggle παρέχει Nvidia P100 16GB 1.32GHz 9.3TFLOPS. β) Παρέχουν τη δυνατότητα σύνδεσης με το Google Drive διευκολύνοντας την εισαγωγή δεδομένων στο πρόγραμμα. Ο λόγος για τον οποίο χρησιμοποιήθηκαν και τα δύο υπολογιστικά περιβάλλοντα ήταν διότι οι δωρεάν δυνατότητες επιτάχυνσης που παρέχουν έχουν περιορισμού στην χρήση. Συγκεκριμένα στο Kaggle η χρήση της GPU μπορεί να γίνει για έως και 35 ώρες εβδομαδιαίως ενώ στο Google Colab η παραχώρηση της GPU γίνεται με πιο δυναμικό τρόπο ανάλογα με την χρήση που όμως μπορεί να οδηγήσει μετά από περιόδους συνεχόμενης χρήσης για αρκετές ώρες/μέρες μετά για ένα μεγάλο διάστημα να μην δίνεται στον χρήστη πρόσβαση στην GPU. Αυτή την δυσκολία αντιμετωπίσαμε κατά την εκτέλεση των πειραμάτων και ενώ ξεκινήσαμε την ανάπτυξη των προγραμμάτων στο colab επειδή αρκετές φορές ξεπεράσαμε το όριο χρήσης αναγκαστήκαμε να δουλέψουμε και στο kaggle.

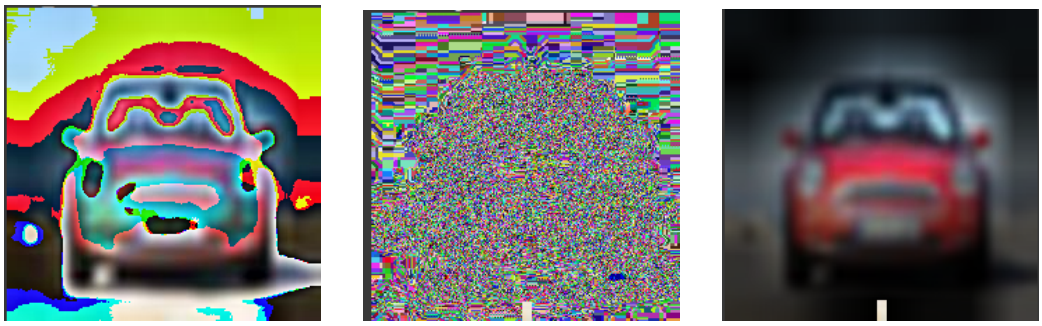
[1] https://colab.research.google.com/
[2] https://www.kaggle.com/

Κεφάλαιο 4

Αξιολόγηση και σύγκριση αποτελεσμάτων

4.1 Πρώιμα αποτελέσματα

Η βασική ιδέα με την οποία πειραματιστήκαμε ήταν αντί να χρησιμοποιούμε τα ανταγωνιστικά παραδείγματα που δημιουργούμε με κάποια από τις μεθόδους επίθεσης, να κρατήσουμε ένα κομμάτι τους με σκοπό να δημιουργήσουμε ανταγωνιστικά παραδείγματα που θα είναι πιο κοντά στην αρχική εικόνα αλλά θα καταφέρνουν και πάλι να ρίχνουν το Accuracy του μοντέλου σε έναν ικανοποιητικό βαθμό. Για να το πετύχουμε αυτό σκεφτήκαμε να χρησιμοποιήσουμε κάποιον αλγόριθμο εξηγήσιμης τεχνητής νοημοσύνης για να εντοπίσουμε τις περιοχές τις εικόνας στις οποίες το μοντέλο δίνει παραπάνω έμφαση για να κάνει την τελική πρόβλεψη. Σκεφτήκαμε λοιπόν να αλλάξουμε με κάποιο τρόπο μόνο αυτές τις περιοχές, ούτως ώστε να δημιουργήσουμε ένα ανταγωνιστικό παράδειγμα το οποίο θα είχε πολύ πιο μικρή απόσταση από την αρχική εικόνα αλλά θα μπορούσε και πάλι σε ικανοποιητικό βαθμό να αλλάξει την τελική πρόβλεψη του μοντέλου. Αρχικά σκεφτήκαμε για κάθε pixel του saliency map να υπολογίζουμε το ποσοστό σημαντικότητας που έχει, καθώς οι τιμές του saliency map κινούνται στο διάστημα [0,254] όπου όσο μεγαλώνει η τιμή σημαίνει ότι τόσο περισσότερο ρόλο στην τελική απόφαση παίζει το συγκεκριμένο pixel. Έτσι διαιρώντας τη τιμή του κάθε pixel με το 254 βρίσκουμε έναν αριθμό P που κυμαίνεται στο διάστημα [0,1] και εκφράζει την σημαντικότητα του συγκεκριμένου pixel. Στην συνέχεια σκεφτήκαμε να δημιουργούμε ένα καινούργιο ανταγωνιστικό παράδειγμα κάθε pixel του οποίου θα προέκυπτε από το αντίστοιχο pixel του αρχικού ανταγωνιστικού παραδείγματος πολλαπλασιασμένο με τον αριθμό P συν το αντίστοιχο pixel της αρχικής εικόνας πολλαπλασιασμένο με το (1-P). Όπως παρατηρήσαμε όμως με τον συγκεκριμένο τύπο όπως και με διάφορες παραλλαγές του τα αποτελέσματα των οποίων φαίνονται στο σχήμα 4.1 παρακάτω, υπήρχε πολύ μεγάλη παραμόρφωση στην εικόνα. Επομένως αναγκαστήκαμε να εγκαταλείψουμε την συγκεκριμένη προσέγγιση και να αναζητήσουμε κάποιον άλλο τρόπο για να δημιουργήσουμε ανταγωνιστικά παραδείγματα όπως αυτά που αναφέραμε παραπάνω. Παρά την άκαρπη, όπως αποδείχθηκε, προσπάθεια να δημιουργήσουμε ανταγωνιστικά παραδείγματα με αυτή την μέθοδο, αναφέρουμε τα παραπάνω κυρίως γιατί μέσω αυτής της αποτυχίας μπορέσαμε να κατευθύνουμε την αναζήτηση μας προς μια πιο σωστή μέθοδο, που μας έδωσε αισθητά καλύτερα αποτελέσματα όπως θα δούμε και στις επόμενες ενότητες. Οι εικόνες που ακολουθούν είναι κάποια δείγματα των μεθόδων που εξηγήσαμε σε αυτή την υποενότητα.



Σχήμα 4.1: Failed cutom adversarial images

4.2 Αποτελέσματα επιθέσεων με ποσοστό από την αρχική εικόνα

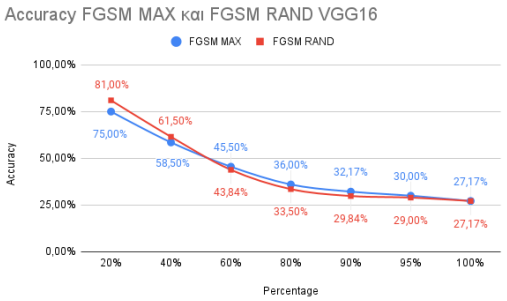
Όπως αναφέραμε και παραπάνω η βασική ιδέα με την οποία πειραματιστήκαμε ήταν αντί να χρησιμοποιούμε τα ανταγωνιστικά παραδείγματα που δημιουργούμε με κάποια από τις μεθόδους επίθεσης, να κρατήσουμε ένα κομμάτι τους με σκοπό να δημιουργήσουμε ανταγωνιστικά παραδείγματα που θα είναι πιο κοντά στην αρχική εικόνα αλλά θα καταφέρνουν και πάλι να ρίχνουν το Accuracy του μοντέλου σε έναν ικανοποιητικό βαθμό. Μετά από δοκιμές που είδαμε και στην προηγούμενη ενότητα, καταλήξαμε να συνθέτουμε ένα καινούργιο ανταγωνιστικό παράδειγμα ξεκινώντας από την αρχική εικόνα και αντικαθιστώντας ένα συγκεκριμένο ποσοστό pixel με τα αντίστοιχα pixel του ανταγωνιστικού παραδείγματος που παίρναμε από την επίθεση. Κάναμε δυο φορές το πείραμα, μια αλλάζοντας το συγκεκριμένο ποσοστό pixel με τυχαίο τρόπο, ενώ την άλλη επιλέγοντας το συγκεκριμένο ποσοστό των pixel που είχαν την μεγαλύτερη αντίστοιχη τιμή στο saliency map. Ο λόγος που κινηθήκαμε με αυτό τον τρόπο ήταν διότι θέλαμε να διερευνήσουμε αν αλλάζοντας τα pixel που παίζουν παραπάνω ρόλο στην λήψη της απόφασης, και επομένως έχουν μεγαλύτερη τιμή στο saliency map, θα πετύχουμε αισθητά καλύτερα αποτελέσματα από ότι αλλάζοντας το ίδιο ποσοστό pixel αλλά με τυχαίο τρόπο. Θέλαμε να εξετάσουμε αν αλλάζοντας έναν μικρότερο αριθμό pixel από ότι η αρχική επίθεση, που όμως γνωρίζουμε ότι είναι pixel που παίζουν κυρίαρχο ρόλο στην τελική απόφαση του μοντέλου θα καταφέρναμε να ρίξουμε το Accuracy του μοντέλου σε ικανοποιητικό βαθμό. Προσπαθήσαμε στην ουσία να χρησιμοποιήσουμε την γνώση που αποκτήσαμε για το μοντέλο από τον αλγόριθμο εξηγήσιμης τεχνητής νοημοσύνης για να βελτιώσουμε τις υπάρχουσες επιθέσεις, προσπαθώντας να αλλάξουμε μόνο τις περιοχές που φαίνεται από το saliency map ότι είναι σημαντικές στην λήψη της τελικής απόφασης.

