



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

Δημιουργία Εικόνας από Κείμενο με Χρήση Γεννητικών Ανταγωνιστικών Δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΘΑΝΑΣΙΟΣ Π. ΜΑΣΟΥΡΗΣ

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

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

Αθήνα, Νοέμβριος 2020



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

Δημιουργία Εικόνας από Κείμενο με Χρήση Γεννητικών Ανταγωνιστικών Δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΘΑΝΑΣΙΟΣ Π. ΜΑΣΟΥΡΗΣ

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

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 02η Νοεμβρίου 2020.

.....
Ανδρέας Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

.....
Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

.....
Γεώργιος Στάμου
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2020

.....
Αθανάσιος Π. Μασούρης

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

Copyright © Αθανάσιος Π. Μασούρης, 2020.
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Η δημιουργία εικόνας από κείμενο αποτελεί αντικείμενο έρευνας που συνδυάζει τους τομείς της Όρασης Υπολογιστών και της Επεξεργασίας Φυσικής Γλώσσας. Στόχος είναι η δημιουργία ενός μοντέλου το οποίο, δοθείσης μιας λεκτικής περιγραφής, θα παράγει εικόνες. Οι εικόνες αυτές θα πρέπει όχι μόνο να είναι ρεαλιστικές, αλλά και οι οπτικές λεπτομέρειες που περιέχουν να αντιστοιχούν στην προαναφερθείσα λεκτική περιγραφή.

Με την εμφάνιση των Γεννητικών Ανταγωνιστικών Δικτύων (Generative Adversarial Networks, GANs) σημειώθηκε σημαντική πρόοδος στη συγκεκριμένη κατεύθυνση. Τα συστήματα που έχουν προταθεί έχουν τη δυνατότητα να παράξουν εικόνες υψηλής ανάλυσης, οι οποίες ταυτίζονται με την εκάστοτε λεκτική περιγραφή, με τη χρήση διαφόρων τεχνικών. Η δομή στοιβαγμένων GANs αποτελεί ίσως τη σημαντικότερη από αυτές τις τεχνικές. Τα υπάρχοντα συστήματα παράγουν μία αρχική εικόνα χαμηλών διαστάσεων, η οποία περνά από έναν αριθμό σταδίων βελτίωσης προκειμένου να παραχθεί η τελική εικόνα υψηλών διαστάσεων.

Στην παρούσα διπλωματική διατριβή προτείνουμε μία νέα αρχιτεκτονική για τη δημιουργία εικόνων υψηλών διαστάσεων. Πιο συγκεκριμένα, χρησιμοποιούμε τη δομή στοιβαγμένων GANs, με τρία επίπεδα, προκειμένου να διασπάσουμε το συνολικό δύσκολο πρόβλημα παραγωγής εικόνας υψηλών διαστάσεων σε επιμέρους διαχειρίσιμα υποπροβλήματα. Στο πρώτο στάδιο παράγεται μία ασπρόμαυρη εικόνα διαστάσεων 128x128. Στο δεύτερο στάδιο προστίθενται χρώματα στην εικόνα του πρώτου σταδίου. Τέλος, στο τρίτο και τελευταίο στάδιο η εικόνα του δεύτερου σταδίου μεγεθύνεται προκειμένου να παραχθεί εικόνα υψηλών διαστάσεων (256x256).

Ακόμη, στα πλαίσια αυτής της εργασίας, εξετάζεται η επίδραση των αναπαραστάσεων κειμένου, που προέκυψαν από τα μοντέλα char-CNN-RNN, GPT-2 και RoBERTa, στην ποιότητα των παραγόμενων εικόνων των μοντέλων gan-int-cls και StackGAN για τα σύνολα δεδομένων Oxford-102 και CUB. Γίνεται, επίσης, εκπαίδευση των δικτύων αυτών στο σύνολο δεδομένων Flickr8k και παρουσιάζονται τα αποτελέσματα.

Λέξεις κλειδιά

Όραση υπολογιστών, Δημιουργία εικόνας από κείμενο, Γεννητικά Ανταγωνιστικά Δίκτυα, Στοιβαγμένα Γεννητικά Ανταγωνιστικά Δίκτυα.

Abstract

The problem of text-to-image synthesis is a research area that combines the fields of Computer Vision and Natural Language Processing. The goal is to create a model which, given a text description, generates images. These images must not only be realistic but also contain visual details that match the aforementioned text description.

The emergence of Generative Adversarial Networks (GANs) marked a period of significant progress in this direction. The systems that have been proposed can generate high-resolution images that match their corresponding text description using a variety of techniques. Stacked GANs probably constitute the most important development in this direction. Existing models generate an initial image of low quality, which passes through a number of sketch-refinement processing stages in order to generate the high-resolution image.

In this diploma dissertation, we propose a novel architecture to generate high-resolution images. In particular, we use the Stacked GANs structure, with three stages, in order to decompose the difficult problem of generating images of high quality into more manageable sub-problems. More specifically, the network of the first stage generates a black and white image of 128x128 resolution. At the second stage, colors are added to the image of the first stage. Finally, at the third and last stage, the image of the second stage is enhanced to high resolution (256x256).

In addition, we examine the impact of different text representations, produced by char-CNN-RNN, GPT-2 and RoBERTa language models, on the quality of generated images from gan-int-cls and Stack-GAN models on Oxford-102 and CUB datasets. We also train these networks on the Flickr8k dataset and produce results.

Key words

Computer Vision, Text-to-image synthesis, Generative Adversarial Networks (GANs), Stacked GANs.

Ευχαριστίες

Η παρούσα διπλωματική εργασία σηματοδοτεί την ολοκλήρωση των προπτυχιακών σπουδών μου. Συνεπώς, θα ήθελα να ευχαριστήσω θερμά όλους όσους με υποστήριξαν στην εκπόνηση αυτής της διπλωματικής αλλά και συνολικά με τις σπουδές μου.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή αυτής της διατριβής, κ. Ανδρέα-Γεώργιο Σταφυλοπάτη, για την ευκαιρία που μου προσέφερε να εκπονήσω τη συγκεκριμένη διπλωματική καθώς και για την εμπιστοσύνη που μου έδειξε. Παράλληλα, θα ήθελα να ευχαριστήσω τον συνεπιβλέποντα, κ. Γεώργιο Αλεξανδρίδη, για τη συνεχή υποστήριξη και καθοδήγησή του. Οι πολύτιμες συμβουλές του και οι χρήσιμες συζητήσεις που πραγματοποιήσαμε με βοήθησαν να ξεπεράσω σημαντικά εμπόδια που συνάντησα κατά την εκπόνηση της εργασίας. Επιπροσθέτως, θα ήθελα να ευχαριστήσω τους καθηγητές, κ. Στέφανο Κόλλια και κ. Γεώργιο Στάμου, που με τίμησαν με την παρουσία τους στην τριμελή επιτροπή εξέτασης.

Σε προσωπικό επίπεδο θα ήθελα πρωτίστως να ευχαριστήσω την οικογένειά μου και κυρίως τους γονείς και τον αδερφό μου. Συγκεκριμένα, ευχαριστώ από τα βάθη της καρδιάς μου τη μητέρα μου, Κατερίνα Σπανού, τον πατέρα μου, Παναγιώτη Μασούρη, καθώς και τον αδερφό μου, Γιάννη Μασούρη, για τη συνεχή ενθάρρυνση και υποστήριξή τους. Τέλος, θα ήθελα να ευχαριστήσω όλους μου τους φίλους, εντός και εκτός σχολής, που στάθηκαν στο πλευρό μου όλα αυτά τα χρόνια.

Αθανάσιος Π. Μασούρης,

Αθήνα, 02η Νοεμβρίου 2020

Περιεχόμενα

Περίληψη	5
Abstract	7
Ευχαριστίες	9
Περιεχόμενα	11
Κατάλογος σχημάτων	15
1. Εισαγωγή	17
1.1 Κίνητρο	17
1.2 Συμβολή	18
1.3 Δομή εργασίας	18
2. Γεννητικά Ανταγωνιστικά Δίκτυα	21
2.1 Αρχιτεκτονικές	21
2.1.1 Βασικό GAN	21
2.1.2 GAN υπό συνθήκη	22
2.1.3 Βαθύ Συνελικτικό GAN	23
2.2 Συναρτήσεις σφάλματος	23
2.2.1 Τετραγωνικό σφάλμα	23
2.2.2 Σφάλμα Διασταυρούμενης Εντροπίας	24
2.3 Μέθοδοι εκπαίδευσης	24
2.3.1 Αλγόριθμος Στοχαστικής Κατάβασης Κλίσης	24
2.3.2 Οπίσθια διάδοση του σφάλματος	24
2.4 Συνήθη προβλήματα των GANs	24
2.4.1 Αδυναμία βελτίωσης παραγόμενων δειγμάτων	24
2.4.2 Κατάρρευση Συστήματος	25
2.5 Τεχνικές βελτιστοποίησης εκπαίδευσης	25
2.5.1 Τροποποιημένη συνάρτηση σφάλματος	25
2.5.2 Διαχωρισμός δεσμών δεδομένων	25
2.5.3 Εξομάλυνση ετικετών εκπαίδευσης	25
2.5.4 Προσθήκη θορύβου στην είσοδο του διευκρινιστή	25
3. Σύνθεση εικόνας από κείμενο με χρήση GANs	27
3.1 GAN-CLS	28
3.1.1 Διανυσματική αναπαράσταση κειμένου	28
3.1.2 Αρχιτεκτονική δικτύου	28
3.1.3 Τροποποιημένη συνάρτηση σφάλματος	29
3.1.4 Παραλλαγή GAN-INT-CLS	30
3.2 StackGAN	30
3.2.1 Επαύξηση Διανυσματικής Αναπαράστασης	30

3.2.2	Πρώτο στάδιο	31
3.2.3	Δεύτερο στάδιο	32
3.3	StackGAN++	33
3.3.1	Αρχιτεκτονική δικτύου	33
3.3.2	Μοντελοποίηση παραγωγής εικόνας υπό συνθήκη	34
3.3.3	Συναρτήσεις βελτιστοποίησης	34
4.	TeleGAN	35
4.1	Επαύξηση Διανυσματικής Αναπαράστασης	35
4.2	Στάδιο-1	35
4.2.1	Αρχιτεκτονική	36
4.2.2	Λεπτομέρειες Υλοποίησης	37
4.3	Στάδιο-2	37
4.3.1	Αρχιτεκτονική	38
4.3.2	Λεπτομέρειες Υλοποίησης	38
4.4	Στάδιο-3	39
4.4.1	Αρχιτεκτονική	41
4.4.2	Λεπτομέρειες Υλοποίησης	41
4.5	Εκπαίδευση	42
5.	Πειραματική Διαδικασία και Αποτελέσματα	43
5.1	Σύνολα Δεδομένων	43
5.2	Μετρικές Αξιολόγησης	44
5.2.1	Inception Score	44
5.2.2	Fréchet Inception Distance	44
5.3	Σύγκριση διανυσματικών αναπαραστάσεων δεδομένων	45
5.4	Σύγκριση με άλλα δίκτυα	46
5.5	Έλεγχος για κατάρρευση συστήματος	47
5.6	Παρουσίαση αποτελεσμάτων σε CUB και Oxford-102	48
5.7	Σύνολο δεδομένων Flickr8k	52
6.	Συμπεράσματα και Μελλοντικές Κατευθύνσεις	53
	Βιβλιογραφία	55
	Παράρτημα	61
A.	Δομικά στοιχεία Συνελκτικών Νευρωνικών Δικτύων	61
A.1	Συνελκτικό επίπεδο	61
A.2	Συγκεντρωτικό επίπεδο	62
A.3	Πλήρως Συνδεδεμένο επίπεδο	63
A.4	Συναρτήσεις Ενεργοποίησης	64
A.4.1	Σιγμοειδής	64
A.4.2	Συνάρτηση υπερβολικής εφαπτομένης	64
A.4.3	Συνάρτηση διορθωμένης γραμμικής μονάδας	64
A.4.4	Διαρρέουσα συνάρτηση διορθωμένης γραμμικής μονάδας	65
A.5	Κανονικοποίηση δέσμης δεδομένων	65
A.6	Αντιμεταθετικό συνελκτικό επίπεδο	65

B. Επιπλέον αποτελέσματα	67
B.1 Απεικόνιση συνθετικών εικόνων στον χώρο	67
B.2 Πρόσθετα αποτελέσματα στο Oxford-102	68
B.3 Πρόσθετα αποτελέσματα στο CUB	68
B.4 Περιπτώσεις αποτυχημένης σύνθεσης εικόνας	69
C. Ευρετήριο Ακρωνυμίων και Συντμήσεων	71