Στην ενότητα αυτή θα παρουσιάσουμε και θα σχολιάσουμε τα αποτελέσματα των πειραμάτων που περιγράψαμε παραπάνω για κάθε μια από τις επιθέσεις που υλοποιήσαμε. Κάθε επίθεση έχει δοκιμαστεί για δύο αρχιτεκτονικές για λόγους πληρότητας, ούτως ώστε να μπορούμε σε ένα βαθμό να γενικεύσουμε να συμπεράσματα μας. Επίσης αξίζει να αναφερθεί ότι για όλα μας τα παραδείγματα χρησιμοποιήσαμε δυο μετρικές, με αυτές αξιολογούμε τα αποτελέσματα μας και οι τιμές αυτών των μετρικών είναι που παρουσιάζονται στα διαγράμματα στις παρακάτω υποενότητες. Οι μετρικές αυτές είναι το Accuracy του Classifier δηλαδή το ποσοστό των σωστών προβλέψεων που πέτυχε το μοντέλο. Η δεύτερη μετρική είναι το NISSIM (Normalized Inverted Structural Similarity Index) υπολογίζεται από το Structural Similarity Index (SSIM) αντιστρέφοντας το εύρος και στη συνέχεια κανονικοποιώντας το. Το SSIM είναι μια μετρική που εστιάζει στην ομοιότητα, άρα αντιστρέφοντας το μας δίνει ανομοιότητα. Η τιμή του SSIM περιορίζεται στο διάστημα (-1, 1],

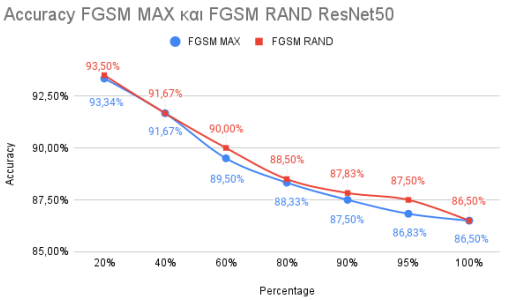
όπου -1 σημαίνει ανόμοιο ενώ 1 σημαίνει παρόμοιο και η τιμή του NISSIM περιορίζεται στο διάστημα (0, 1] όπου το 0 σημαίνει όμοιο και το 1 σημαίνει ανόμοιο. Στην ιδανική περίπτωση θέλουμε αυτή την τιμή όσο το δυνατόν πιο κοντά στο 0. Η μετρική αυτή μας δείχνει την ανομοιότητα της αρχικής εικόνας με το ανταγωνιστικό παράδειγμα που δημιουργήσαμε.

4.2.1 FGSM

Σε αυτή την υποενότητα θα παρουσιάσουμε τα αποτελέσματα για την επίθεση FGSM. Πρόκειται για 4 διαγράμματα εκ των οποίων στα 2 πρώτα φαίνονται οι τιμές της πρώτης μετρικής που αναφέραμε παραπάνω, δηλαδή του Accuracy. Ενώ στα άλλα 2 φαίνονται οι τιμές της δεύτερης μετρικής, δηλαδή του NISSIM. Σε κάθε διάγραμμα παρουσιάζονται σε διαφορετική γραμμή τα αποτελέσματα για την μέθοδο αλλαγής pixel με τυχαίο τρόπο και με βάση την μέγιστη αντίστοιχη τιμή του saliency map. Τα αποτελέσματα για την μέθοδο αλλαγής pixel με τυχαίο τρόπο συμβολίζονται με την κόκκινη γραμμή, ενώ με βάση την μέγιστη αντίστοιχη τιμή του saliency map με μπλε γραμμή. Ακόμα για κάθε μετρική παρουσιάζονται 2 διαγράμματα καθώς το καθένα αντιστοιχεί στις δυο διαφορετικές αρχιτεκτονικές που υλοποιήσαμε, δηλαδή την VGG16 και την ResNet50. Τέλος για να μελετήσουμε ολοκληρωμένα την υπόθεση μας σε κάθε διάγραμμα παρουσιάζουμε τα αποτελέσματα για 7 διαφορετικά ποσοστά όπως φαίνονται στον οριζόντιο άξονα. Όπως εξηγήσαμε και παραπάνω, για κάθε τιμή, έχουμε αντικαταστήσει στην αρχική εικόνα αυτό το ποσοστό των pixel με τα αντίστοιχα pixel του ανταγωνιστικού παραδείγματος. Επομένως το 0 αντιστοιχεί στην αρχική εικόνα και το 100 στο ανταγωνιστικό παράδειγμα.



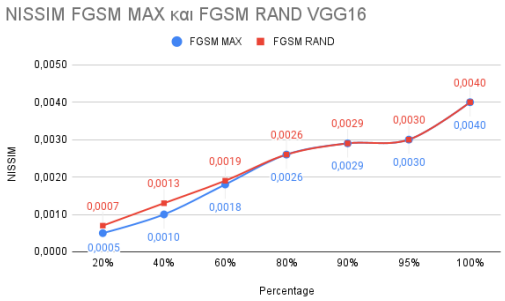
Σχήμα 4.2: Accuracy FGSM VGG16



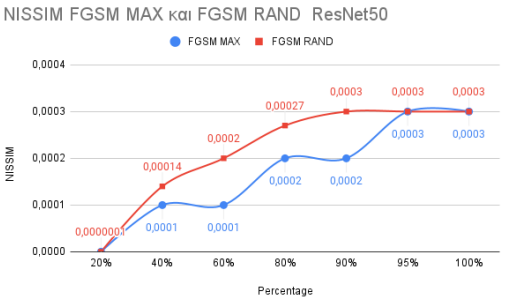
Σχήμα 4.3: Accuracy FGSM ResNet50

Όπως βλέπουμε και στα δύο διαγράμματα 4.2, 4.3 το Accuracy πέφτει καθώς αυξάνεται η τιμή του ποσοστού των pixel που αλλάζουμε. Ειδικά για την αρχιτεκτονική VGG16 ενώ το Accuracy ξεκινάει από 93% ήδη με πολύ μικρό ποσοστό αλλαγής pixel πέφτει στο 81%. Ακόμη παρατηρούμε ότι και στα δύο γραφήματα οι τιμές για αλλαγή των pixel με τυχαίο τρόπο και για αλλαγή με βάση τις μέγιστες αντίστοιχες τιμές του saliency map είναι πολύ κοντά, δεν παρατηρείται δηλαδή μια σταθερή και αξιοσημείωτη διαφορά μεταξύ των δύο μεθόδων. Τέλος παρατηρούμε ότι η αρχιτεκτονική ResNet50 παρουσιάζει αρκετά μεγαλύτερη ανθεκτικότητα σε σχέση με την VGG16 στην συγκεκριμένη επίθεση, καθώς το Accuracy παραμένει σε πολύ υψηλές τιμές και δεν πέφτει κάτω από 86,5% σε αντίθεση με το 27,17% του VGG16

Βλέποντας από τα διαγράμματα 4.4, 4.5 τις τιμές του NISSIM για τα διάφορα ποσοστά αλλαγής pixel αρχικά παρατηρούμε ότι οι τιμές ξεκινάνε από πολύ κοντά στο μηδέν και σταδιακά ανεβαίνουν. Αυτό είναι λογικό καθώς ξεκινάμε αλλάζοντας μόνο το 20 των pixel της αρχικής ει-



Σχήμα 4.4: NISSIM FGSM VGG16



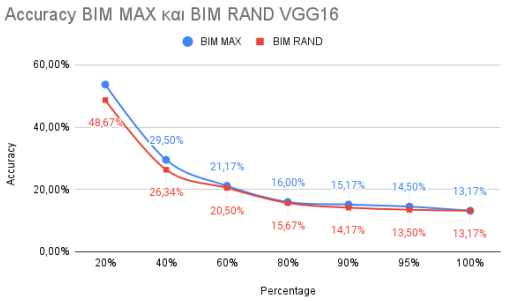
Σχήμα 4.5: NISSIM FGSM ResNet50

κόνας επομένως είναι λογικό η απόσταση των ανταγωνιστικών παραδειγμάτων από την αρχική εικόνα να είναι πολύ μικρή και το NISSIM πολύ κοντά στο 0. Ακόμα παρατηρούμε ότι για την VGG16 η απόσταση μεταξύ των τιμών για αλλαγή των pixel με τυχαίο τρόπο και για αλλαγή με βάση τις μέγιστες αντίστοιχες τιμές του saliency map είναι πολύ κοντά. Όσον αφορά την ResNet50 ενώ φαίνεται να υπάρχει διαφορά μεταξύ των δύο γραμμών, παρόλα αυτά αν παρατηρήσουμε η διαφορά μεταξύ των τιμών είναι της τάξη του 0,0001 δηλαδή των τεσσάρων δεκαδικών ψηφίων, που είναι τόσο μικρή που μπορεί να οφείλεται στην τυχαιότητα.

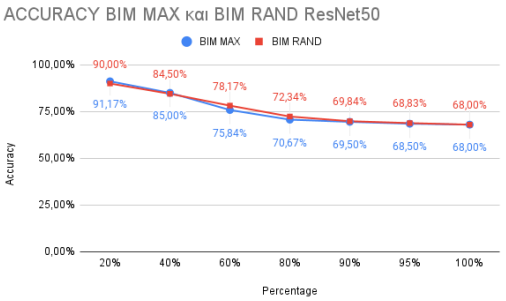
4.2.2 BIM

Σε αυτή την υποενότητα θα παρουσιάσουμε τα αποτελέσματα για την επίθεση BIM. Πρόκειται για 4 διαγράμματα εκ των οποίων στα 2 πρώτα φαίνονται οι τιμές της πρώτης μετρικής που αναφέραμε παραπάνω, δηλαδή του Accuracy. Ενώ στα άλλα 2 φαίνονται οι τιμές της δεύτερης μετρικής, δηλαδή του NISSIM. Σε κάθε διάγραμμα παρουσιάζονται σε διαφορετική γραμμή τα αποτελέσματα για την μέθοδο αλλαγής pixel με τυχαίο τρόπο και με βάση την μέγιστη αντίστοιχη τιμή του saliency map. Τα αποτελέσματα για την μέθοδο αλλαγής pixel με τυχαίο τρόπο συμβολίζονται με την κόκκινη γραμμή, ενώ με βάση την μέγιστη αντίστοιχη τιμή του saliency map με μπλε γραμμή. Ακόμα για κάθε μετρική παρουσιάζονται 2 διαγράμματα καθώς το καθένα αντιστοιχεί στις δυο διαφορετικές αρχιτεκτονικές που υλοποιήσαμε, δηλαδή την VGG16 και την ResNet50. Τέλος για να μελετήσουμε ολοκληρωμένα την υπόθεση μας σε κάθε διάγραμμα παρουσιάζουμε τα αποτελέσματα για 7 διαφορετικά ποσοστά όπως φαίνονται στον οριζόντιο άξονα. Όπως εξηγήσαμε και παραπάνω, για κάθε τιμή, έχουμε αντικαταστήσει στην αρχική εικόνα αυτό το ποσοστό των pixel με τα αντίστοιχα pixel του ανταγωνιστικού παραδείγματος. Επομένως το 0 αντιστοιχεί στην αρχική εικόνα και το 100 στο ανταγωνιστικό παράδειγμα.

Όπως βλέπουμε και στα δύο διαγράμματα 4.6, 4.7 το Accuracy πέφτει καθώς αυξάνεται η τιμή του ποσοστού των pixel που αλλάζουμε. Ειδικά για την αρχιτεκτονική VGG16 ένω το Accuracy ξεκινάει από 93% ήδη με μόλις 20% ποσοστό αλλαγής pixel πέφτει κάτω από το 50% και με ποσοστό αλλαγής pixel 40% πέφτει κάτω από το 30%. Ακόμη παρατηρούμε ότι και στα δύο γραφήματα οι τιμές για αλλαγή των pixel με τυχαίο τρόπο και για αλλαγή με βάση τις μέγιστες αντίστοιχες τιμές του saliency map είναι πολύ κοντά, δεν παρατηρείται δηλαδή μια σταθερή και αξιοσημείωτη διαφορά μεταξύ των δύο μεθόδων. Τέλος παρατηρούμε ότι η αρχιτεκτονική ResNet50 παρουσιάζει αρκετά μεγαλύτερη ανθεκτικότητα σε σχέση με την VGG16 στην συγκεκριμένη επίθεση, καθώς το

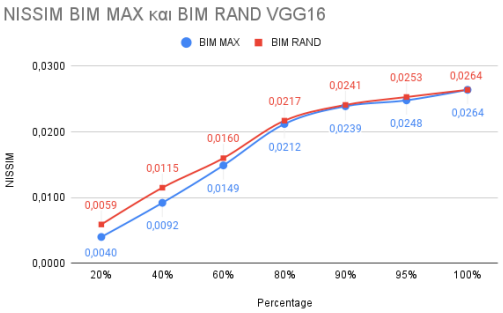


Σχήμα 4.6: Accuracy BIM VGG16

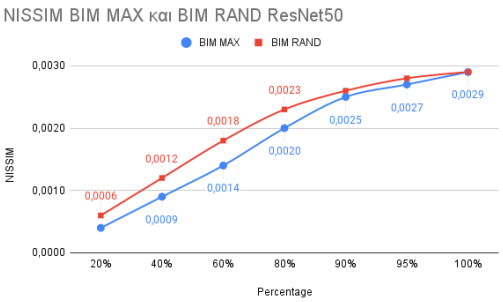


Σχήμα 4.7: Accuracy BIM ResNet50

Accuracy παραμένει σε πολύ υψηλές τιμές και δεν πέφτει κάτω από 68 σε αντίθεση με το 13,17 του VGG16



Σχήμα 4.8: NISSIM BIM VGG16



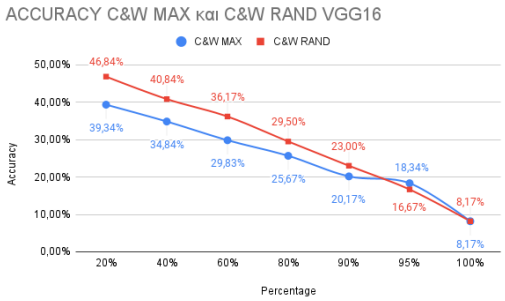
Σχήμα 4.9: NISSIM BIM ResNet50

Βλέποντας από τα διαγράμματα 4.8, 4.9 τις τιμές του NISSIM για τα διάφορα ποσοστά αλλαγής pixel αρχικά παρατηρούμε ότι οι τιμές ξεκινάνε από πολύ κοντά στο μηδέν και σταδιακά ανεβαίνουν. Αυτό είναι λογικό καθώς ξεκινάμε αλλάζοντας μόνο το 20 των pixel της αρχικής εικόνας επομένως είναι λογικό η απόσταση των ανταγωνιστικών παραδειγμάτων από την αρχική εικόνα να είναι πολύ μικρή και το NISSIM πολύ κοντά στο 0. Ακόμα παρατηρούμε ότι και για την VGG16 και για την ResNet50 η απόσταση μεταξύ των τιμών για αλλαγή των pixel με τυχαίο τρόπο και για αλλαγή με βάση τις μέγιστες αντίστοιχες τιμές του saliency map είναι πολύ κοντά. Χωρίς να παρατηρείται κάποια ιδιαίτερη διαφορά σταθερή διαφορά μεταξύ των δύο μεθόδων.

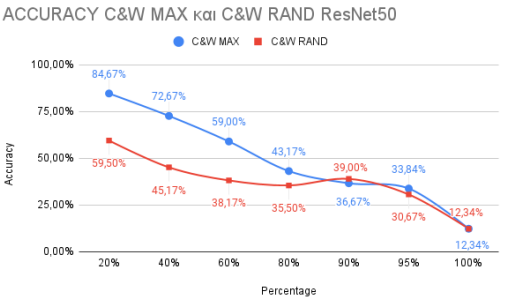
4.2.3 CW

Σε αυτή την υποενότητα θα παρουσιάσουμε τα αποτελέσματα για την επίθεση CW. Πρόκειται για 4 διαγράμματα εκ των οποίων στα 2 πρώτα φαίνονται οι τιμές της πρώτης μετρικής που αναφέραμε παραπάνω, δηλαδή του Accuracy. Ενώ στα άλλα 2 φαίνονται οι τιμές της δεύτερης μετρικής, δηλαδή του NISSIM. Σε κάθε διάγραμμα παρουσιάζονται σε διαφορετική γραμμή τα αποτελέσματα για την μέθοδο αλλαγής pixel με τυχαίο τρόπο και με βάση την μέγιστη αντίστοιχη τιμή του saliency map. Τα αποτελέσματα για την μέθοδο αλλαγής pixel με τυχαίο τρόπο συμβολίζονται με την κόκκινη γραμμή, ενώ με βάση την μέγιστη αντίστοιχη τιμή του saliency map με μπλε γραμμή. Ακόμα για κάθε μετρική παρουσιάζονται 2 διαγράμματα καθώς το καθένα αντιστοιχεί στις δυο διαφορετικές αρχιτεκτονικές που υλοποιήσαμε, δηλαδή την VGG16 και την ResNet50. Τέλος για να μελετήσουμε ολοκληρωμένα την υπόθεση μας σε κάθε διάγραμμα παρουσιάζουμε τα

αποτελέσματα για 7 διαφορετικά ποσοστά όπως φαίνονται στον οριζόντιο άξονα. Όπως εξηγήσαμε και παραπάνω, για κάθε τιμή, έχουμε αντικαταστήσει στην αρχική εικόνα αυτό το ποσοστό των pixel με τα αντίστοιχα pixel του ανταγωνιστικού παραδείγματος. Επομένως το 0 αντιστοιχεί στην αρχική εικόνα και το 100 στο ανταγωνιστικό παράδειγμα.