Κατάλογος σχημάτων

1.1	Συνθετικές εικόνες του StyleGAN2 (Πηγή: https://github.com/NVlabs/stylegan2)	18
1.2	Συνθετικές εικόνες του DM-GAN (Πηγή: [Zhu19])	19
2.1	Διαγραμματική απεικόνιση του βασικού GAN.	21
2.2	Διαγραμματική απεικόνιση του CGAN.	22
2.3	Ο γεννήτορας του DCGAN (Πηγή: [Radf15])	23
2.4	Διαχωρισμός δεσμών δεδομένων (mini batches) σε δέσμες από πραγματικά και συνθετικά δείγματα αντίστοιχα. (Πηγή: https://github.com/soumith/ganhacks)	26
3.1	Επιφανειακή (high-level) απεικόνιση της αρχιτεκτονικής του δικτύου GAN-INT-CLS. (Πηγή: [Reed16b])	29
3.2	Απεικόνιση της αρχιτεκτονικής του δικτύου StackGAN. (Πηγή: [Zhan16])	32
3.3	Απεικόνιση της αρχιτεκτονικής του δικτύου StackGAN++. (Πηγή: [Zhan17])	33
4.1	Απεικόνιση της αρχιτεκτονικής του δικτύου του πρώτου σταδίου του TeleGAN.	36
4.2	Απεικόνιση της αρχιτεκτονικής του δικτύου του δεύτερου σταδίου του TeleGAN.	39
4.3	Απεικόνιση της αρχιτεκτονικής του δικτύου του τρίτου σταδίου του TeleGAN.	41
5.1	Έλεγχος για κατάρρευση συστήματος με τη χρήση του αλγορίθμου t-SNE.	48
5.2	Συνθετικές εικόνες του TeleGAN στο σύνολο δεδομένων CUB.	49
5.3	Συνθετικές εικόνες του TeleGAN στο σύνολο δεδομένων Oxford-102.	50
5.4	Συνθετικές εικόνες του TeleGAN στα σύνολα δεδομένων CUB και Oxford-102	51
5.5	Συνθετικές εικόνες των GAN-INT-CLS, StackGAN και TeleGAN στο σύνολο δεδομένων Flickr8k.	52
A.1	Αρχιτεκτονική Συνελκτικού Νευρωνικού Δικτύου	61
A.2	Παράδειγμα εφαρμογής συνέλιξης ενός τυχαίου φίλτρου με μία τυχαία είσοδο.	62
A.3	Παράδειγμα εφαρμογής συγκεντρωτικού επιπέδου με συγκέντρωση μεγίστου.	63
A.4	Παράδειγμα εφαρμογής συγκεντρωτικού επιπέδου με συγκέντρωση μέσης τιμής.	63
A.5	(a) Σιγμοειδής συνάρτηση, (b) Υπερβολική εφαπτομένη	64
A.6	(a) ReLU, (b) Leaky ReLU	65
A.7	Παράδειγμα εφαρμογής αντιμεταθετικής συνέλιξης ενός τυχαίου φίλτρου με μία τυχαία είσοδο.	66
B.1	Απεικόνιση συνθετικών εικόνων του TeleGAN, στα σύνολα δεδομένων Oxford-102 και CUB, σε χάρτη δύο διαστάσεων με τη χρήση του αλγορίθμου t-SNE. Σε κάθε χάρτη απεικονίζονται 5,000 συνθετικές εικόνες.	67
B.2	Επιπλέον συνθετικές εικόνες του TeleGAN στο σύνολο δεδομένων Oxford-102.	68
B.3	Επιπλέον συνθετικές εικόνες του TeleGAN στο σύνολο δεδομένων CUB.	68
B.4	Αποτυχημένες συνθετικές εικόνες του TeleGAN στο σύνολο δεδομένων Oxford-102.	69
B.5	Αποτυχημένες συνθετικές εικόνες του TeleGAN στο σύνολο δεδομένων CUB.	69

Κεφάλαιο 1

Εισαγωγή

Οι Walter Pitts και Warren McCulloch ανέλυσαν για πρώτη φορά τη λειτουργία ενός τεχνητού νευρώνα το 1943. Στο έργο τους [McCu43] παρουσίασαν τη δυνατότητα του προαναφερθέντα νευρώνα να εκτελέσει απλές λογικές πράξεις και περιέγραψαν αυτό που σήμερα αποκαλούμε νευρωνικό δίκτυο. Ο Alan Turing επτά χρόνια αργότερα αναφέρθηκε για πρώτη φορά στη δυνατότητα ενός υπολογιστή να «σκεφτεί». Στο έργο του [Tur09] εξήγησε ότι είναι δύσκολο να οριστεί η έννοια της «σκέψης» και για τον σκοπό αυτόν περιέγραψε μία διαδικασία, το γνωστό Turing Test. Η διαδικασία ελέγχου είναι αρκετά απλή: *για να επιτύχει ένας υπολογιστής, θα πρέπει μετά από αλληλεπίδρασή του με έναν άνθρωπο, να μην είναι σε θέση ο άνθρωπος να διακρίνει αν πρόκειται για υπολογιστή ή άνθρωπο.*

Στη σημερινή εποχή η εξέλιξη της τεχνητής νοημοσύνης και της βαθιάς μάθησης έχει οδηγήσει σε συστήματα τα οποία είναι σε θέση όχι μόνο να ξεγελάσουν έναν άνθρωπο, δηλαδή να επιτύχουν υπό προϋποθέσεις στο Turing Test, αλλά και να ξεπεράσουν τις ανθρώπινες επιδόσεις σε ένα μεγάλο εύρος προβλημάτων. Ιδιαίτερο ενδιαφέρον παρουσιάζει η επικράτηση τέτοιου είδους συστημάτων σε εφαρμογές της *όρασης υπολογιστών*. Ορισμένα χαρακτηριστικά παραδείγματα συστημάτων που πέτυχαν καλύτερη απόδοση έναντι ανθρώπων σε διάφορα προβλήματα της όρασης υπολογιστών είναι: το σύστημα των ερευνητών της Microsoft [He16] στη δοκιμασία αναγνώρισης αντικειμένων του ImageNet [Russ15], το δίκτυο LipNet [Assa16] που κλήθηκε να «διαβάσει» τα χείλη ανθρώπων σε ένα βίντεο χωρίς ήχο καθώς και ιδιαίτερα σημαντικές ιατρικές εφαρμογές όπως η αναγνώριση ασθενειών από εικόνες εξετάσεων [Liu19a].

1.1 Κίνητρο

Ανεξάρτητα από τη δυνατότητα επίτευξης καλύτερης επίδοσης από αυτή του ανθρώπου, ο τομέας της όρασης υπολογιστών έχει σημειώσει σημαντική πρόοδο. Τα συστήματα είναι σε θέση να αναγνωρίσουν σε μεγάλο βαθμό τι απεικονίζεται σε μία εικόνα. Ωστόσο, αυτό δε σημαίνει ότι έχει επιτευχθεί μία καθολική κατανόηση του οπτικού κόσμου, ή ακόμα καλύτερα μίας κατανομής των οπτικών χαρακτηριστικών, καθώς σε αυτή την περίπτωση θα ήταν σε θέση όχι μόνο να αναγνωρίσουν αλλά και να παράξουν ένα οπτικό αποτέλεσμα.

Με βάση αυτή την παρατήρηση, μία καλύτερη προσέγγιση για την κατανόηση του οπτικού κόσμου είναι τα *γεννητικά μοντέλα*. Πρόκειται ουσιαστικά για μοντέλα τα οποία εκπαιδεύονται σε ένα συγκεκριμένο σύνολο δεδομένων εικόνων και κατορθώνουν να μάθουν μία κατανομή οπτικών χαρακτηριστικών με βάση την οποία παράγουν συνθετικές εικόνες που είναι πανομοιότυπες με τις πραγματικές. Διάφορα είδη γεννητικών μοντέλων [Suts09, Lee09, King13] έχουν κατορθώσει να παράξουν ρεαλιστικές εικόνες και κατ' επέκταση να συμβάλλουν στην επίλυση του προβλήματος. Η εμφάνιση όμως των *Γεννητικών Ανταγωνιστικών Δικτύων* (generative adversarial networks - GANs) [Good14] ήταν καθοριστική για την εξέλιξη του τομέα. Ιδιαίτερα σημαντικό ρόλο διαδραμάτισε και η παρουσίαση μίας ιδιαίτερης αρχιτεκτονικής, αυτή του *Βαθιού Συνελικτικού Γεννητικού Ανταγωνιστικού Δίκτυου* (deep convolutional GAN - DCGAN) [Radf15]. Τα σύγχρονα GANs έχουν τη δυνατότητα να παράξουν συνθετικές εικόνες που δε διακρίνονται από τις πραγματικές. Στο Σχήμα 1.1 παρουσιάζονται ορισμένα αποτελέσματα του StyleGAN [Karr19a, Karr19b].

Ιδιαίτερο ενδιαφέρον παρουσιάζει και μία επέκταση του παραπάνω προβλήματος, η οποία αποτε-



Σχήμα 1.1: Συνθετικές εικόνες του StyleGAN2 (Πηγή: <https://github.com/NVlabs/stylegan2>)

λεί και αντικείμενο της συγκεκριμένης διπλωματικής εργασίας. Πιο συγκεκριμένα η επέκταση αυτή αφορά το πρόβλημα της *Σύνθεσης Εικόνας από Κείμενο*, ή διαφορετικά «*Δοσμένης μίας λεκτικής περιγραφής, σύνθεσε εικόνα που να αντιστοιχεί στην περιγραφή αυτή*». Το πλήθος εφαρμογών ενός τέτοιου συστήματος, καθιστά την προσπάθεια υλοποίησής του ιδιαίτερα σημαντική και ενδιαφέρουσα. Ορισμένα παραδείγματα εφαρμογών είναι η σχεδίαση προϊόντων προσαρμοσμένων σε λεκτικές περιγραφές των πελατών, η σχεδίαση της προσωπογραφίας ενός δράστη από την περιγραφή ενός θύματος και η απεικόνιση του αποτελέσματος μίας συνταγής μαγειρέματος [Han20]. Επίσης, η επέκταση των συστημάτων αυτών με ακολουθιακή δομή μπορεί να χρησιμοποιηθεί προκειμένου να οπτικοποιηθεί μία σύντομη ιστορία [Li19b].

Οι παραπάνω πιθανές εφαρμογές σε συνδυασμό με τα εντυπωσιακά αποτελέσματα που έχουν παρουσιάσει σύγχρονες προσεγγίσεις του προβλήματος [Reed16b, Zhan16, Zhan17, Xu18, Li19a, Qiao19, Zhu19] αποτέλεσαν το κίνητρο για την ερευνητική ενασχόληση με το συγκεκριμένο πρόβλημα. Στο Σχήμα 1.2 παρουσιάζονται ορισμένα αποτελέσματα του [Zhu19].

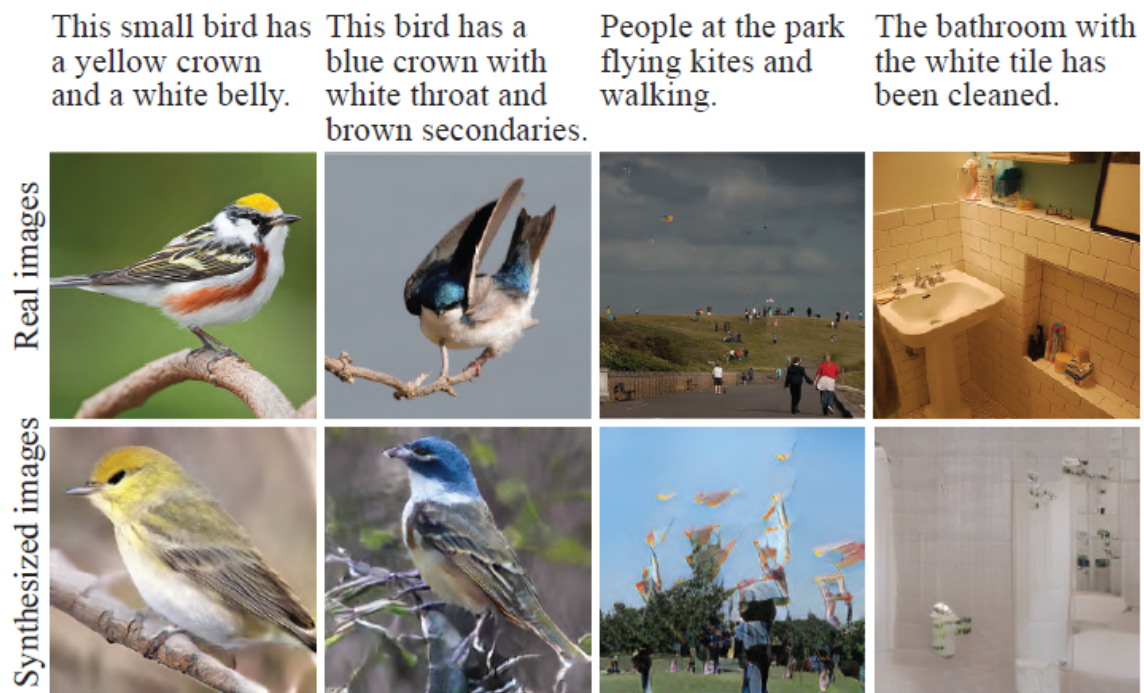
1.2 Συμβολή

Το κύριο αντικείμενο της παρούσας διατριβής είναι η ερευνητική ενασχόληση με το πρόβλημα της σύνθεσης εικόνας από κείμενο. Για την πραγματοποίηση αυτού του σκοπού, πραγματοποιήθηκαν τα εξής:

- Παραγωγή πηγαίου κώδικα για τα μοντέλα gan-int-cls, StackGAN, StackGANv2 και αναπαραγωγή των αποτελεσμάτων τους για τα σύνολα δεδομένων CUB και Oxford 102 Flowers με σκοπό την επαλήθευση της επίδοσής τους.
- Εκπαίδευση των παραπάνω μοντέλων με αναπαραστάσεις κειμένων που προέκυψαν από τα σύγχρονα μοντέλα GPT-2 και RoBERTa και έλεγχος της επίδρασης της επιλεγόμενης αναπαράστασης ως προς την ποιότητα των παραγόμενων εικόνων.
- Εκπαίδευση των παραπάνω μοντέλων στο σύνολο δεδομένων Flickr8k με αναπαραστάσεις κειμένων που προέκυψαν από το μοντέλο BERT και έλεγχος επίδοσης.
- Πρόταση μίας πρωτότυπης αρχιτεκτονικής Γεννητικού Ανταγωνιστικού Δικτύου για τη δημιουργία εικόνων υψηλών διαστάσεων από λεκτικές περιγραφές.

1.3 Δομή εργασίας

Η παρούσα διπλωματική εργασία δομείται σε 6 κεφάλαια. Πέρα από την εισαγωγή (Κεφάλαιο 1), στο Κεφάλαιο 2 καλύπτεται το θεωρητικό υπόβαθρο των Γεννητικών Ανταγωνιστικών Δικτύων προκειμένου να είναι δυνατή η κατανόηση της αρχιτεκτονικής των συστημάτων που θα παρουσιαστούν στη συνέχεια. Στο Κεφάλαιο 3 γίνεται μία σύντομη εισαγωγή στο πρόβλημα της Σύνθεσης



Σχήμα 1.2: Συνθετικές εικόνες του DM-GAN (Πηγή: [Zhu19])

Εικόνες από Κείμενο με τη χρήση GANs και παρουσιάζονται οι αρχιτεκτονικές των ερευνητικών δημοσιεύσεων που θα αξιολογηθούν στα πλαίσια αυτής της εργασίας. Στο Κεφάλαιο 4 παρουσιάζεται η αρχιτεκτονική και οι λεπτομέρειες υλοποίησης του συστήματος που προτείνεται ως συνεισφορά της συγκεκριμένης διατριβής. Στο Κεφάλαιο 5 αναλύονται σε βάθος τα πειράματα που πραγματοποιήθηκαν καθώς και τα αποτελέσματα που προέκυψαν. Τέλος, το Κεφάλαιο 6 ολοκληρώνει την παρούσα εργασία παρουσιάζοντας τα συμπεράσματα που προέκυψαν από τη συγκεκριμένη έρευνα που πραγματοποιήθηκε καθώς και πιθανές μελλοντικές κατευθύνσεις.

Στα παραρτήματα παρουσιάζονται επίσης η θεωρία για τα δομικά στοιχεία των συνελκτικών νευρωνικών δικτύων, προκειμένου να διευκολυνθεί η αναγνώστης στην κατανόηση των αρχιτεκτονικών των Γεννητικών Ανταγωνιστικών Δικτύων που παρουσιάζονται στην εργασία και περαιτέρω αποτελέσματα από τις πειραματικές διαδικασίες.

Κεφάλαιο 2

Γεννητικά Ανταγωνιστικά Δίκτυα

Το *γεννητικό ανταγωνιστικό δίκτυο* (generative adversarial network - GAN) [Good14] είναι ένα ειδικό μοντέλο εκπαίδευσης γεννητικού δικτύου επηρεασμένο από τη θεωρία παιγνίων. Το μοντέλο αυτό αποτελείται από δύο διακριτά νευρωνικά δίκτυα τα οποία ανταγωνίζονται μεταξύ τους. Το ένα δίκτυο ονομάζεται *γεννήτορας* (generator) και αποτελεί το γεννητικό δίκτυο το οποίο είναι υπεύθυνο για τη σύνθεση νέων δειγμάτων. Το άλλο δίκτυο ονομάζεται *διευκρινιστής* (discriminator) και αποτελεί το διευκρινιστικό δίκτυο, του οποίου σκοπός είναι να αναγνωρίζει αν ένα δείγμα προέρχεται από την κατανομή των σύνθετων δειγμάτων του γεννήτορα ή την κατανομή των πραγματικών δεδομένων.

2.1 Αρχιτεκτονικές

2.1.1 Βασικό GAN

Το Σχήμα 2.1 απεικονίζει διαγραμματικά το βασικό GAN (vanilla GAN), όπως προτάθηκε από τον Ian Goodfellow και τους συνεργάτες του [Good14]. Αποτελείται από τον γεννήτορα, που παίρνει ως είσοδο τυχαίο θόρυβο και παράγει συνθετικά/ψεύτικα δείγματα και τον διευκρινιστή, που παίρνει ως είσοδο τα δείγματα και αναγνωρίζει αν έχουν κατασκευαστεί από τον γεννήτορα ή είναι όντως πραγματικά δεδομένα. Και τα δύο δίκτυα, του γεννήτορα και του διευκρινιστή, αποτελούνται από *πολυεπίπεδα perceptrons* (multilayer perceptrons - MLPs).



Σχήμα 2.1: Διαγραμματική απεικόνιση του βασικού GAN.

Προκειμένου να βρεθεί η κατανομή p_g του γεννήτορα ως προς τα δεδομένα x , ορίζεται ως είσοδος ο θόρυβος z με κατανομή $p_z(z)$ και δημιουργείται η συσχέτιση $G(z; \theta_g)$ με τον χώρο των δεδομένων, με τη χρήση της διαφοροποιήσιμης συνάρτησης G που αντιπροσωπεύεται από ένα MLP με παραμέτρους θ_g . Ορίζεται, επίσης, ένα MLP $D(x; \theta_d)$ με έξοδο ένα μονοδιάστατο διάνυσμα. Το $D(x)$ αντιπροσωπεύει την πιθανότητα το x να προήλθε από την κατανομή των πραγματικών δεδομένων αντί για την κατανομή p_g του γεννήτορα.

Στο βασικό GAN εκπαιδεύεται ο διευκρινιστής D με σκοπό να μεγιστοποιήσει την πιθανότητα ανάθεσης της σωστής ετικέτας τόσο για τα δείγματα που προήλθαν από την κατανομή των πραγματικών δεδομένων, όσο και για τα δείγματα που προήλθαν από τον γεννήτορα G . Ταυτόχρονα, εκπαιδεύεται ο γεννήτορας με σκοπό την ελαχιστοποίηση του $\log(1 - D(G(z)))$, δηλαδή της πιθανότητας να αναγνωρίσει ο διευκρινιστής τα δείγματα που παράγει ο γεννήτορας ως μη πραγματικά. Συνεπώς,

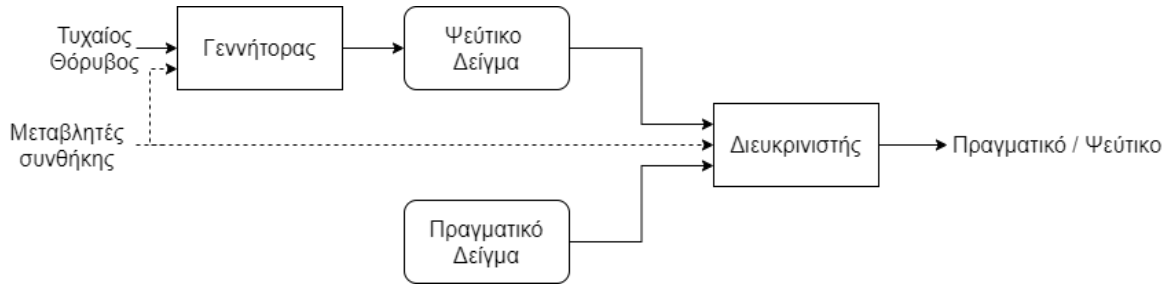
ο γεννήτορας G και ο διευκρινιστής D παίζουν το ακόλουθο minimax παιχνίδι με τη συνάρτηση τιμής $V(G, D)$, όπως περιγράφεται στην Εξίσωση 2.1.

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (2.1)$$

Σκοπός της εκπαίδευσης του GAN είναι το παραπάνω παιχνίδι να καταλήξει σε *minimax ισορροπία* (minimax equilibrium). Στην κατάσταση αυτή, οι παίκτες εξουδετερώνουν ο ένας τον άλλον, με αποτέλεσμα κανένας από τους δύο να μην μπορεί να βελτιώσει την επίδοσή του, δίχως να αλλάξει τις παραμέτρους του παιχνιδιού. Θεωρητικά, στην περίπτωση των GANs, όταν το σύστημα βρεθεί στην κατάσταση αυτή, ο γεννήτορας θα παράγει δείγματα τα οποία δε θα μπορούν να διακριθούν από τα πραγματικά και ο διευκρινιστής θα δίνει συνεχώς έξοδο 0.5, καθώς δε θα είναι σε θέση να αναγνωρίσει από ποια κατανομή προέρχεται το κάθε δείγμα.

2.1.2 GAN υπό συνθήκη

Το Σχήμα 2.2 απεικονίζει διαγραμματικά το *GAN υπό συνθήκη* (conditional GAN - CGAN) [Mirz14]. Ο γεννήτορας παίρνει ως είσοδο τον τυχαίο θόρυβο και τις μεταβλητές συνθήκης και παράγει συνθετικά/ψεύτικα δείγματα. Ο διευκρινιστής παίρνει ως είσοδο τα δείγματα και τις μεταβλητές συνθήκης και αναγνωρίζει αν προέρχονται από τον γεννήτορα ή τα πραγματικά δεδομένα.



Σχήμα 2.2: Διαγραμματική απεικόνιση του CGAN.

Το CGAN αποτελεί μία προέκταση του βασικού GAN στην οποία τόσο ο γεννήτορας όσο και ο διευκρινιστής λαμβάνουν ως επιπλέον είσοδο *μεταβλητές συνθήκης* (conditioning variables, c). Αυτή η παραλλαγή επιτρέπει στον γεννήτορα να παράξει συνθετικά δείγματα βασισμένα στην εκάστοτε συνθήκη που θα λάβει ως είσοδο. Η συνθήκη αυτή μπορεί να αποτελεί κείμενο (το οποίο έχει προβληθεί στον διανυσματικό χώρο), εικόνα, ετικέτες κλάσης ή οποιαδήποτε άλλη πρόσθετη πληροφορία.

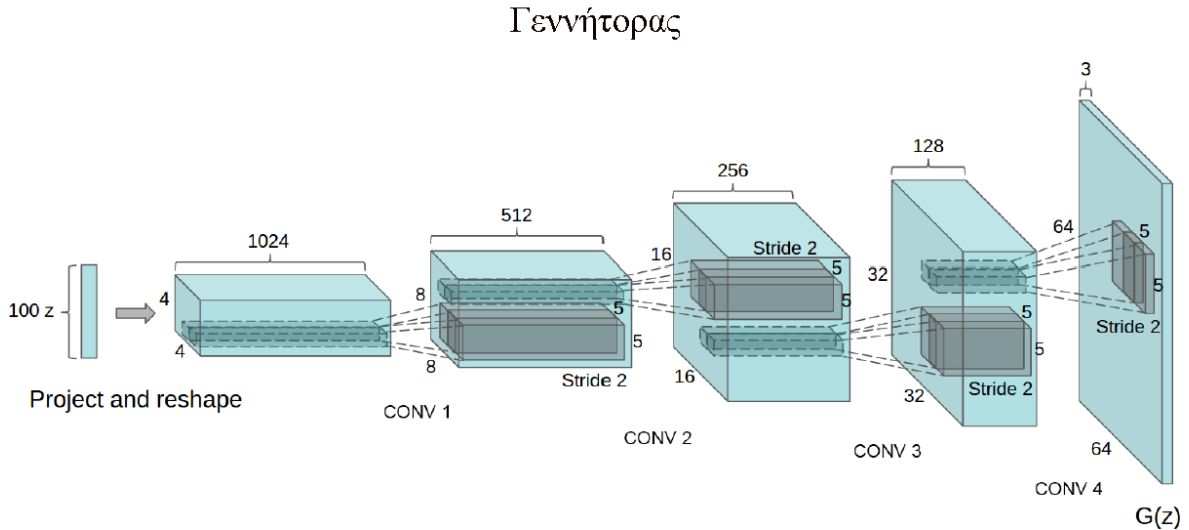
Στην πλευρά του γεννήτορα, οι μεταβλητές συνθήκης c συνήθως συνενώνονται με τον τυχαίο θόρυβο στο πρώτο επίπεδο του δικτύου. Η ίδια διαδικασία ακολουθείται και στην πλευρά του διευκρινιστή, αλλά αυτή τη φορά η συνένωση πραγματοποιείται ανάμεσα στις μεταβλητές συνθήκης και στο δείγμα που λαμβάνει ως είσοδο ο διευκρινιστής.

Ο στόχος της εκπαίδευσης του CGAN είναι ο ίδιος με εκείνον της εκπαίδευσης του βασικού GAN. Η μοναδική διαφορά εντοπίζεται στις εισόδους των MLPs που λαμβάνουν πλέον ως είσοδο και τις μεταβλητές συνθήκης c . Συνεπώς το minimax παιχνίδι που παίζουν ο γεννήτορας και ο διευκρινιστής στο CGAN έχει τη συνάρτηση τιμής $V(G, D)$, όπως περιγράφεται στην Εξίσωση 2.2.

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x|c)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z|c)))] \quad (2.2)$$

2.1.3 Βαθύ Συνελικτικό GAN

Το βαθύ συνελικτικό GAN (deep convolutional GAN - DCGAN) [Radf15], ο γεννήτορας του οποίου απεικονίζεται στο Σχήμα 2.3, αποτελεί μία από τις σημαντικότερες βελτιώσεις στο μοντέλο του GAN. Στο DCGAN τα MLPs που χρησιμοποιούνταν στο αρχικό μοντέλο για τον γεννήτορα και τον διευκρινιστή έχουν αντικατασταθεί από *Βαθιά Συνελικτικά Δίκτυα* (deep convolutional networks), οι αρχιτεκτονικές των οποίων είναι αντίστροφες μεταξύ τους. Στο Παράρτημα Α παρουσιάζονται περισσότερες πληροφορίες σχετικά με τα βαθιά συνελικτικά δίκτυα.



Σχήμα 2.3: Ο γεννήτορας του DCGAN όπως απεικονίζεται στο [Radf15] για το LSUN σύνολο δεδομένων. Αποτελείται από ένα Βαθύ Συνελικτικό Δίκτυο με τέσσερα κρυφά συνελικτικά επίπεδα. Λαμβάνει ως είσοδο θόρυβο z και παράγει εικόνα διαστάσεων 64x64 στην έξοδο.

Ολόκληρο το δίκτυο εκπαιδεύεται με *οπίσθια διάδοση σφάλματος* (backpropagation of error). Σε αντιστοιχία με το βασικό GAN, στο DCGAN ο διευκρινιστής D εκπαιδεύεται για να μεγιστοποιήσει την πιθανότητα σωστής αναγνώρισης της προέλευσης του δείγματος, δηλαδή αναγνωρίζει αν προέρχεται από την πραγματική κατανομή των δεδομένων ή τα συνθετικά δείγματα του γεννήτορα, ενώ ο γεννήτορας G εκπαιδεύεται για να «εξαπατήσει» τον διευκρινιστή. Συνεπώς εμπλέκονται πάλι στο ίδιο minimax παιχνίδι με συνάρτηση τιμής $V(G, D)$, όπως περιγράφεται στην Εξίσωση 2.1.