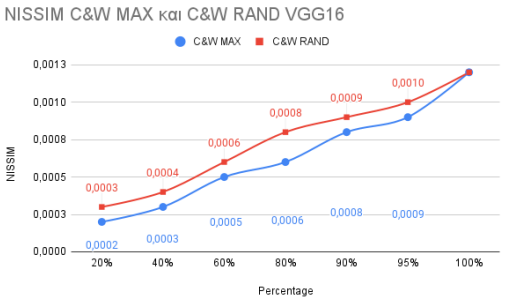


Σχήμα 4.10: Accuracy CW VGG16

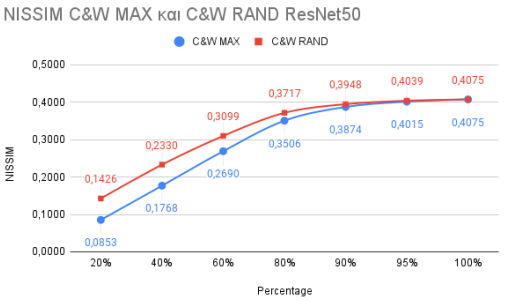


Σχήμα 4.11: Accuracy CW ResNet50

Όπως βλέπουμε στα παραπάνω διαγράμματα 4.10, 4.11 το Accuracy πέφτει καθώς αυξάνεται η τιμή του ποσοστού των pixel που αλλάζουμε. Ειδικά για την αρχιτεκτονική VGG16 ένω το Accuracy ξεκινάει από 93% ήδη με μόλις 20% ποσοστό αλλαγής pixel πέφτει κάτω από 47% και 40% για Rand και Max αντίστοιχα ενώ για την ResNet50 ενώ ξεκινάει από 93,78% πέφτει κάτω από 60% και 85% για Rand και Max αντίστοιχα. Ακόμα παρατηρούμε ότι στην συγκεκριμένη επίθεση παρατηρείται διαφορά μεταξύ των δύο μεθόδων. Αρχικά για την αρχιτεκτονική VGG16 για χαμηλές τιμές του ποσοστού αλλαγής pixel βλέπουμε ότι η μέθοδος αλλαγής pixel με βάση τις μέγιστες αντίστοιχες τιμές του saliency map δίνει χαμηλότερο Accuracy. Αντίθετα για την αρχιτεκτονική ResNet50 παρατηρούμε ότι για χαμηλές τιμές του ποσοστού αλλαγής pixel είναι η μέθοδος αλλαγής pixel με τυχαίο τρόπο που δίνει χαμηλότερο Accuracy. Αυτό που μπορούμε να συμπεράνουμε από τα συγκεκριμένα αλλά και από τα διαγράμματα για τις άλλες επιθέσεις, στα οποία δεν παρατηρήσαμε κάποια σημαντική διαφορά μεταξύ των δύο μεθόδων, είναι ότι ο αρχικός μας ισχυρισμός δεν επαληθεύεται, όχι απόλυτα τουλάχιστον. Δηλαδή δεν επιβεβαιώνεται πλήρως αυτό που αρχικά υποθέσαμε, ότι άμα αλλάξουμε τα pixel που παίζουν περισσότερο ρόλο στην λήψη της απόφασης θα πετύχουμε καλύτερα αποτελέσματα από ότι αν αλλάξουμε pixel με τυχαίο τρόπο.



Σχήμα 4.12: NISSIM CW VGG16



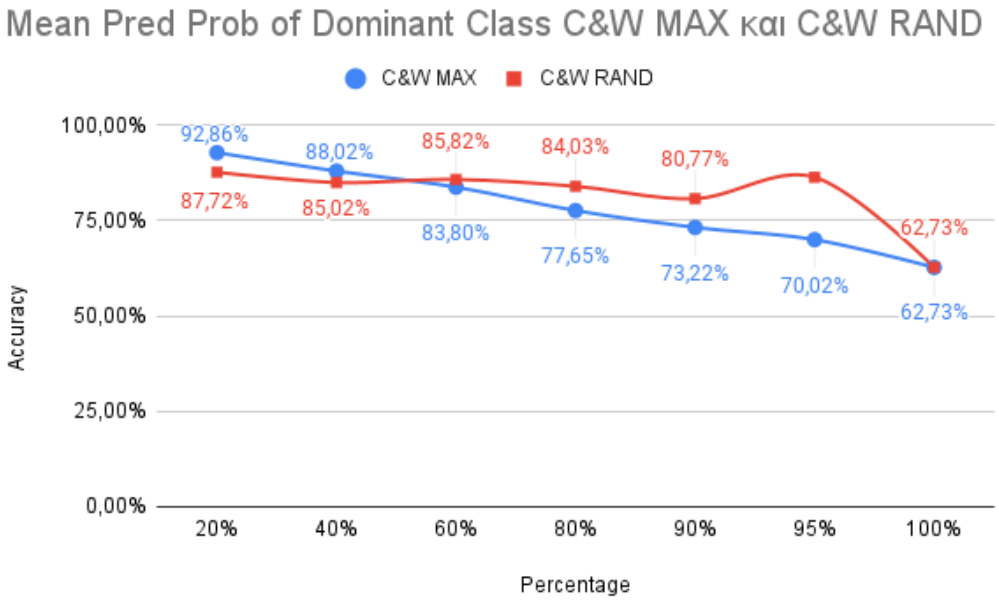
Σχήμα 4.13: NISSIM CW ResNet50

Βλέποντας από τα διαγράμματα 4.12, 4.13 τις τιμές του NISSIM για τα διάφορα ποσοστά αλλαγής pixel αρχικά παρατηρούμε ότι οι τιμές ξεκινάνε από πολύ κοντά στο μηδέν και σταδιακά ανεβαίνουν. Αυτό είναι λογικό καθώς ξεκινάμε αλλάζοντας μόνο το 20 των pixel της αρχικής εικόνας επομένως είναι λογικό η απόσταση των ανταγωνιστικών παραδειγμάτων από την αρχική εικόνα να είναι πολύ μικρή και το NISSIM πολύ κοντά στο 0. Ακόμα παρατηρούμε ότι και για την VGG16 και για την ResNet50 η απόσταση μεταξύ των τιμών για αλλαγή των pixel με τυχαίο τρόπο

και για αλλαγή με βάση τις μέγιστες αντίστοιχες τιμές του saliency map είναι πολύ κοντά. Χωρίς να παρατηρείται κάποια ιδιαίτερη σταθερή διαφορά μεταξύ των δύο μεθόδων που θα μπορούσε να χαρακτηριστεί ως σημαντική.

4.3 Πτώση ποσοστού κυρίαρχης κλάσης

Πέρα από τις δύο μετρικές που χρησιμοποιήσαμε παραπάνω και παρουσιάσαμε στα αποτελέσματα θεωρήσαμε ότι θα είχε ενδιαφέρον να μελετήσουμε το πώς επηρεάζεται το ποσοστό πρόβλεψης της κυρίαρχης κλάσης. Δηλαδή προσπαθήσαμε να παρατηρήσουμε το πως αλλάζει η σιγουριά του μοντέλου για την κλάση που τελικά επιλέγει ως κυρίαρχη. Όταν τροφοδοτούμε τον Classifier με μια εικόνα, μας επιστρέφει ένα ποσοστό για κάθε κλάση, το ποσοστό αυτό μιας κλάσης μας δείχνει πόσο σίγουρο είναι το μοντέλο ότι η εικόνα ανήκει στην συγκεκριμένη κλάση. Η κλάση που έχει το μεγαλύτερο ποσοστό είναι η κυρίαρχη κλάση και είναι η τελική πρόβλεψη του μοντέλου. Εμείς για κάθε ποσοστό αλλαγής pixel υπολογίσαμε την μέση τιμή του ποσοστού τις εκάστοτε κυρίαρχης κλάσης για κάθε εικόνα στο dataset μας. Όπως και παραπάνω τα αποτελέσματα για την μέθοδο αλλαγής pixel με τυχαίο τρόπο συμβολίζονται με την κόκκινη γραμμή, ενώ με βάση την μέγιστη αντίστοιχη τιμή του saliency map με μπλε γραμμή. Τα διαγράμματα που ακολουθούν, παρουσιάζουν τις μέσες τιμές του ποσοστού πρόβλεψης της κυρίαρχης κλάσης, για τις διάφορες τιμές του ποσοστού αλλαγής pixel, για την επίθεση CW στις αρχιτεκτονικές VGG16 και ResNet50. Ο σχολιασμός που ακολουθεί παρακάτω αποτελεί παρατηρήσεις πάνω σε κάποια πρώτα πειράματα που κάναμε με σκοπό να μελετηθεί το φαινόμενο.