2.2 Συναρτήσεις σφάλματος

2.2.1 Τετραγωνικό σφάλμα

Το *τετραγωνικό σφάλμα* (quadratic loss) υπολογίζεται από τον τύπο της Εξίσωσης 2.3

$$QuadraticLoss(x, y) = \frac{1}{n} \sum_{i=1}^n |y_{pred,i} - y_i|^2 \quad (2.3)$$

όπου το y_{pred} αντιπροσωπεύει τις n προβλέψεις του δικτύου και το y αντιπροσωπεύει τις πραγματικές κλάσεις των δεδομένων της εισόδου.

2.2.2 Σφάλμα Διασταυρούμενης Εντροπίας

Η *διασταυρούμενη εντροπία* (cross entropy) αποτελεί μία ευρέως χρησιμοποιούμενη εναλλακτική του τετραγωνικού σφάλματος. Χρησιμοποιείται όταν η έξοδος του δικτύου αναπαριστά κατανομή πιθανότητας. Υπολογίζεται από τον τύπο της Εξίσωσης 2.4

$$CrossEntropyLoss(x, y) = - \sum_{i=1}^n y_i * \log \frac{e^{y_{pred,i}}}{\sum_{j=1}^n e^{y_{pred,j}}} \quad (2.4)$$

όπου y_{pred} είναι ένα διάνυσμα από n προβλέψεις του νευρωνικού δικτύου και y είναι ένα δυαδικό διάνυσμα γεμάτο με μηδενικά, εκτός από άσσους που αντιπροσωπεύουν τις πραγματικές κλάσεις των δεδομένων εισόδου.

2.3 Μέθοδοι εκπαίδευσης

2.3.1 Αλγόριθμος Στοχαστικής Κατάβασης Κλίσης

Ο αλγόριθμος στοχαστικής κατάβασης κλίσης (stochastic gradient descent - SGD) αποτελεί μια παραλλαγή του απλού αλγορίθμου κατάβασης κλίσης (gradient descent - GD). Στη μέθοδο αυτή τα βάρη του δικτύου ενημερώνονται για κάθε δείγμα εισόδου ξεχωριστά σύμφωνα με τον τύπο της Εξίσωσης 2.5

$$\theta = \theta - \eta \cdot \nabla_{\theta} J(\theta, x_i, y_i) \quad (2.5)$$

όπου x_i το δείγμα εισόδου, y_i η αντίστοιχη ετικέτα, θ τα βάρη του δικτύου και η ο ρυθμός μάθησης. Στην πράξη ο αλγόριθμος SGD συχνά χρησιμοποιείται σε *μικρές δέσμες δεδομένων* (mini-batches).

2.3.2 Οπίσθια διάδοση του σφάλματος

Η *οπίσθια διάδοση του σφάλματος* (backpropagation of error - BP) αποτελεί μία συχνά χρησιμοποιούμενη μέθοδο εκπαίδευσης νευρωνικών δικτύων και χρησιμοποιείται συνήθως σε συνδυασμό με μία μέθοδο βελτιστοποίησης όπως SGD ή Adam [King14]. Στην οπίσθια διάδοση υπολογίζουμε πρώτα την πρόβλεψη και το αντίστοιχο σφάλμα για κάθε δείγμα εισόδου. Αθροίζουμε όλα τα σφάλματα για να υπολογίσουμε το τελικό σφάλμα και στη συνέχεια διαδίδουμε το τελικό σφάλμα προς τα πίσω στο δίκτυο υπολογίζοντας τις μερικές παραγώγους $\frac{\partial J}{\partial \theta}$ και $\frac{\partial J}{\partial \theta_0}$ για τη συνάρτηση σφάλματος J και όλα τα βάρη θ και τις αποκλίσεις θ_0 στο νευρωνικό δίκτυο.

2.4 Συνήθη προβλήματα των GANs

Τα δύο συνηθέστερα προβλήματα σύγκλισης κατά την εκπαίδευση των GANs είναι η αδυναμία βελτίωσης των παραγόμενων δειγμάτων του γεννήτορα καθώς και η *κατάρρευση συστήματος* (mode collapse) [Barn18].

2.4.1 Αδυναμία βελτίωσης παραγόμενων δειγμάτων

Υπάρχουν δύο λόγοι που μπορεί να οδηγήσουν σε αδυναμία βελτίωσης ενός GAN ως προς την ποιότητα των δειγμάτων που παράγει. Ο πρώτος αφορά την εκάστοτε μέθοδο εκπαίδευσης, η οποία πιθανόν να αποτρέψει το δίκτυο να προσεγγίσει τη βέλτιστη τιμή (π.χ. παραμονή σε τοπικό ελάχιστο).

Ο δεύτερος λόγος, αφορά ένα γνωστό πρόβλημα της μηχανικής μάθησης, τις *εκλειπόμενες κλίσεις* (vanishing gradients). Στην περίπτωση αυτή, οι κλίσεις (μερικές παράγωγοι της Ενότητας 2.3) που υπολογίζονται για την εκπαίδευση του γεννήτορα συρρικνώνονται σε σημείο μηδενισμού, με αποτέλεσμα να μη μπορούν να συνεισφέρουν στη μεταβολή των βαρών του δικτύου. Αυτό το φαινόμενο είναι πιθανό να προκληθεί από την επικράτηση ενός πολύ καλού διευκρινιστή, που δεν επιτρέπει στον γεννήτορα να βελτιωθεί [Arjo17].

2.4.2 Κατάρρευση Συστήματος

Συνήθως, ο σκοπός της εκπαίδευσης ενός GAN είναι η δυνατότητα σύνθεσης ποικιλόμορφων δειγμάτων. Πιο συγκεκριμένα, το επιθυμητό αποτέλεσμα είναι για κάθε τυχαία είσοδο, να παράγεται ένα διαφορετικό δείγμα ως έξοδος. Ωστόσο, ένα σύνηθες πρόβλημα της εκπαίδευσης των GANs είναι η προσκόλληση του γεννήτορα σε μία συγκεκριμένη μορφή/κατανομή δείγματος επειδή το αποτέλεσμα στην έξοδο είναι αληθοφανές και καταφέρνει να ξεγελάσει τον διευκρινιστή. Αντίστοιχα, ο διευκρινιστής εκπαιδεύεται πλέον ώστε να αναγνωρίζει μόνο τη συγκεκριμένη μορφή δειγμάτων. Σε κάθε επόμενη επανάληψη, ο γεννήτορας υπερπροσαρμόζεται πάνω στην εκάστοτε κατάσταση του διευκρινιστή, με τα δείγματα να παραμένουν όμοια μεταξύ τους. Αυτή η μορφή αποτυχημένης εκπαίδευσης ονομάζεται *κατάρρευση συστήματος* (mode collapse). Ως καταρρεύσεις συστήματος αναφέρονται επίσης και περιπτώσεις στις οποίες ο γεννήτορας ενός GAN παράγει συνθετικές εικόνες χωρίς κάποια οπτική ουσία (nonsensical images), τις οποίες αδυνατεί να αναγνωρίσει ως συνθετικές ο διευκρινιστής.

2.5 Τεχνικές βελτιστοποίησης εκπαίδευσης

2.5.1 Τροποποιημένη συνάρτηση σφάλματος

Όπως είδαμε στην Ενότητα 2.1.1 για το βασικό GAN, ο γεννήτορας εκπαιδεύεται με σκοπό να ελαχιστοποιήσει την πιθανότητα να αναγνωρίσει ο διευκρινιστής τα δείγματα που παράγει ως μη πραγματικά, δηλαδή η συνάρτηση σφάλματος για τον γεννήτορα είναι $\min(\log(1 - D))$. Ωστόσο, στην πράξη είναι καλύτερη η χρησιμοποίηση της συνάρτησης $\max(\log(D))$, καθώς στην πρώτη εμφανίζεται συχνά το πρόβλημα των εκλειπόμενων κλίσεων [Good14].

2.5.2 Διαχωρισμός δεσμών δεδομένων

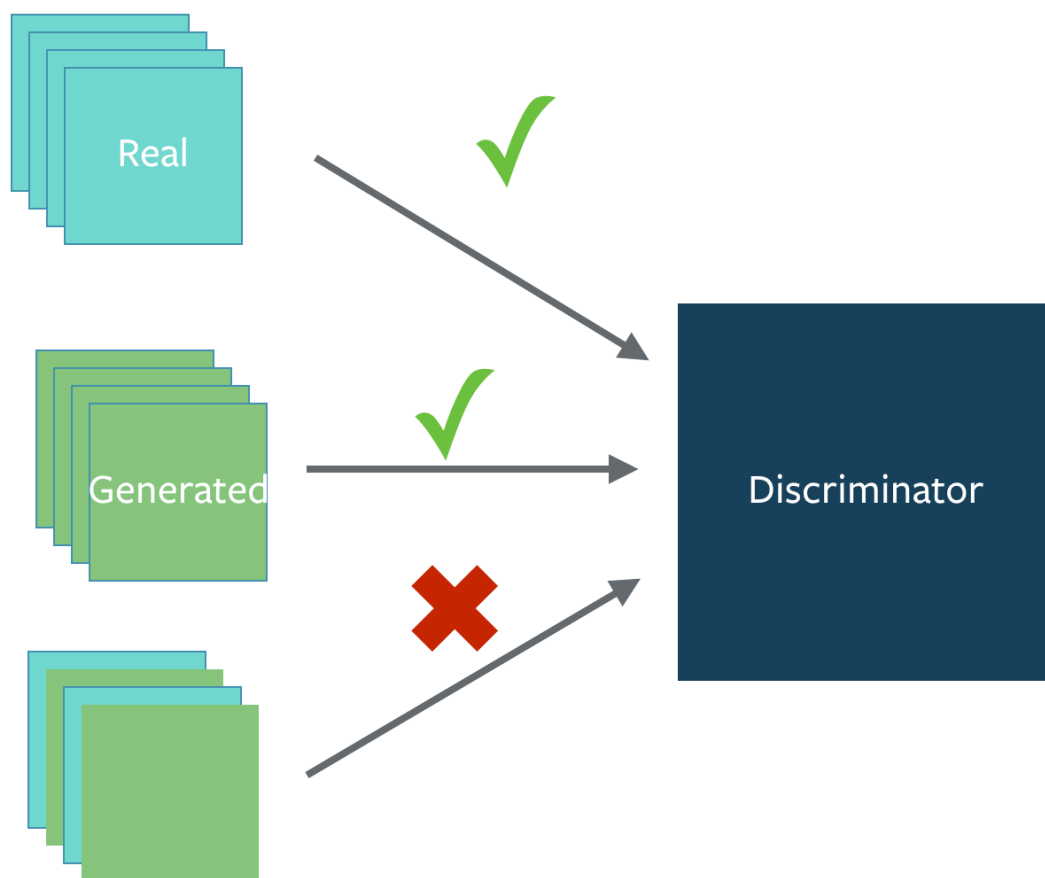
Έχει αποδειχθεί ότι η εκπαίδευση του διευκρινιστή με ξεχωριστές δέσμες δεδομένων από πραγματικά και συνθετικά δείγματα αντίστοιχα, ευνοεί την εκπαίδευση του GAN, σε αντίθεση με τον συνδυασμό των δύο παραπάνω κατανομών σε ενιαία δέσμη δεδομένων. Η μέθοδος αυτή, δε, είναι ιδιαίτερα χρήσιμη όταν χρησιμοποιείται *κανονικοποίηση δέσμης δεδομένων* (batch normalization) λόγω των αποκλίσεων των πραγματικών και των συνθετικών κατανομών των δειγμάτων, ειδικά στα πρώτα βήματα της εκπαίδευσης του δικτύου (Σχήμα 2.4).

2.5.3 Εξομάλυνση ετικετών εκπαίδευσης

Είναι σύνηθες να χρησιμοποιείται η ετικέτα κλάσης 1 για να αντιπροσωπεύει τα δείγματα της πραγματικής κατανομής και η ετικέτα κλάσης 0 για τα συνθετικά δείγματα του γεννήτορα κατά την εκπαίδευση του διευκρινιστή. Στην πράξη, ωστόσο, έχει εντοπιστεί ότι βοηθάει στην εκπαίδευση του δικτύου η εισαγωγή ενός παράγοντα κανονικοποίησης, μία τεχνική που ονομάζεται *εξομάλυνση ετικετών εκπαίδευσης* (label smoothing). Πιο συγκεκριμένα, στην τεχνική αυτή οι ετικέτες δεν έχουν τις αυστηρές τιμές 0 και 1, αλλά παίρνουν τιμές ομοιόμορφα που κυμαίνονται στα διαστήματα $[0.0, 0.3]$ και $[0.7, 1.2]$ αντίστοιχα [Mull19].

2.5.4 Προσθήκη θορύβου στην είσοδο του διευκρινιστή

Μία ακόμα τεχνική που έχει αποδειχθεί ότι βοηθάει στην εκπαίδευση των GANs είναι η προσθήκη θορύβου, ελαττωμένου με τον χρόνο, στα δείγματα εισόδου του διευκρινιστή. Με τη μέθοδο αυτή αποτρέπεται το πρόβλημα της επικράτησης του διευκρινιστή ως προς τον γεννήτορα, δεδομένου ότι τα πραγματικά δείγματα που δέχεται ως είσοδο ο πρώτος, στην αρχή της εκπαίδευσης, περιέχουν και αυτά θόρυβο [Arjo17].