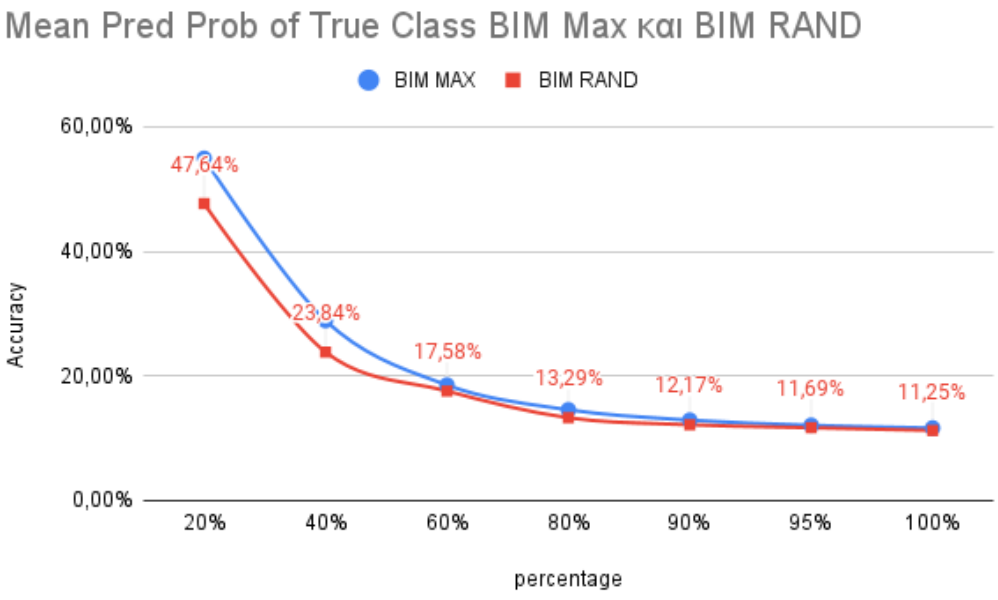
Σχήμα 4.14: Μέσο ποσοστό πρόβλεψης κυρίαρχης κλάσης

Στο διάγραμμα 4.14 βλέπουμε ότι το μέσο ποσοστό της εκάστοτε κυρίαρχης κλάσης σταδιακά πέφτει καθώς αυξάνεται το ποσοστό των pixel που αλλάζουμε. Η σιγουριά δηλαδή του μοντέλου για την κλάση που έχει επιλέξει ως κυρίαρχη πέφτει μέχρι και σχεδόν τα 2/3 της αρχικής. Αυτό σημαίνει ότι είτε η επίθεση είναι επιτυχημένη είτε όχι, είτε δηλαδή άλλαξε η κυρίαρχη κλάση είτε

όχι, το πόσο σίγουρο ήταν το μοντέλο για την τελική του απόφαση, έπεσε αρκετά. Το ενδιαφέρον με το παραπάνω αποτέλεσμα είναι ότι ενώ στην πλειοψηφία των περιπτώσεων η επίθεση είναι επιτυχημένη (βλ. Accuracy για CW ειδικά για μεγάλα ποσοστά αλλαγής pixel) δεν καταφέρνει το μοντέλο να πετύχει ίδια ποσοστά σιγουριάς για την τελική του απόφαση σε σχέση με την αρχική-πραγματική του απόφαση. Ακόμα παρατηρούμε ότι η μέθοδος με την τυχαία αλλαγή των pixel παρουσιάζει πιο σταθερά ποσοστά παρόλα αυτά στο τέλος και αυτή έχει αρκετά μικρότερη σιγουριά. Επειδή όπως αναφέρουμε και παραπάνω δεν πραγματοποιήσαμε εκτενή έρευνα πάνω στο συγκεκριμένο πείραμα, δεν μπορούμε να εξηγήσουμε πλήρως γιατί η τυχαία αλλαγή pixel παρουσιάζει μεγαλύτερη σταθερότητα, παρόλα αυτά θα είχε ενδιαφέρον να μελετηθεί παραπάνω.

4.4 Πτώση ποσοστού πραγματικής κλάσης

Αντίστοιχα με την προηγούμενη υποενότητα θεωρήσαμε ότι θα έχει ενδιαφέρον πέρα από το ποσοστό πρόβλεψης της κυρίαρχης κλάσης, να μελετήσουμε το πώς επηρεάζεται το ποσοστό πρόβλεψης της πραγματικής κλάσης. Δηλαδή το πως αλλάζει η σιγουριά του μοντέλου για την κλάση που είναι η πραγματική κλάση της εικόνας. Επομένως για κάθε ποσοστό αλλαγής pixel υπολογίσαμε την μέση τιμή του ποσοστού τις εκάστοτε πραγματικής κλάσης για κάθε εικόνα στο dataset μας. Όπως και παραπάνω τα αποτελέσματα για την μέθοδο αλλαγής pixel με τυχαίο τρόπο συμβολίζονται με την κόκκινη γραμμή, ενώ με βάση την μέγιστη αντίστοιχη τιμή του saliency map με μπλε γραμμή. Τα διαγράμματα που ακολουθούν, παρουσιάζουν τις μέσες τιμές του ποσοστού πρόβλεψης της πραγματικής κλάσης, για τις διάφορες τιμές του ποσοστού αλλαγής pixel, για την επίθεση CW στις αρχιτεκτονικές VGG16 και ResNet50. Ο σχολιασμός που ακολουθεί παρακάτω αποτελεί παρατηρήσεις πάνω σε κάποια πρώτα πειράματα που κάναμε με σκοπό να μελετηθεί το φαινόμενο.



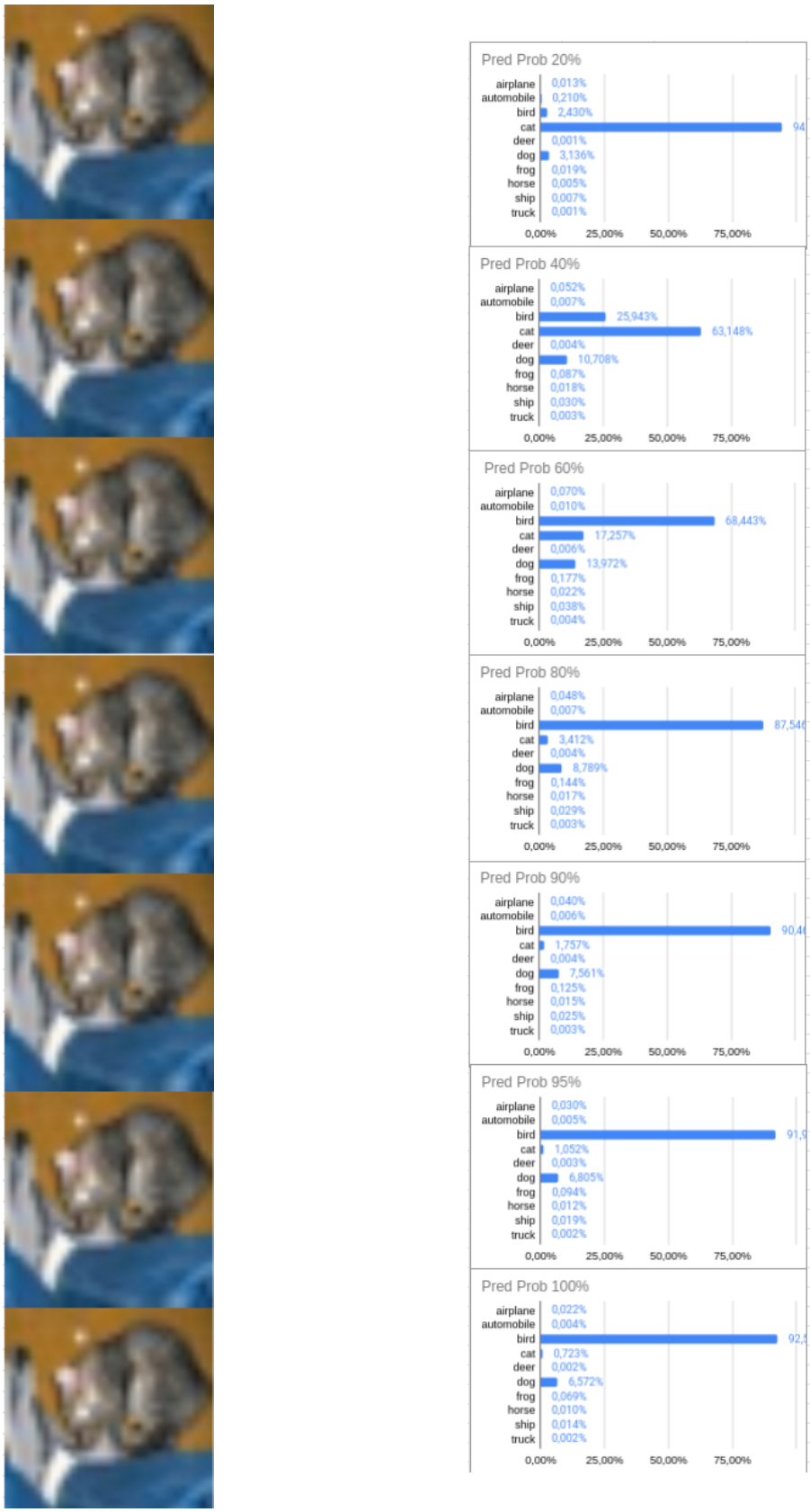
Σχήμα 4.15: Μέσο ποσοστό πρόβλεψης πραγματικής κλάσης