Σχήμα 2.4: Διαχωρισμός δεσμών δεδομένων (mini batches) σε δέσμες από πραγματικά και συνθετικά δείγματα αντίστοιχα. (Πηγή: <https://github.com/soumith/ganhacks>)

Κεφάλαιο 3

Σύνθεση εικόνας από κείμενο με χρήση GANs

Ένα από τα πιο απαιτητικά προβλήματα στον τομέα της *όρασης υπολογιστών* (computer vision - CV), με επεκτάσεις και στον τομέα της *επεξεργασίας φυσικής γλώσσας* (natural language processing - NLP), αποτελεί η σύνθεση εικόνας από κείμενο. Όπως αναφέρθηκε και στην εισαγωγή (Κεφάλαιο 1, το ζητούμενο είναι, δοσμένης μιας λεκτικής περιγραφής, η σύνθεση εικόνας που να αντιστοιχεί στην περιγραφή αυτή. Πολύ εύκολα μπορεί να παρατηρήσει κανείς ότι πρόκειται για το αντίστροφο του γνωστού προβλήματος *δημιουργίας λεζάντας για εικόνα* (image captioning). Στα δύο αυτά ζητούμενα εντοπίζονται κάποιες κοινές δυσκολίες, με βασικότερη εξ αυτών την ποικιλομορφία των παραγόμενων αποτελεσμάτων. Ωστόσο, η δυσκολία στην αντιμετώπιση τους δεν είναι η ίδια. Η διαδοχικότητα μεταξύ των λέξεων στη φυσική γλώσσα καθιστά την παραγωγή λεζάντας, δηλαδή κειμένου, πιο εύκολη διαδικασία από τη σύνθεση εικόνας.

Παρά τις δυσκολίες αυτές, έχει σημειωθεί σημαντική πρόοδος και στους δύο τομείς, με τη συμβολή των GANs, και ιδιαίτερα των CGAN [Mirz14] και DCGAN [Radf15] να είναι καθοριστική. Ο Scott Reed και οι συνεργάτες του στο [Reed16b], πρότειναν μία απλή και αποτελεσματική αρχιτεκτονική για την αντιμετώπιση του προβλήματος της σύνθεσης εικόνας από κείμενο, χρησιμοποιώντας ένα βαθύ συνελκτικό γεννητικό ανταγωνιστικό δίκτυο υπό συνθήκη. Οι συνθετικές εικόνες που παράγει το δίκτυό τους είναι πειστικές, έχοντας όμως χαμηλή ανάλυση (64x64) και ελάχιστες λεπτομέρειες. Η μετάβαση σε υψηλότερη ανάλυση (π.χ. 256x256) καθιστά το πρόβλημα περισσότερο δύσκολο στην αντιμετώπισή του. Ο Han Zhang και οι συνεργάτες του προσέγγισαν το πρόβλημα με μία πρωτότυπη αρχιτεκτονική *στοιβαγμένων GANs* (stacked GAN - StackGAN) [Zhan16]. Μέτα από δύο διαδοχικά στάδια εκπαίδευσης, κατάφεραν να παρουσιάσουν αποτελέσματα εικόνων υψηλής ανάλυσης (256x256) με περισσότερη λεπτομέρεια και μεγαλύτερη αντιστοιχία μεταξύ εικόνας και λεκτικής περιγραφής. Μάλιστα, στο [Zhan17] εξέλιξαν την αρχιτεκτονική τους, με ταυτόχρονη εκπαίδευση τριών, πλέον, στοιβαγμένων GANs σε δένδρική δομή. Οι εικόνες που παράγονται από τη δεύτερη γενιά του δικτύου τους (StackGAN++) έχουν αντίστοιχα υψηλή ανάλυση (256x256), με δυνατότητα μάλιστα επέκτασης σε υψηλότερες αναλύσεις. Τα αποτελέσματά τους είναι εντυπωσιακά λόγω της αληθοφάνειας και της λεπτομέρειας που παρουσιάζουν.

Επιπρόσθετες αρχιτεκτονικές που έχουν προταθεί για το συγκεκριμένο πρόβλημα και παρουσιάζουν πολύ καλά αποτελέσματα είναι οι ακόλουθες:

- **AttnGAN** [Xu18]: Έχει παρόμοια αρχιτεκτονική με αυτή του StackGAN++, στην οποία έχει προστεθεί *μηχανισμός εστίασης* (attention mechanism) σε επίπεδο λέξεων της περιγραφής ή συγκεκριμένη περιοχή της παραγόμενης εικόνας.
- **Obj-GAN** [Li19a]: Το μοντέλο αυτό επικεντρώνεται στη σύνθεση εικόνων από λεκτικές περιγραφές που αφορούν περίπλοκες σκηνές. Δίνεται έμφαση στα αντικείμενα της περιγραφής (object driven) και ακολουθεί και αυτό μία εξελικτική πορεία κατά τη σύνθεση. Αρχικά, σχεδιάζονται κουτιά για την τοποθεσία των αντικειμένων, στη συνέχεια σχεδιάζονται σε αυτά οι φιγούρες των αντικειμένων σχηματικά (χωρίς λεπτομέρειες) και τέλος γίνεται η σύνθεση της εικόνας προσθέτοντας στιλιστικές λεπτομέρειες και φόντο.
- **MirrorGAN** [Qiao19]: Το μοντέλο αυτό προσπαθεί να συνδυάσει τα δύο ζητούμενα που αναφέρθηκαν στην αρχή της ενότητας. Αποτελείται από τρία διαφορετικά τμήματα. Στο πρώτο

τμήμα παράγονται διανύσματα αναπαράστασης της περιγραφής, με τοπικά και καθολικά χαρακτηριστικά. Στο δεύτερο τμήμα πραγματοποιείται η σύνθεση της εικόνας με χρήση στοιβαγμένων GANs και μηχανισμών εστίασης, ενώ στο τρίτο και τελευταίο τμήμα υπάρχει ένα μοντέλο αναπαραγωγής της λεκτικής περιγραφής από την παραγόμενη εικόνα του δεύτερου τμήματος (image captioning). Έπειτα προστίθεται ένα σφάλμα για την ομοιότητα της παραγόμενης λεκτικής περιγραφής με την αρχική, επιτυγχάνοντας έτσι καλύτερη αντιστοιχία μεταξύ σύνθετης εικόνας και περιγραφής.

- **DM-GAN** [Zhu19]: Ένα σύνθετο πρόβλημα των παραπάνω μοντέλων είναι ότι η ποιότητα της παραγόμενης εικόνας βασίζεται σημαντικά στην εικόνα του πρώτου σταδίου της αρχιτεκτονικής. Συνεπώς, αν η εικόνα του πρώτου σταδίου δεν είναι καλή, η ποιότητα της σύνθετης εικόνας του τελευταίου σταδίου δε θα είναι ικανοποιητική. Το Dynamic Memory GAN επιλύει το συγκεκριμένο πρόβλημα με τη χρήση *πυλών μνήμης* (memory gates) οι οποίες δέχονται ως είσοδο τη σύνθετη εικόνα του εκάστοτε σταδίου και τη λεκτική περιγραφή, προκειμένου να εντοπίσουν και να διορθώσουν τμήματα τα οποία δεν έχουν αποτυπωθεί σωστά.

3.1 GAN-CLS

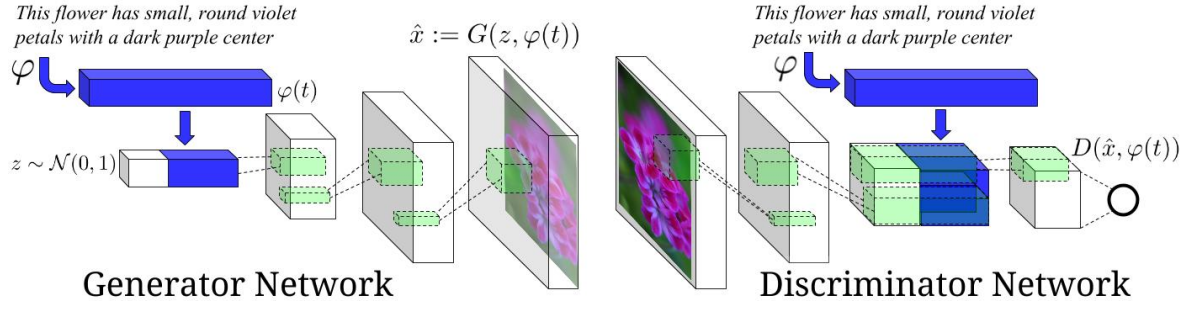
Η πρώτη προσέγγιση στο πρόβλημα της σύνθεσης εικόνας από κείμενο προτάθηκε από τον Scott Reed και τους συνεργάτες του, με ενθαρρυντικά αποτελέσματα [Reed16b]. Στην προσέγγισή τους, χώρισαν το πρόβλημα σε δύο επιμέρους υποπροβλήματα. Το πρώτο από αυτά ήταν η δημιουργία αναπαράστασης χαρακτηριστικών κειμένου που να αποδίδει με τον βέλτιστο τρόπο τις οπτικές λεπτομέρειες της περιγραφής. Το δεύτερο υποπρόβλημα ήταν η χρησιμοποίηση των αναπαραστάσεων αυτών για τη σύνθεση ρεαλιστικών εικόνων με τη χρήση ενός GAN.

3.1.1 Διανυσματική αναπαράσταση κειμένου

Οι λεκτικές περιγραφές πρέπει να μετατραπούν σε διανύσματα χαρακτηριστικών προκειμένου να χρησιμοποιηθούν ως είσοδοι σε οποιοδήποτε μοντέλο. Στην περίπτωση αυτή οι συγγραφείς χρησιμοποίησαν *διανυσματικές αναπαραστάσεις κειμένου* (embeddings) που παρήχθησαν από τον κωδικοποιητή char-CNN-RNN. Στον κωδικοποιητή αυτό, οι εικόνες και οι λεκτικές περιγραφές αντιστοιχίζονται σε ένα κοινό διανυσματικό χώρο με τέτοιο τρόπο ώστε τα ζεύγη που ταιριάζουν μεταξύ τους να βρίσκονται πιο κοντά. Για την αντιστοίχιση αυτή χρησιμοποιείται ένα *Συνελικτικό Νευρωνικό Δίκτυο* (CNN) για την επεξεργασία των εικόνων και ένα υβριδικό Συνελικτικό-Ανατροφοδοτούμενο Νευρωνικό Δίκτυο (CNN-RNN) για την επεξεργασία των περιγραφών [Reed16a].

3.1.2 Αρχιτεκτονική δικτύου

Στο Σχήμα 3.1 παρουσιάζεται μία επιφανειακή απεικόνιση της αρχιτεκτονικής του GAN υπό συνθήκη κειμένου που χρησιμοποιείται. Όπως είδαμε στην Ενότητα 2.1.2 η σύνθεση δειγμάτων από τον γεννήτορα σε ένα CGAN, εξαρτάται άμεσα από τις μεταβλητές συνθήκης c που λαμβάνει στην είσοδό του. Στο GAN-CLS, τα διανύσματα αναπαράστασης που παράγονται από το char-CNN-RNN χρησιμοποιούνται ως μεταβλητές συνθήκης. Πιο συγκεκριμένα, παρατηρούμε ότι το κωδικοποιημένο κείμενο $\phi(t)$ συνενώνεται τόσο με το διάνυσμα θορύβου z για να δοθεί ως είσοδος στον γεννήτορα, όσο και με χάρτες χαρακτηριστικών του διεκρινιστή, προκειμένου να χρησιμοποιηθεί σε επόμενα στάδια συνελικτικής επεξεργασίας.



Σχήμα 3.1: Επιφανειακή (high-level) απεικόνιση της αρχιτεκτονικής του δικτύου GAN-INT-CLS. (Πηγή: [Reed16b])

Στο GAN-CLS χρησιμοποιούνται βαθιά συνελκτικά δίκτυα τόσο για τον γεννήτορα όσο και για τον διευκρινιστή. Η αρχιτεκτονική τους είναι όμοια με εκείνη του DCGAN που παρουσιάστηκε στην Ενότητα 2.1.3 [Radf15].

Γεννήτορας

Στον γεννήτορα αρχικά δειγματοληπτείται θόρυβος $z \in \mathbb{R} \sim \mathcal{N}(0, 1)$ και κωδικοποιείται η λεκτική περιγραφή t χρησιμοποιώντας τον κωδικοποιητή ϕ . Στη συνέχεια, η διανυσματική αναπαράσταση της περιγραφής $\phi(t)$ συμπιέζεται σε μικρότερες διαστάσεις (στην πράξη 128) με τη χρήση ενός πλήρους συνδεδεμένου επιπέδου (fully connected layer) με συνάρτηση ενεργοποίησης διαρρέουσας διορθωμένης γραμμικής μονάδας (Leaky ReLu) και συνενώνεται με το διάνυσμα θορύβου z . Το συνενωμένο διάνυσμα, προβάλλεται γραμμικά και περνά από μία σειρά από αντιμεταθετικά και μη συνελκτικά επίπεδα έως ότου να καταλήξει σε ένα τανυστή με διαστάσεις $64 \times 64 \times 3$. Οι τιμές του τανυστή αυτού περνούν μέσα από μια συνάρτηση ενεργοποίησης υπερβολικής εφαπτομένης προκειμένου να απεικονιστούν στο πεδίο τιμών $[-1, 1]$.

Διευκρινιστής

Στον διευκρινιστή η εικόνα που λαμβάνεται ως είσοδος περνά από μία σειρά από συνελκτικά επίπεδα προκειμένου να μειωθεί η χωρική της διάσταση. Μόλις καταλήξει σε διαστάσεις 4×4 συνενώνεται με τη συμπιεσμένη διανυσματική αναπαράσταση της περιγραφής $\phi(t)$, η οποία παράχθηκε χρησιμοποιώντας ένα διαφορετικό πλήρως συνδεδεμένο επίπεδο με συνάρτηση ενεργοποίησης Leaky ReLu, σε σχέση με αυτό του γεννήτορα. Πιο συγκεκριμένα, η αναπαράσταση $\phi(t)$ επαναλαμβάνεται χωρικά και ενώνεται με τον χάρτη χαρακτηριστικών του σημείου εκείνου στη διάσταση του βάθους. Ο συνενωμένος τανυστής περνά από επιπλέον συνελκτικά επίπεδα έως ότου καταλήξει σε μονόμετρο μέγεθος. Στη συνέχεια, η τιμή αυτή περνά μέσα από σιγμοειδή συνάρτηση ενεργοποίησης για να απεικονιστεί στο πεδίο τιμών $[0, 1]$, προκειμένου να αναπαριστά έγκυρη τιμή πιθανότητας.

3.1.3 Τροποποιημένη συνάρτηση σφάλματος

Ο διευκρινιστής του GAN-CLS διαφοροποιείται από εκείνον του απλού CGAN ως προς το γεγονός ότι πρέπει να αναγνωρίζει δύο κατηγορίες λαθών: (i) τα λάθη που οφείλονται σε μη ρεαλιστικές εικόνες (ανεξάρτητα της περιγραφής στη συνθήκη) και (ii) τα λάθη που οφείλονται σε ρεαλιστικές εικόνες που δεν ταιριάζουν με την περιγραφή που δίνεται ως συνθήκη. Προκειμένου λοιπόν ο διευκρινιστής του GAN-CLS να είναι σε θέση να αναγνωρίσει την αντιστοιχία μεταξύ κειμένου-εικόνας, η συνάρτηση σφάλματός της Εξίσωσης 2.2 τροποποιείται όπως παρακάτω (Εξίσωση 3.1)

$$L_D = \log(D(x, h)) + \frac{\log(1 - D(x, \hat{h})) + \log(1 - D(G(z, h), h))}{2} \quad (3.1)$$

όπου $x \sim p_{data}$ έχει ληφθεί από την κατανομή των πραγματικών δεδομένων, $h \sim p_{matching_embeddings}$ από την κατανομή των κωδικοποιημένων διανυσμάτων των περιγραφών που ταιριάζουν με την εικόνα, $\hat{h} \sim p_{mismatching_embeddings}$ από την κατανομή των κωδικοποιημένων διανυσμάτων των περιγραφών που δεν ταιριάζουν με την εικόνα και $z \sim p_z$ από την κατανομή του θορύβου.

3.1.4 Παραλλαγή GAN-INT-CLS

Έχει αποδειχθεί ότι τα βαθιά δίκτυα έχουν τη δυνατότητα να αναγνωρίσουν και να μάθουν αναπαραστάσεις για ζεύγη εισόδων των οποίων οι διανυσματικές αναπαραστάσεις βρίσκονται κοντά, ως προς την πολλαπλότητα των δεδομένων. Στο GAN-INT-CLS [Reed16b], ωθούμενοι από αυτή την ιδιότητα, παράγονται πρόσθετες διανυσματικές αναπαραστάσεις κειμένων από την παρεμβολή μεταξύ αναπαραστάσεων του συνόλου εκπαίδευσης. Στην πράξη αυτό επιτυγχάνεται προσθέτοντας έναν επιπλέον όρο στη συνάρτηση βελτιστοποίησης που καλείται να ελαχιστοποιήσει ο γεννήτορας, όπως απεικονίζεται στην Εξίσωση 3.2

$$\mathbb{E}_{t_1, t_2 \sim p_{data}} [\log(1 - D(G(z, \beta t_1 + (1 - \beta)t_2)))] \quad (3.2)$$

όπου $z \sim p_z$ έχει ληφθεί από την κατανομή του θορύβου και το β παρεμβάλλεται μεταξύ των διανυσματικών αναπαραστάσεων κειμένου t_1 και t_2 . Στην πράξη χρησιμοποιείται η τιμή $\beta = 0.5$, λόγω καλών αποτελεσμάτων.

Οι αναπαραστάσεις t_1 και t_2 είναι πιθανό να προέρχονται από διαφορετικές εικόνες ακόμα και κατηγορίες. Ωστόσο, επειδή ο διευκρινιστής μαθαίνει να αναγνωρίζει και την αντιστοιχία μεταξύ κειμένου-εικόνας, ο γεννήτορας εκπαιδεύεται να καλύπτει τα κενά μεταξύ των αναπαραστάσεων που προκύπτουν από την πολλαπλότητα των δεδομένων.

3.2 StackGAN

Είναι γνωστό ότι η εκπαίδευση των GANs είναι δύσκολη διαδικασία καθώς είναι ασταθής και συχνά ευαίσθητη στην επιλογή των υπερπαραμέτρων. Τα προβλήματα αυτά είναι περισσότερο έντονα όταν η έξοδος του GAN είναι εικόνα υψηλής ανάλυσης (πχ 256x256). Επιπλέον, όπως αναφέρθηκε στην Ενότητα 2.4.2, ένα σύνθετο φαινόμενο κατά την εκπαίδευση των GANs είναι η κατάρρευση συστήματος, κατά την οποία πολλά από τα παραγόμενα δείγματα του γεννήτορα είναι όμοια μεταξύ τους.

Λόγω των προβλημάτων αυτών, οι αρχικές προσεγγίσεις για σύνθεση εικόνας από κείμενο περιορίστηκαν στην παραγωγή εικόνων χαμηλής ανάλυσης. Για παράδειγμα στο GAN-CLS [Reed16b] η ανάλυση των παραγόμενων εικόνων είναι 64x64. Μάλιστα, η μετάβαση σε υψηλότερη ανάλυση (256x256) οδηγεί σε αποτελέσματα εικόνων που ανταποκρίνονται ελάχιστα στην λεκτική περιγραφή, δίχως να περιέχουν λεπτομέρειες για το κεντρικό αντικείμενο που απεικονίζουν. Αυτό οφείλεται στο γεγονός ότι η αρχιτεκτονική του GAN-CLS παράγει απευθείας την εικόνα σε υψηλή ανάλυση δίχως να υπάρχει περιθώριο βελτίωσης.

Στο StackGAN [Zhan16] το συγκεκριμένο πρόβλημα επιλύεται διαχωρίζοντας την παραγωγή της εικόνας υψηλής ανάλυσης σε δύο στάδια, χρησιμοποιώντας δύο στοιβαγμένα GANs. Στο πρώτο στάδιο, το Stage-I GAN σχεδιάζει μία αρχική εικόνα χαμηλής ανάλυσης (64x64) βασισμένη στη λεκτική περιγραφή, σε αντιστοιχία με το GAN-CLS. Το Stage-II GAN, στη συνέχεια, παίρνει ως είσοδο τα αποτελέσματα του Stage-I GAN και τις λεκτικές περιγραφές και παράγει εικόνα υψηλότερης ανάλυσης (256x256) με περισσότερες λεπτομέρειες στην απεικόνιση και μεγαλύτερη αντιστοιχία ως προς τις περιγραφές.

3.2.1 Επαύξηση Διανυσματικής Αναπαράστασης

Εκτός από την παραγωγή εικόνας σε πολλαπλές αναλύσεις, οι συγγραφείς προτείνουν και μία πρωτότυπη τεχνική επαύξησης του χώρου των διανυσματικών αναπαραστάσεων των περιγραφών. Η τεχνική αυτή ονομάζεται *επαύξηση συνθήκης* (conditioning augmentation - CA).

Για κάθε διανυσματική αναπαράσταση μιας περιγραφής ϕ_t , επανξιμένες αναπαραστάσεις μπορούν να παραχθούν δειγματοληπτώντας την Γκαουσιανή κατανομή $\mathcal{N}(\mu(\phi_t), \Sigma(\phi_t))$, όπου $\mu(\phi_t)$ η μέση τιμή και $\Sigma(\phi_t)$ ο διαγώνιος πίνακας συνδιακύμανσης. Με την τεχνική αυτή, λοιπόν, παράγονται περισσότερα δείγματα εκπαίδευσης έχοντας ένα μικρό αριθμό από ζεύγη εικόνας-κειμένου και επομένως το σύστημα παρουσιάζει μεγαλύτερη αντοχή ως προς τις διαταραχές που προκύπτουν από τον διανυσματικό χώρο των συνθηκών. Για να ενισχυθεί ακόμα περισσότερο η αντοχή αυτή και να αποφευχθεί η υπερπροσαρμογή του συστήματος, εισάγεται ένας όρος κανονικοποίησης στη συνάρτηση βελτιστοποίησης του γεννήτορα. Ο όρος κανονικοποίησης αποτελείται από την απόκλιση Kullback-Leibler ανάμεσα στην Γκαουσιανή κατανομή των διανυσματικών αναπαραστάσεων και στην κανονική Γκαουσιανή κατανομή. Στην Εξίσωση 3.3 παρουσιάζεται ο τύπος υπολογισμού της απόκλισης.

$$D_{KL}(\mathcal{N}(\mu(\phi_t), \Sigma(\phi_t)) \parallel \mathcal{N}(0, 1)) \quad (3.3)$$

3.2.2 Πρώτο στάδιο

Σκοπός του πρώτου σταδίου είναι η παραγωγή μίας αρχικής εικόνας με αδρά σχήματα και σωστά χρώματα για το κεντρικό αντικείμενο της απεικόνισης. Έστω ϕ_t η διανυσματική αναπαράσταση της δοσμένης λεκτικής περιγραφής. Οι μεταβλητές συνθήκης $\hat{c}_0$ παράγονται από την τεχνική CA που αναφέρθηκε παραπάνω. Το CGAN του πρώτου σταδίου, λαμβάνοντας ως είσοδο την τυχαία μεταβλητή θορύβου z και τη συνθήκη $\hat{c}_0$, εκπαιδεύει τον διευκρινιστή D_0 για να μεγιστοποιήσει το $\mathcal{L}_{D_0}$ στην Εξίσωση 3.4 και τον γεννήτορα G_0 για να ελαχιστοποιήσει το $\mathcal{L}_{G_0}$ στην Εξίσωση (3.5)

$$\mathcal{L}_{D_0} = \mathbb{E}_{(I_0, t) \sim p_{data}} [\log D_0(I_0, \phi_t)] + \mathbb{E}_{z \sim p_z, t \sim p_{data}} [\log (1 - D_0(G_0(z, \hat{c}_0), \phi_t))] \quad (3.4)$$

$$\mathcal{L}_{G_0} = \mathbb{E}_{z \sim p_z, t \sim p_{data}} [\log (1 - D_0(G_0(z, \hat{c}_0), \phi_t))] + \lambda D_{KL}(\mathcal{N}(\mu_0(\phi_t), \Sigma_0(\phi_t)) \parallel \mathcal{N}(0, 1)) \quad (3.5)$$

όπου η πραγματική εικόνα I_0 και η λεκτική περιγραφή t προέρχονται από την κατανομή των πραγματικών δεδομένων p_{data} , z είναι το διάνυσμα τυχαίου θορύβου που δειγματοληπτείται από την Γκαουσιανή κατανομή p_z και λ είναι μία παράμετρος κανονικοποίησης για την εξισορρόπηση των δύο όρων της Εξίσωσης 3.5. Οι συγγραφείς επέλεξαν την τιμή $\lambda = 1$ για τα πειράματα. Επίσης χρησιμοποίησαν την τεχνική της επαναπαραμετροποίησης, που προτάθηκε στους μεταβλητούς αυτοκωδικοποιητές (Variational Autoencoders - VAEs) [King13], με την οποία τα $\mu_0(\phi_t)$ και $\Sigma_0(\phi_t)$ μαθαίνονται σε συνδυασμό με το υπόλοιπο δίκτυο.

Αρχιτεκτονική

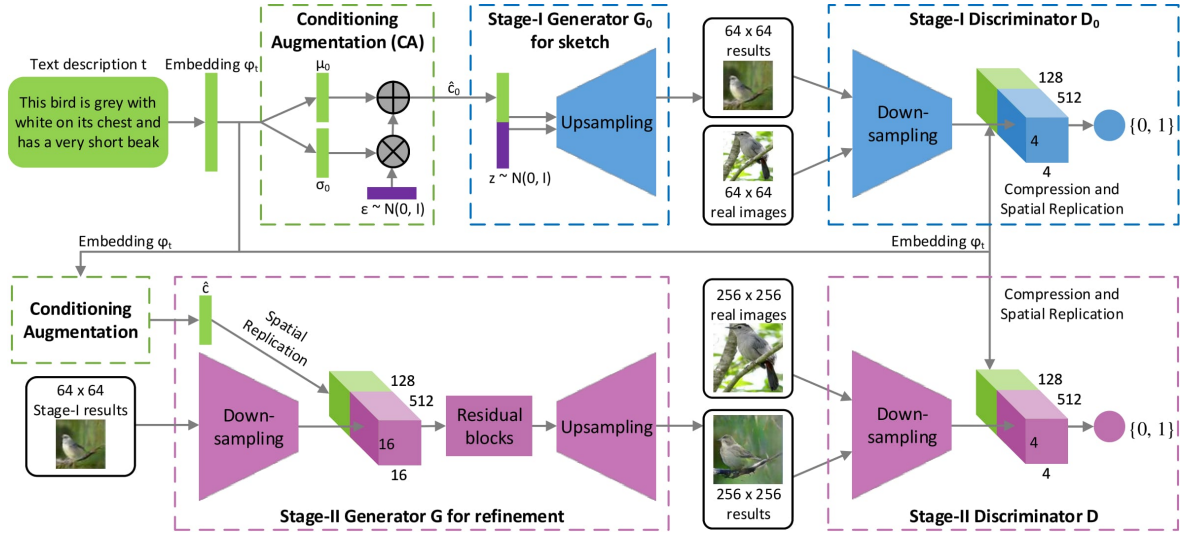
Η αρχιτεκτονική του γεννήτορα του πρώτου σταδίου G_0 παρουσιάζει ομοιότητες με εκείνη του GAN-CLS (Ενότητα 3.1) με την προσθήκη της τεχνικής CA για τις μεταβλητές συνθήκης $\hat{c}_0$. Πιο συγκεκριμένα, για να παραχθεί το $\hat{c}_0$, η διανυσματική αναπαράσταση του κειμένου ϕ_t τροφοδοτείται σε ένα πλήρως συνδεδεμένο δίκτυο για να παραχθούν τα μ_0 και σ_0 , όπου σ_0 είναι οι τιμές της διαγωνίου του Σ_0 , για την Γκαουσιανή κατανομή $\mathcal{N}(\mu_0(\phi_t), \Sigma_0(\phi_t))$. Στη συνέχεια το $\hat{c}_0$ υπολογίζεται με βάση τον τύπο της Εξίσωσης 3.6

$$\hat{c}_0 = \mu_0 + \sigma_0 \odot \epsilon \quad (3.6)$$

όπου το $\odot$ εκφράζει τον πολλαπλασιασμό στοιχείο προς στοιχείο των διανυσμάτων και $\epsilon \sim \mathcal{N}(0, 1)$.