Στο διάγραμμα 4.15 βλέπουμε ότι το μέσο ποσοστό της πραγματικής κλάσης πέφτει δραματικά καθώς αυξάνεται το ποσοστό των pixel που αλλάζουμε. Η σιγουριά δηλαδή του μοντέλου για την πραγματική κλάση της εικόνας πέφτει μέχρι και το 11%. Αυτό σημαίνει ότι είτε η επίθεση είναι επιτυχημένη είτε όχι, είτε δηλαδή άλλαξε η κυρίαρχη κλάση είτε όχι, το πόσο σίγουρο ήταν το μοντέλο για την πραγματική κλάση, έπεσε δραματικά. Το αποτέλεσμα αυτό είναι λογικό καθώς ο βασικός στόχος της επίθεσης είναι να ρίξει το ποσοστό της πραγματικής κλάσης ούτως ώστε να μην κατηγοριοποιηθεί σωστά η εικόνα. Επειδή το ποσοστό πέφτει τόσο χαμηλά θα μπορούσαμε να υποθέσουμε ότι, ακόμα και τις φορές που η επίθεση δεν είναι επιτυχημένη και η πραγματική κλάση παραμένει κυρίαρχη, το ποσοστό της πέφτει αρκετά. Μάλιστα έχει ενδιαφέρον ότι με περίπου 50% ποσοστό αλλαγής pixel έχει πέσει το ποσοστό σχεδόν μέχρι την τελική του τιμή, που σημαίνει ότι από εκείνο το ποσοστό και μετά η επίθεση απλώς ανεβάζει κάποια άλλη κλάση.

4.5 Ποσοστό όλων των κλάσεων

Σε συνέχεια των δυο προηγούμενων υποενοτήτων, προσπαθήσαμε να μελετήσουμε το πως αλλάζουν τα ποσοστά πρόβλεψης όλων των κλάσεων και όχι μόνο της κυρίαρχης και της πραγματικής. Κάτι τέτοιο θεωρήσαμε ότι θα έχει ενδιαφέρων για να μπορέσουμε να κάνουμε κάποιες πρώτες παρατηρήσεις σχετικά με την κατεύθυνση των επιθέσεων και τον τρόπο που επιλέγουν κάθε φορά να αλλάξουν την κυρίαρχη κλάση. Στο σημείο αυτό να θυμίσουμε ότι επιλέξαμε να ασχοληθούμε με μη στοχευμένες επιθέσεις, που σημαίνει ότι δεν έχουν μια κλάση-στόχο Y που θέλουμε το μοντέλο να κατηγοριοποιήσει την εικόνα ως τέτοια, αλλά θέλουμε το μοντέλο να κατηγοριοποιήσει εσφαλμένα την εικόνα για οποιαδήποτε κλάση εκτός της πραγματικής κλάσης X 2. Επομένως δεν ξέρουμε εκ των προτέρων ποιας κλάσης το ποσοστό πρόβλεψης θα προσπαθήσει να ανεβάσει η επίθεση. Ακόμα πρέπει να έχουμε στο μυαλό μας ότι η διαδικασία αυτή χωρίζεται επί της ουσίας σε δύο κατηγορίες. Στην πρώτη που η τελική επίθεση είναι επιτυχημένη, κάποια στιγμή όσο αυξάνεται το ποσοστό αλλαγής pixel, η κυρίαρχη κλάση θα αλλάξει και από την πραγματική θα γίνει κάποια άλλη, δηλαδή θα συμβεί φλιπ στην κυρίαρχη κλάση. Ενώ στην δεύτερη που η τελική επίθεση δεν είναι επιτυχημένη, η κυρίαρχη κλάση δεν θα αλλάξει και θα παραμείνει η πραγματική ως κυρίαρχη μέχρι τέλους. Και στις δύο αυτές περιπτώσεις αξίζει να παρατηρήσουμε πώς αλλάζει το ποσοστό πρόβλεψης όλων των κλάσεων όσο μεγαλώνει το ποσοστό αλλαγής pixel, και να προσπαθήσουμε να το σχολιάσουμε. Ο σχολιασμός που ακολουθεί παρακάτω αποτελεί παρατηρήσεις πάνω σε κάποια πρώτα πειράματα που κάναμε με σκοπό να μελετηθεί το φαινόμενο.

4.5.1 Με φλιπ



Σχήμα 4.16: Flip image and bar plot

Τα ραβδογράμματα 4.16 αφορούν σε εικόνες που συμβαίνει φλιπ στην κυρίαρχη κλάση, δηλαδή που η επίθεση είναι επιτυχημένη και η κυρίαρχη κλάση αλλάζει από την πραγματική κλάση σε κάποια άλλη.

Όπως βλέπουμε και από τα ραβδογράμματα 4.16 η επίθεση αρχικά προσπαθεί να ρίξει το ποσοστό πρόβλεψης της πραγματικής κλάσης, χωρίς όμως να στοχεύει να ανεβάσει το ποσοστό πρόβλεψης κάποιας άλλης συγκεκριμένης κλάσης. Αυτό που παρατηρούμε είναι ότι ανεβαίνει το ποσοστό για τις κλάσεις που έχουν τα 2-3 μεγαλύτερα ποσοστά πρόβλεψης μετά την πραγματική κλάση, μέχρι την στιγμή που μια από αυτές υπερισχύει. Από εκείνο το σημείο και μετά το ποσοστό πρόβλεψης της κλάσης που υπερίσχυσε ανεβαίνει, ενώ τα ποσοστά των άλλων κλάσεων που αρχικά ανεβήκαν, αρχίζουν και πέφτουνε. Το ποσοστό πρόβλεψης της πραγματικής κλάσης σταθερά πέφτει. Ενδιαφέρον έχουν τα ραβδογράμματα για την δεύτερη εικόνα, για την όποια όπως βλέπουμε η πραγματική κλάση ξεκινάει με αρκετά μικρό ποσοστό, στην συνέχεια η κλάση ”γάτα” υπερισχύει και γίνεται κυρίαρχη προσωρινά, μέχρι που η κλάση ”σκύλος” την ξεπερνάει και το ποσοστό της κλάσης ”γάτα” ξαναπέφτει. Ιδιαίτερο ενδιαφέρον έχει το τελικό ποσοστό πρόβλεψης της κυρίαρχης κλάσης, το οποίο όπως βλέπουμε στις περισσότερες περιπτώσεις είναι αρκετά μικρότερο από το αρχικό ποσοστό πρόβλεψης της πραγματικής κλάσης, πράγμα που επαληθεύει τις παρατηρήσεις που κάναμε στην υποενότητα 4.3. σχετικά με το ποσοστό πρόβλεψης της κυρίαρχης κλάσης.

4.5.2 Χωρίς φλιπ

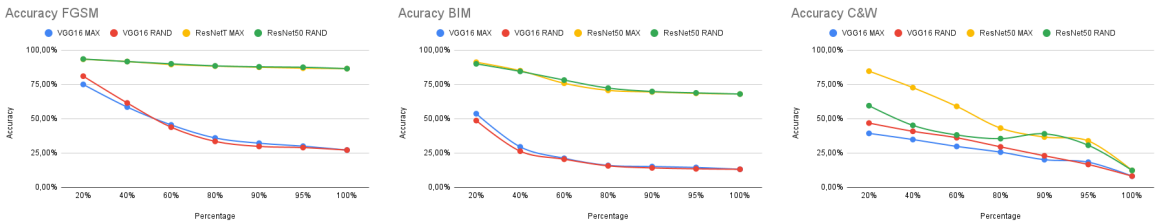
Τα ραβδογράμματα 4.17 αφορούν σε εικόνες που δε συμβαίνει φλιπ στην κυρίαρχη κλάση, δηλαδή που η επίθεση δεν είναι επιτυχημένη και η πραγματική κλάση παραμένει κυρίαρχη μέχρι το τέλος.

Όπως βλέπουμε και από τα ραβδογράμματα 4.17 η επίθεση προσπαθεί να ρίξει το ποσοστό πρόβλεψης της πραγματικής κλάσης, πάλι χωρίς να στοχεύει να ανεβάσει το ποσοστό πρόβλεψης κάποιας άλλης συγκεκριμένης κλάσης. Όπως και στην προηγούμενη υποενότητα κινείται προς την κατεύθυνση των κλάσεων με το μεγαλύτερο ποσοστό πρόβλεψης μετά την πραγματική. Η διαφορά με πριν είναι ότι δεν καταφέρνει να ρίξει το ποσοστό της πραγματικής κλάσης ούτε αντίστοιχα να ανεβάσει το ποσοστό πρόβλεψης κάποιας άλλης κλάσης, σε σημείο που να αλλάξει η κυρίαρχη κλάση και έτσι δεν συμβαίνει ποτέ το φλιπ. Αξίζει όμως να παρατηρήσουμε ότι όπως και πριν το τελικό ποσοστό πρόβλεψης της κυρίαρχης κλάσης (και πραγματικής στην προκείμενη περίπτωση) πέφτει συγκριτικά με το αρχικό ποσοστό της, κάτι που και πάλι επαληθεύει τις παρατηρήσεις που κάναμε στην υποενότητα 4.3. σχετικά με το ποσοστό πρόβλεψης της κυρίαρχης κλάσης.

Σχήμα 4.17: No flip image and bar plot

4.6 Συγκριτικά αποτελέσματα μεταξύ των επιθέσεων

Στη συγκεκριμένη ενότητα θα προσπαθήσουμε να βγάλουμε κάποια συγκριτικά αποτελέσματα μεταξύ των επιθέσεων ανεξαρτήτως της αρχιτεκτονικής που τις υλοποιήσαμε. Σκοπός είναι να βγάλουμε κάποια γενικά συμπεράσματα για την απόδοση των επιθέσεων ανεξαρτήτως της αρχιτεκτονικής που υλοποιήθηκαν και να διερευνήσουμε αν κάποια επίθεση ήταν γενικά πιο αποτελεσματική. Όπως βλέπουμε και από τα διαγράμματα 4.18 η FGSM είναι η μέθοδος που δυσκολεύεται περισσότερο να ρίξει το accuracy του μοντέλου. Αργεί αρκετά να ρίξει σημαντικά το Accuracy και η τελική τιμή που φτάνει είναι αρκετά υψηλή σε σχέση με τις υπόλοιπες επιθέσεις, οι παραπάνω παρατηρήσεις είναι λογικές καθώς όπως αναλύσαμε και στο κεφ 3.3.1. η FGSM πρόκειται για την πιο απλή μέθοδο ανταγωνιστικής επίθεσης η οποία δεν καταφέρνει τα καλύτερα αποτελέσματα. Ακόμα παρατηρούμε ότι η BIM παρουσιάζει εμφανώς καλύτερα αποτελέσματα από την FGSM καταφέρνει και ρίχνει από την αρχή αρκετά το accuracy και πετυχαίνει πολύ χαμηλότερες τιμές στο τελικό accuracy. Πάλι τα αποτελέσματα επιβεβαιώνουν την θεωρεία καθώς η BIM αποτελεί στην ουσία μια εξέλιξη της FGSM πραγματοποιώντας την ίδια λειτουργία επαναληπτικά και καταφέρνοντας καλύτερα αποτελέσματα συγκριτικά με την πρώτη. Τέλος το πιο σημαντικό ίσως συμπέρασμα που παρατηρούμε είναι ότι η CW παρουσιάζει εντυπωσιακά καλύτερα αποτελέσματα σε σχέση με τις άλλες δύο. Καταφέρνει να ρίξει δραματικά το accuracy από πολύ νωρίς και ρίχνει το τελικό accuracy αρκετά πιο χαμηλά σε σχέση με τις άλλες δύο. Η CW είναι μια αρκετά πιο σύνθετη επίθεση σε σχέση με τις άλλες δύο και με πολύ καλύτερα αποτελέσματα, επομένως είναι λογικά τα αποτελέσματα που παρατηρήσαμε.

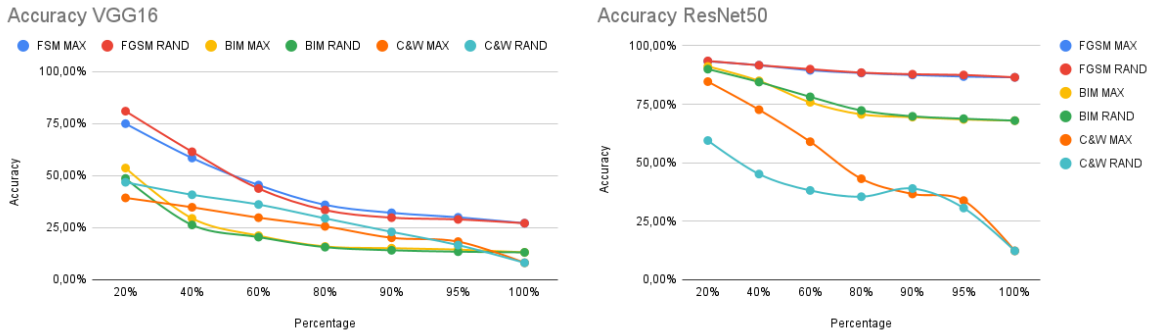


Σχήμα 4.18: Accuracy on FGSM, BIM and CW

4.7 Συγκριτικά αποτελέσματα μεταξύ των αρχιτεκτονικών

Στη συγκεκριμένη ενότητα θα προσπαθήσουμε να βγάλουμε κάποια συγκριτικά αποτελέσματα μεταξύ των αρχιτεκτονικών ανεξαρτήτως της επίθεσης που εφαρμόσαμε. Σκοπός είναι να βγάλουμε κάποια γενικά συμπεράσματα για την ανθεκτικότητα των αρχιτεκτονικών ανεξαρτήτως της επίθεσης που εφαρμόσαμε και να διερευνήσουμε αν κάποια αρχιτεκτονική ήταν γενικά πιο ανθεκτική. Από τα διαγράμματα 4.19 αρχικά παρατηρούμε ότι η αρχιτεκτονική ResNet50 είναι αρκετά πιο ανθεκτική σε σχέση με την VGG16. Αυτό μπορούμε εύκολα να το διαπιστώσουμε καθώς γενικά η ResNet50 κρατάει τα ποσοστά του Accuracy σε πολύ πιο υψηλά επίπεδα σε σχέση με την VGG16. Ειδικά για τις επιθέσεις FGSM και BIM η ResNet50 δείχνει πάρα πολύ μεγάλη ανθεκτικότητα καθώς κρατάει το ποσοστό Accuracy σε πολύ υψηλές τιμές. Αρκετά περισσότερο φαίνεται να επηρεάζεται από την CW καθώς σε αυτή τελικά το ποσοστό accuracy πέφτει πολύ χα-

μηλά, παρόλα αυτά όμως συγκριτικά με την VGG16 ακόμα και σε αυτή την επίθεση παρουσιάζει μεγαλύτερη ανθεκτικότητα καθώς διατηρεί το accuracy σε πιο ψηλά ποσοστά.



Σχήμα 4.19: Accuracy on VGG16 and ResNet50

Κεφάλαιο 5

Επίλογος

Στο παρών κεφάλαιο συνοψίζεται η δουλειά που έγινε στη διπλωματική και στη συνέχεια παρουσιάζονται επιγραμματικά τα συμπεράσματα που προέκυψαν από τα προηγούμενα κεφάλαια. Τέλος παρουσιάζονται μερικές προτάσεις για τις μελλοντικές κατευθύνσεις της επιστημονικής μελέτης.

5.1 Σύνοψη

Στην παρούσα διπλωματική εργασία ασχοληθήκαμε με μεθόδους ανταγωνιστικών επιθέσεων, εξηγήσιμης τεχνητής νοημοσύνης και τον συνδυασμό αυτών των δυο ερευνητικών πεδίων. Σκοπός μας ήταν να μελετήσουμε και να αναλύσουμε το πώς η γνώση που αποκτάμε από τις μεθόδους εξηγήσιμης τεχνητής νοημοσύνης μπορούν να συνδυαστούν με τις υπάρχουσες μεθόδους ανταγωνιστικών επιθέσεων για την δημιουργία δικών μας ανταγωνιστικών παραδειγμάτων. Δημιουργήσαμε ένα μοντέλο κατηγοριοποίησης εικόνων χρησιμοποιώντας δύο αρχιτεκτονικές με σκοπό να μελετήσουμε το πως ανταποκρίνονται στις διαφορετικές επιθέσεις, συγκεκριμένα χρησιμοποιήσαμε την αρχιτεκτονική VGG16 και την ResNet50. Δημιουργήσαμε saliency map για κάθε εικόνα , υλοποιώντας μια μέθοδο εξηγήσιμης τεχνητής νοημοσύνης και συγκεκριμένα του GradCAM. Αναλυτικότερα χρησιμοποιήσαμε δυο μεθόδους για την δημιουργία ανταγωνιστικών παραδειγμάτων και μελετήσαμε το πως επηρεάζουν την ακρίβεια του μοντέλου πειραματιζόμενοι με διάφορες τιμές. Κατασκευάσαμε τα ανταγωνιστικά παραδείγματα χρησιμοποιώντας τρεις έτοιμες μεθόδους και πιο συγκεκριμένα τις FGSM, BIM, CW, τις συγκεκριμένες επιθέσεις τις τροποποιήσαμε με δύο μεθόδους. Τις μεθόδους αυτές τις εξηγούμε αναλυτικά και σε προηγούμενα κεφάλαια, επιγραμματικά αντικαθιστούμε σε κάθε εικόνα συγκεκριμένο ποσοστό pixel με τα αντίστοιχα από τα έτοιμα ανταγωνιστικά παραδείγματα στην μια μέθοδο με τυχαίο τρόπο στην άλλη αντικαθιστώντας τα pixel που έχουν τις μέγιστες τιμές στο αντίστοιχο saliency map, για κάθε μέθοδο πειραματιστήκαμε με 7 διαφορετικά ποσοστά αλλαγής pixel. Εκτελέσαμε συγκριτικά πειράματα ανάμεσα στις δύο μεθόδους των τριών επιθέσεων πάνω στις δύο αρχιτεκτονικές. Αναλύσαμε το θεωρητικό υπόβαθρο γύρω από τις ανταγωνιστικές επιθέσεις και την εξηγήσιμη τεχνητή νοημοσύνη. Μέσα από την μελέτη και εφαρμογή όσων διαβάσαμε, εξοικειωθήκαμε με τα εργαλεία και τις μεθόδους για τα δυο επιστημονικά πεδία που αναφέραμε παραπάνω και κατανοήσαμε σε μεγάλο βαθμό τον τρόπο λειτουργίας τους.