Στη συνέχεια το διάνυσμα $\hat{c}_0$ συνενώνεται με το διάνυσμα του τυχαίου θορύβου z και περνάει μέσα από μία σειρά από *επίπεδα αύξησης διαστάσεων της εισόδου* (up-sampling layers) και συνεπικτικά επίπεδα προκειμένου να παραχθεί η εικόνα του πρώτου σταδίου. Στο πάνω μέρος του Σχήματος 3.2 παρουσιάζεται εποπτικά αυτή η διαδικασία.

Στον διευκρινιστή του πρώτου σταδίου D_0 , η εικόνα που λαμβάνεται ως είσοδος περνά από μία σειρά από *επίπεδα μείωσης διαστάσεων της εισόδου* (down-sampling layers), μέχρι μία συγκεκριμένη χωρική διάσταση. Η διανυσματική αναπαράσταση του κειμένου ϕ_t συμπίεζεται αρχικά σε μικρότερες διαστάσεις με τη χρήση ενός πλήρους συνδεδεμένου επιπέδου και στη συνέχεια επαναλαμβάνεται χωρικά προκειμένου να παραχθεί ένας τανυστής με την παραπάνω χωρική διάσταση. Τότε, ο χάρτης



Σχήμα 3.2: Απεικόνιση της αρχιτεκτονικής του δικτύου StackGAN. Ο Γεννήτορας του πρώτου σταδίου σχεδιάζει μία εικόνα χαμηλής ανάλυσης με αδρά σχήματα και βασικά χρώματα του κεντρικού αντικείμενου της απεικόνισης από το δοσμένο κείμενο και τον τυχαίο θόρυβο z . Ο Γεννήτορας του δεύτερου σταδίου, χρησιμοποιώντας το αποτέλεσμα του πρώτου σταδίου, διορθώνει τις όποιες ατέλειες και παράγει μια πιο ρεαλιστική εικόνα υψηλότερης ανάλυσης με περισσότερες λεπτομέρειες. (Πηγή: [Zhan16])

χαρακτηριστικών συνενώνεται με τον τανυστή της διανυσματικής περιγραφής στη διάσταση του καναλιού. Ο τανυστής που προκύπτει από τη συνένωση τροφοδοτείται σε ένα 1×1 συνελκτικό επίπεδο προκειμένου να μάθει από κοινού χαρακτηριστικά της εικόνας και της περιγραφής. Τέλος, με τη χρήση ενός πλήρους συνδεδεμένου επιπέδου δίνεται ως έξοδος η τιμή της απόφασης του διευκρινιστή.

3.2.3 Δεύτερο στάδιο

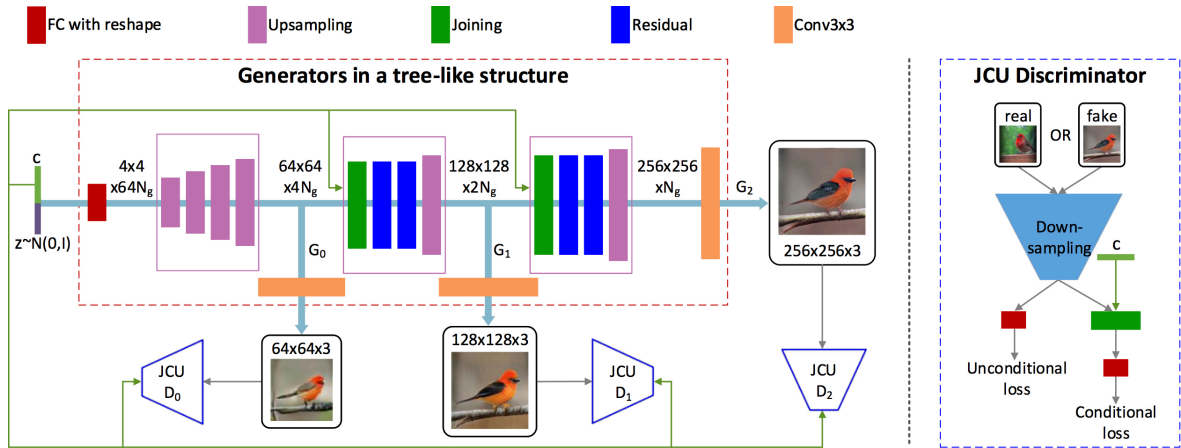
Η εικόνα που παράγεται από το πρώτο στάδιο του StackGAN είναι χαμηλών διαστάσεων και ελλιπής σε λεπτομέρειες. Σκοπός του δεύτερου σταδίου, λοιπόν, είναι να συνεχίσει τη δουλειά που πραγματοποίησε το πρώτο στάδιο, διορθώνοντας τις όποιες ατέλειες, προσθέτοντας περισσότερες πληροφορίες για το κεντρικό αντικείμενο της περιγραφής και αυξάνοντας την ανάλυση της εικόνας. Συνεπώς, το CGAN του δεύτερου σταδίου διαφοροποιείται από τη βασική αρχιτεκτονική. Πιο συγκεκριμένα, το διάνυσμα του τυχαίου θορύβου z δε χρησιμοποιείται σε αυτό το στάδιο. Η τυχαιότητα του συστήματος εισάγεται με την εικόνα του πρώτου σταδίου που δίνεται ως είσοδος στον γεννήτορα G του δεύτερου σταδίου. Επιπλέον, χρησιμοποιείται και το διάνυσμα μεταβλητών συνθήκης $\hat{c}$, το οποίο παράγεται με αντίστοιχο τρόπο με το $\hat{c}_0$ του πρώτου σταδίου και από την ίδια διανυσματική αναπαράσταση ϕ_t , αλλά με διαφορετικό πλήρως συνδεδεμένο επίπεδο προκειμένου να παραχθούν διαφορετικές μέσες τιμές και τυπικές αποκλίσεις.

Το CGAN του δεύτερου σταδίου, λαμβάνοντας ως είσοδο την σύνθετη εικόνα χαμηλής ανάλυσης του πρώτου σταδίου s_0 και τη συνθήκη $\hat{c}$, εκπαιδεύει τον διευκρινιστή D για να μεγιστοποιήσει το $\mathcal{L}_D$ στην Εξίσωση 3.7 και τον γεννήτορα G για να ελαχιστοποιήσει το $\mathcal{L}_G$ στην Εξίσωση (3.8)

$$\mathcal{L}_D = \mathbb{E}_{(I,t) \sim p_{data}} [\log D(I, \phi_t)] + \mathbb{E}_{s_0 \sim p_{G_0}, t \sim p_{data}} [\log (1 - D(G(s_0, \hat{c}), \phi_t))] \quad (3.7)$$

$$\mathcal{L}_G = \mathbb{E}_{s_0 \sim p_{G_0}, t \sim p_{data}} [\log (1 - D(G(s_0, \hat{c}), \phi_t))] + \lambda D_{KL}(\mathcal{N}(\mu_0(\phi_t), \Sigma_0(\phi_t)) || \mathcal{N}(0, 1)) \quad (3.8)$$

όπου η πραγματική εικόνα I και η λεκτική περιγραφή t προέρχονται από την κατανομή των πραγματικών δεδομένων p_{data} και λ είναι μία παράμετρος κανονικοποίησης για την εξισορρόπηση των δύο όρων της Εξίσωσης 3.8. Οι συγγραφείς επέλεξαν την τιμή $\lambda = 1$ για τα πειράματα.



Σχήμα 3.3: Απεικόνιση της αρχιτεκτονικής του δικτύου StackGAN++. (Πηγή: [Zhan17])

Αρχιτεκτονική

Η αρχιτεκτονική του γεννήτορα του δεύτερου σταδίου G έχει τη μορφή κωδικοποιητή - αποκωδικοποιητή. Αρχικά, η εικόνα του πρώτου σταδίου που λαμβάνει ως είσοδο περνά από μία σειρά από επίπεδα μείωσης διαστάσεων μέχρι μία συγκεκριμένη χωρική διάσταση. Αντίστοιχα με το πρώτο στάδιο, η διανυσματική αναπαράσταση του κειμένου ϕ_t συμπιέζεται αρχικά σε μικρότερες διαστάσεις με τη χρήση ενός πλήρους συνδεδεμένου επιπέδου και στη συνέχεια επαναλαμβάνεται χωρικά προκειμένου να παραχθεί ένας ταυστής με την παραπάνω χωρική διάσταση. Έπειτα, οι χάρτες χαρακτηριστικών της εικόνας και της περιγραφής συνενώνονται στη διάσταση του καναλιού και περνούν μέσα από πολλαπλά Residual Blocks [He16]. Τελικά, χρησιμοποιώντας μία σειρά από επίπεδα αύξησης διαστάσεων, παράγεται μία ρεαλιστική εικόνα υψηλής ανάλυσης (256x256). Στο κάτω μέρος του Σχήματος 3.2 παρουσιάζεται εποπτικά αυτή η διαδικασία.

Ο διευκρινιστής του δεύτερου σταδίου D , είναι όμοιος με τον διευκρινιστή του πρώτου σταδίου D_0 , με την προσθήκη επιπλέον επιπέδων μείωσης διαστάσεων λόγω υψηλότερης ανάλυσης της εικόνας εισόδου.

3.3 StackGAN++

Η πρώτη έκδοση του StackGAN, όπως παρουσιάστηκε στην Ενότητα 3.2, αποτελείται από δύο ξεχωριστά δίκτυα. Το δίκτυο πρώτου σταδίου ήταν υπεύθυνο για την παραγωγή εικόνας χαμηλής ανάλυσης με αδρά χαρακτηριστικά, ενώ το δεύτερο ήταν υπεύθυνο για τη βελτιστοποίηση της εικόνας του πρώτου σταδίου και την αύξηση της ανάλυσής της. Αντιθέτως, η δεύτερη έκδοση του StackGAN (StackGAN++) αποτελείται από μία ενιαία αρχιτεκτονική. Και σε αυτή την περίπτωση η εικόνα περνά από πολλαπλά στάδια βελτιστοποίησης, με τη διαφορά ότι τα στάδια αυτά αποτελούν τμήματα του ίδιου δικτύου. Επιπροσθέτως, χρησιμοποιείται η τεχνική CA που περιγράφηκε στην Ενότητα 3.2.1.

3.3.1 Αρχιτεκτονική δικτύου

Το StackGAN++ αποτελείται από πολλαπλούς γεννήτορες και διευκρινιστές, οι οποίοι τοποθετούνται σε μία δομή όμοια με δενδρική. Πιο συγκεκριμένα, οι εικόνες, από τη χαμηλή μέχρι την υψηλή ανάλυση, παράγονται από διαφορετικά κλαδιά του δένδρου. Σε κάθε κλαδί, ο γεννήτορας παράγει μία εικόνα των εκάστοτε διαστάσεων του κλαδιού και ο διευκρινιστής είναι υπεύθυνος να αναγνωρίσει αν η εικόνα προέρχεται από τον γεννήτορα ή την πραγματική κατανομή εικόνων αυτής της διάστασης. Οι γεννήτορες εκπαιδεύονται συλλογικά για να προσεγγίσουν τις πολλαπλές κατανομές δεδομένων. Στο Σχήμα 3.3 παρουσιάζεται η εν λόγω αρχιτεκτονική, η οποία μπορεί να επεκταθεί και για περισσότερα από τρία κλαδιά.

3.3.2 Μοντελοποίηση παραγωγής εικόνας υπό συνθήκη

Όπως αναφέρθηκε προηγουμένως, το μοντέλο του StackGAN++ λαμβάνει ως είσοδο ένα διάνυσμα θορύβου $z \sim p_{noise}$ και της μεταβλητής συνθήκης c που παράγονται από την τεχνική της επαύξησης της διανυσματικής αναπαράστασης των λεκτικών περιγραφών. Έπειτα, περνώντας μέσα από τα διαφορετικά κλαδιά της δενδρικής δομής του, παράγονται εικόνες διαφορετικών διαστάσεων. Τα κρυφά χαρακτηριστικά h_i του κάθε κλαδιού υπολογίζονται με τον μη γραμμικό μετασχηματισμό της Εξίσωσης 3.9

$$h_0 = F_0(c, z); \quad h_i = F_i(h_{i-1}, c), \quad i = 1, 2, \dots, m-1 \quad (3.9)$$

όπου το h_i αναπαριστά τα κρυφά χαρακτηριστικά του i -οστού κλαδιού, m είναι ο συνολικός αριθμός από κλαδιά και τα F_i αντιπροσωπεύουν νευρωνικά δίκτυα. Προκειμένου σε κάθε κλαδί να αξιοποιείται πληροφορία της λεκτικής περιγραφής που είχε αγνοηθεί προηγουμένως, η διανυσματική αναπαράσταση της λεκτικής περιγραφής συνενώνεται με τα κρυφά χαρακτηριστικά κάθε κλαδιού. Κατ' επέκταση, οι εικόνες πολλαπλών διαστάσεων παράγονται με τον μετασχηματισμό της Εξίσωσης 3.10.

$$s_i = G_i(h_i), \quad i = 1, 2, \dots, m-1 \quad (3.10)$$

Κίνητρο για τη συγκεκριμένη προσέγγιση μοντελοποίησης του προβλήματος αποτέλεσε το γεγονός ότι, αν η κατανομή των σύνθετων δεδομένων του γεννήτορα κάποιου κλαδιού έχει κοινά χαρακτηριστικά με την πραγματική κατανομή δεδομένων στις αντίστοιχες διαστάσεις, τότε η επικάλυψη μεταξύ των κατανομών μπορεί να παράξει κατάλληλες κλίσεις (gradients), οι οποίες θα ευνοήσουν τη σταθεροποίηση της εκπαίδευσης του συνολικού δικτύου.