5.2 Μελλοντικές Κατευθύνσεις Επιστημονικής Μελέτης

Σε συνέχεια της δουλειάς που κάναμε στην παρούσα διπλωματική εργασία προέκυψαν κάποιες ιδέες σχετικά με μελλοντικές κατευθύνσεις που θα μπορούσε να ακολουθήσει η επιστημονική έρευνα σχετικά με τα αντικείμενα με τα οποία καταπιαστήκαμε. Αρχικά μια πιο συστηματική μελέτη πάνω στις δυο μεθόδους που προτείναμε με υλοποίηση παραπάνω αρχιτεκτονικών αλλά και περισσότερων μεθόδων επίθεσης θα μπορούσε να δώσει κάποια πιο γενικεύσιμα συμπεράσματα. Ακόμα θα μπορούσε να μελετηθεί παραπάνω η πτώση του ποσοστού της πραγματικής και της κυρίαρχης κλάσης για τα οποία κάναμε κάποιες παρατηρήσεις αλλά δεν μπορέσαμε να τρέξουμε ικανοποιητικό αριθμό παραδειγμάτων για να μπορέσουμε να τα μελετήσουμε συστηματικά. Τέλος περαιτέρω συστηματική μελέτη μπορεί να γίνει στο πως επηρεάζονται τα ποσοστά όλων των κλάσεων όσο εξελίσσονται οι επιθέσεις, η οποία θα μπορούσε να βοηθήσει να κατανοήσουμε καλύτερα το πως κατευθύνονται οι διάφορες επιθέσεις και διαλέγουν ποιων κλάσεων το ποσοστό πρόβλεψη θα ανεβάσουν.

Βιβλιογραφία

Facebook removed 3 billion fake accounts in just 6 months. URL https://nypost.com/2019/05/23/facebook-removed-3-billion-fake-accounts-in-just-6-months/.

Ferreti Marchetti Colajanni Apruzzese, Adreolini. Modeling realistic adversarial attacks against network intrusion detection systems. 06 2021.

Anish Athalye. Ai image recognition fooled by single pixel change. URL https://www.bbc.com/news/technology-41845878.

Joseph Tygar Barreno, Nelson. The security of machine learning. *Mach. Learn.*, 81(2):121–148, 2010. doi: 10.1007/s10994-010-5188-5.

Moran Baruch, Gilad Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. *CoRR*, abs/1902.06156, 2019. URL http://arxiv.org/abs/1902.06156.

B.Biggio. *Multiple Classifier Systems for Robust Classifier Design in Adversarial Environment*, 1:27–41, 2010.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines, 2012. URL https://arxiv.org/abs/1206.6389.

Battista Biggio, Igino Corona, Blaine Nelson, Benjamin I. P. Rubinstein, Davide Maiorca, Giorgio Fumera, Giorgio Giacinto, and Fabio Roli. Security evaluation of support vector machines in adversarial environments. *CoRR*, abs/1401.7727, 2014. URL http://arxiv.org/abs/1401.7727.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. *CoRR*, abs/1708.06131, 2017a. URL http://arxiv.org/abs/1708.06131.

Battista Biggio, Giorgio Fumera, and Fabio Roli. Security evaluation of pattern classifiers under attack. *CoRR*, abs/1709.00609, 2017b. URL http://arxiv.org/abs/1709.00609.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Byzantine-tolerant machine learning. *CoRR*, abs/1703.02757, 2017. URL http://arxiv.org/abs/1703.02757.

Scheffer Bruckner, Kanzow. Static prediction games for adversarial learning problems. 2012.

Gabriel Dias Cantareira, Rodrigo Fernandes de Mello, and Fernando V. Paulovich. Explainable adversarial attacks in deep neural networks using activation profiles. *CoRR*, abs/2103.10229, 2021. URL https://arxiv.org/abs/2103.10229.

Wagner Carlini. Towards evaluating the robustness of neural networks. 08 2016.

Tanmay Chakraborty, Utkarsh Trehan, Khawla Mallat, and Jean-Luc Dugelay. Generalizing adversarial explanations with grad-cam. *ArXiv*, abs/2204.05427, 2022.

Lingjiao Chen, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. DRACO: Byzantine-resilient distributed training via redundant gradients. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 903–912. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/chen18l.html.

Deepesh Data and Suhas Diggavi. Byzantine-resilient high-dimensional sgd with local iterations on heterogeneous data. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2478–2488. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/data21a.html.

Ben Dickson. Machine learning: What are membership inference attacks? URL https://bdtechtalks.com/2021/04/23/machine-learning-membership-inference-attacks/.

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame, 2019. URL https://arxiv.org/abs/1906.07983.

El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in Byzantium. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3521–3530. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/mhamdi18a.html.

El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê Nguyên Hoang, and Sébastien Rouault. Collaborative learning as an agreement problem. *CoRR*, abs/2008.00742, 2020. URL https://arxiv.org/abs/2008.00742.

Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. Explainable AI: current status and future directions. *CoRR*, abs/2107.07045, 2021. URL https://arxiv.org/abs/2107.07045.

Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. Planting undetectable backdoors in machine learning models. *ArXiv*, abs/2204.06974, 2022.

Joao Gomes. Adversarial attacks and defences for convolutional neural networks. URL https://medium.com/onfido-tech/adversarial-attacks-and-defences-for-convolutional-neural-networks-66915ece52e7.

Szegedy Goodfellow, Shlens. Explaining and harnessing adversarial examples. 2015.

Chen Laurish Zschech Heinrich, Graf. Fool me once, shame on you, fool me twice, shame on me: A taxonomy of attack and de-fense patterns for ai security. 2020.

Marques-Silva Ignatiev, Narodytska. On relating explanations and adversarial examples. 2019.

Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 19–35, 2018. doi: 10.1109/SP.2018.00057.

Nicolas Papernot Kalpesh Krishna. How to steal modern nlp systems with gibberish? URL https://www.cleverhans.io/2020/04/06/stealing-bert.html.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=jXKKDEi5vJt.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. Distributed momentum for byzantine-resilient stochastic gradient descent. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=H8UHdhWG6A3.

Guirguis Laurish Zschech Mhamdi, Guerraoui. Genuinely distributed byzantine machine learning. 04 2022. doi: https://doi.org/10.1007/s00446-022-00427-9.

Blaine Nelson, Benjamin I. P. Rubinstein, Ling Huang, Anthony D. Joseph, Steven J. Lee, Satish Rao, and J. D. Tygar. Query strategies for evading convex-inducing classifiers. *CoRR*, abs/1007.0484, 2010. URL http://arxiv.org/abs/1007.0484.

Chris Price. Facebook removes 15 billion fake accounts in two years. URL https://www.techdigest.tv/2021/09/facebook-removes-15-billion-fake-accounts-in-two-years.html.

Nicolas Pröllochs, Stefan Feuerriegel, and Dirk Neumann. Learning interpretable negation rules via weak supervision at document level: A reinforcement learning approach. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 407–413, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1038. URL https://aclanthology.org/N19-1038.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning, 2019.

Govindaraju Rodrigues, Ling. *Journal of Visual Languages and Computing*, 23:828–841, 6 2009. doi: https://doi.org/10.48550/arXiv.1710.08864.

Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9389–9398. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/schwarzschild21a.html.

Xiaoyu Li Junhong Duan Dejun Mu Xiao Jing Sensen Guo, Jinxiong Zhao. Security and privacy challenges in internet of things and mobile edge computing. 2021. doi: https://doi.org/10.1155/2021/5578335.

Leslie F. Sikos. *AI in Cybersecurity*. Intelligent Systems Reference Library. Springer Cham, 2019. ISBN 978-3-319-98842-9.

Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial machine learning-industry perspectives. In *2020 IEEE Security and Privacy Workshops (SPW)*, pages 69–75, 2020. doi: 10.1109/SPW50608.2020.00028.

Kouichi Su, Vargas. *IEEE Transactions on Evolutionary Computation*, 20:169–179, 10 2009. doi: https://doi.org/10.1016/j.jvlc.2009.01.010.

Ken Tsui. Perhaps the simplest introduction of adversarial examples ever. URL https://towardsdatascience.com/perhaps-the-simplest-introduction-of-adversarial-examples-ever-c0839a759b8d.

Praca Vitorino, Oliveira. Adaptative perturbation patterns: Realistic adversarial learning for robust intrusion detection. 03 2022.