3.3.3 Συναρτήσεις βελτιστοποίησης

Στο StackGAN++ οι συναρτήσεις βελτιστοποίησης αποτελούνται από δύο ειδών σφάλματα: (i) το σφάλμα *δίχως συνθήκη* (unconditional loss), το οποίο εξετάζει τη γνησιότητα της εικόνας, δηλαδή αν πρόκειται για σύνθετη ή πραγματική και (ii) το σφάλμα *υπό συνθήκη* (conditional loss), που αξιολογεί την αντιστοιχία μεταξύ εικόνας και συνθήκης (λεκτικής περιγραφής).

Γεννήτορας

Η συνάρτηση σφάλματος του κάθε γεννήτορα είναι παρουσιάζεται στην Εξίσωση 3.11

$$\mathcal{L}_{G_i} = \underbrace{-\mathbb{E}_{s_i \sim p_{G_i}} [\log D_i(s_i)]}_{\text{unconditional loss}} + \underbrace{-\mathbb{E}_{s_i \sim p_{G_i}} [\log D_i(s_i, c)]}_{\text{conditional loss}} \quad (3.11)$$

ενώ το συνολικό σφάλμα για την από κοινού τους εκπαίδευση απεικονίζεται στην Εξίσωση 3.12.

$$\mathcal{L}_G = \sum_{i=1}^m \mathcal{L}_{G_i} \quad (3.12)$$

Διευκρινιστής

Αντίστοιχα, η συνάρτηση σφάλματος του κάθε διευκρινιστή αποτελείται από τους δύο όρους της Εξίσωσης 3.13.

$$\begin{aligned} \mathcal{L}_{D_i} = & \underbrace{-\mathbb{E}_{x_i \sim p_{data_i}} [\log D_i(x_i)] - \mathbb{E}_{s_i \sim p_{G_i}} [\log (1 - D_i(s_i))]}_{\text{unconditional loss}} + \\ & \underbrace{-\mathbb{E}_{x_i \sim p_{data_i}} [\log (D_i(x_i, c))] - \mathbb{E}_{s_i \sim p_{G_i}} [\log (1 - D_i(s_i, c))]}_{\text{conditional loss}} \end{aligned} \quad (3.13)$$

Κεφάλαιο 4

TeleGAN

Στην παρούσα διατριβή προτείνουμε ένα πρωτότυπο μοντέλο, το TeleGAN, εμπνευσμένο από το StackGAN [Zhan16] αλλά και από την εξελικτική πορεία της ποιότητας των παρουσιαζόμενων εικόνων στην τηλεόραση, το οποίο συνδυάζει τρία διαφορετικά στάδια. Το βασικό πλεονέκτημα της παρουσιαζόμενης αρχιτεκτονικής είναι ότι επιτρέπει διορθωτικές επεμβάσεις σε κάθε στάδιο ξεχωριστά. Στα πειράματά μας (Ενότητα 5) παρατηρήσαμε ότι η βελτίωση των αποτελεσμάτων ενός σταδίου, οδηγεί στη συνολική βελτίωση του παραγόμενου αποτελέσματος. Στην Ενότητα 6 αναφέρονται και ορισμένες τροποποιήσεις του παρουσιαζόμενου μοντέλου που πιθανόν να οδηγήσουν στη βελτίωση των συνθετικών εικόνων.

Όμοια με το StackGAN, το TeleGAN διασπά το πολύπλοκο πρόβλημα της σύνθεσης ρεαλιστικής εικόνας υψηλών διαστάσεων σε επιμέρους διαχειρίσιμα υποπροβλήματα, χρησιμοποιώντας στοιβαγμένα GANs. Ωστόσο, ο τρόπος διάσπασης στο μοντέλο μας είναι διαφορετικός. Πιο συγκεκριμένα η σύνθεση εικόνας από λεκτική περιγραφή διασπάται στα εξής τρία στάδια:

- **Στάδιο-1:** Σύνθεση ασπρόμαυρης εικόνας με το σχήμα του αντικειμένου, διαστάσεων 128x128, λαμβάνοντας ως είσοδο τη λεκτική περιγραφή και ένα διάνυσμα θορύβου.
- **Στάδιο-2:** Επιχρωματισμός της εικόνας του πρώτου σταδίου και παραγωγή μίας ρεαλιστικής έγχρωμης εικόνας διαστάσεων 128x128, χρησιμοποιώντας και πάλι τη λεκτική περιγραφή.
- **Στάδιο-3:** Αύξηση των διαστάσεων της εικόνας του δεύτερου σταδίου και παραγωγή μίας ρεαλιστικής εικόνας υψηλών διαστάσεων (256x256).

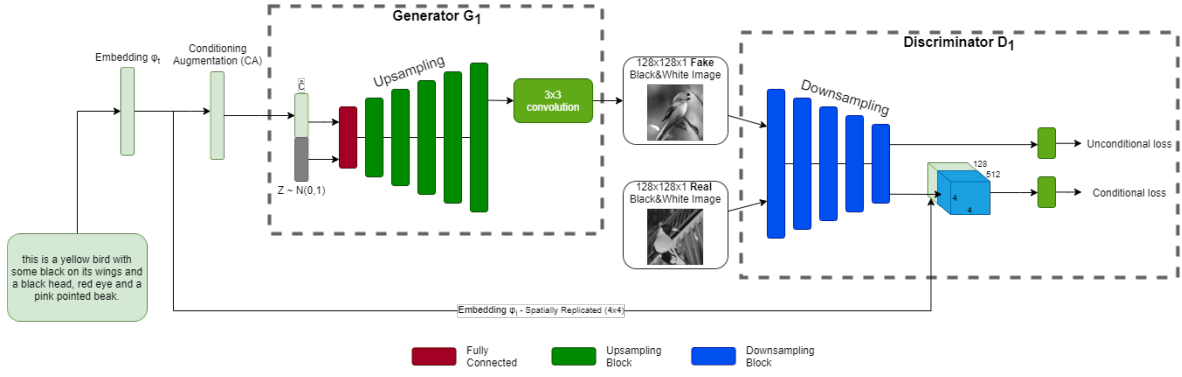
Το μοντέλο μας αξιοποιεί το έργο της ερευνητικής κοινότητας στους τομείς του *επιχρωματισμού* (colorization) [Isol18, Naze18] και της *υπερ-ανάλυσης* (super-resolution) [Ledi17] για το δεύτερο και το τρίτο στάδιο αντίστοιχα.

4.1 Επαύξηση Διανυσματικής Αναπαράστασης

Στο δίκτυο μας χρησιμοποιούμε την τεχνική CA [Zhan16], που όπως εξηγήσαμε στην Ενότητα 3.2.1, αυξάνει τις διαστάσεις των μεταβλητών συνθήκης c . Αυτό επιτυγχάνεται μέσω δειγματοληψίας μεταβλητών $\hat{c}$ από τη Γκαουσιανή κατανομή $\mathcal{N}(\mu(\phi_t), \Sigma(\phi_t))$, όπου ϕ_t η διανυσματική αναπαράσταση της περιγραφής, $\mu(\phi_t)$ η μέση τιμή και $\Sigma(\phi_t)$ ο διαγώνιος πίνακας συνδιακύμανσης. Επίσης, χρησιμοποιούμε τον όρο κανονικοποίησης της Εξίσωσης 3.3.

4.2 Στάδιο-1

Το CGAN του πρώτου σταδίου είναι υπεύθυνο για την παραγωγή εικόνων διαστάσεων 128x128, δίνοντας έμφαση στο σχήμα των αντικειμένων. Για τον σκοπό αυτό αφαιρούμε την πληροφορία του χρώματος, προκειμένου το δίκτυο να εστιάσει στη σχεδίαση των σχημάτων. Συνεπώς, οι εικόνες που παράγει ο γεννήτορας (G_1) είναι ασπρόμαυρες και οι εικόνες της κατανομής των πραγματικών δεδομένων μετατρέπονται σε ασπρόμαυρες προτού δοθούν ως είσοδοι, μαζί με τη διανυσματική αναπαράσταση της εκάστοτε λεκτικής περιγραφής, στο διευκρινιστή (D_1).



Σχήμα 4.1: Απεικόνιση της αρχιτεκτονικής του δικτύου του πρώτου σταδίου του TeleGAN. Ο γεννήτορας του πρώτου σταδίου σχεδιάζει μία ασπρόμαυρη εικόνα ανάλυσης 128x128 του κεντρικού αντικειμένου της απεικόνισης από το δοσμένο κείμενο και τον τυχαίο θόρυβο z .

Έστω ϕ_t η διανυσματική αναπαράσταση της περιγραφής εισόδου που προκύπτει από το προεκπαιδευμένο μοντέλο κωδικοποίησης του [Reed16a], $\hat{c}_1$ οι μεταβλητές συνθήκης που προκύπτουν από την Γκαουσιανή κατανομή $\mathcal{N}(\mu_1(\phi_t), \Sigma_1(\phi_t))$ της τεχνικής CA, z τυχαία μεταβλητή από την Γκαουσιανή κατανομή θορύβου p_z και $1 \sim p_{data}$ εικόνα της κατανομής των πραγματικών δεδομένων, η οποία έχει μετασχηματιστεί κατάλληλα για το πρώτο στάδιο (128x128, ασπρόμαυρη). Επίσης, έστω $s_1 \sim p_{G_1}$ εικόνα που παράγεται από την κατανομή του γεννήτορα ($s_1 = G_1(z, \hat{c}_1)$). Υιοθετώντας τις συναρτήσεις βελτιστοποίησης του [Zhan17], το CGAN του πρώτου σταδίου εκπαιδεύει εναλλάξ τον διευκρινιστή D_1 και τον γεννήτορα G_1 για να μεγιστοποιήσει το $\mathcal{L}_{D_1}$ στην Εξίσωση 4.1 και να ελαχιστοποιήσει το $\mathcal{L}_{G_1}$ στην Εξίσωση 4.2

$$\mathcal{L}_{D_1} = \underbrace{-\mathbb{E}_{1 \sim p_{data}} [\log D_1(1)] - \mathbb{E}_{s_1 \sim p_{G_1}} [\log (1 - D_1(s_1))]}_{\text{unconditional loss}} + \underbrace{-\mathbb{E}_{1 \sim p_{data}} [\log (D_1(1, \hat{c}_1))] - \mathbb{E}_{s_1 \sim p_{G_1}} [\log (1 - D_1(s_1, \hat{c}_1))]}_{\text{conditional loss}} \quad (4.1)$$

$$\mathcal{L}_{G_1} = \underbrace{-\mathbb{E}_{s_1 \sim p_{G_1}} [\log D_1(s_1)]}_{\text{unconditional loss}} + \underbrace{-\mathbb{E}_{s_1 \sim p_{G_1}} [\log D_1(s_1, \hat{c}_1)]}_{\text{conditional loss}} + \lambda D_{KL}(\mathcal{N}(\mu_1(\phi_t), \Sigma_1(\phi_t)) \parallel \mathcal{N}(0, 1)) \quad (4.2)$$

όπου λ είναι μία παράμετρος κανονικοποίησης για την εξισορρόπηση των όρων της Εξίσωσης 4.2. Σε αντιστοιχία με το [Zhan16] χρησιμοποιείται η μέθοδος επαναπαραμετροποίησης του [King13], στην οποία τα $\mu_1(\phi_t)$ και $\Sigma_1(\phi_t)$ μαθαίνονται σε συνδυασμό με το υπόλοιπο δίκτυο.

Με τη χρήση του σφάλματος δίχως συνθήκη εξετάζεται η γνησιότητα της εικόνας, δηλαδή αν η εικόνα προέρχεται από την πραγματική κατανομή ή την κατανομή του γεννήτορα G_1 , ενώ με το σφάλμα υπό συνθήκη αξιολογείται η αντιστοιχία μεταξύ εικόνας και λεκτικής περιγραφής.

4.2.1 Αρχιτεκτονική

Η αρχιτεκτονική του γεννήτορα G_1 είναι όμοια με εκείνη του [Zhan16] που περιγράφηκε στην Ενότητα 3.2.2. Για να παραχθεί το $\hat{c}_1$, η διανυσματική αναπαράσταση του κειμένου ϕ_t τροφοδοτείται σε ένα πλήρως συνδεδεμένο δίκτυο για να παραχθούν τα μ_1 και σ_1 , όπου σ_1 είναι οι τιμές της διαγωνίου του Σ_1 , για την γκαουσιανή κατανομή $\mathcal{N}(\mu_1(\phi_t), \Sigma_1(\phi_t))$. Στη συνέχεια το $\hat{c}_1$ υπολογίζεται με βάση τον τύπο της Εξίσωσης 4.3.

$$\hat{c}_1 = \mu_1 + \sigma_1 \odot \epsilon \quad (4.3)$$

όπου το $\odot$ εκφράζει τον πολλαπλασιασμό στοιχείο προς στοιχείο των διανυσμάτων και $\epsilon \sim \mathcal{N}(0, 1)$.

Έπειτα το διάνυσμα $\hat{c}_1$ συνενώνεται με το διάνυσμα του τυχαίου θορύβου z και το διάνυσμα που παράγεται περνά μέσα από ένα πλήρως συνδεδεμένο επίπεδο προκειμένου να μετασχηματιστεί σε κατάλληλες διαστάσεις. Τέλος, το διάνυσμα αυτό, περνάει μέσα από μία σειρά από επίπεδα αύξησης διαστάσεων της εισόδου και συνελκτικά επίπεδα προκειμένου να παραχθεί η ασπρόμαυρη εικόνα του πρώτου σταδίου.

Στον διευκρινιστή D_1 , η εικόνα που λαμβάνεται ως είσοδος περνά μέσα από μία σειρά από επίπεδα μείωσης διαστάσεων, μέχρι να καταλήξει σε χωρική διάσταση 4×4 . Έπειτα, χρησιμοποιείται ο χάρτης χαρακτηριστικών που έχει προκύψει προκειμένου να παραχθεί η πρόβλεψη ως προς την ρεαλιστικότητα της εικόνας (πραγματική ή συνθετική). Στη συνέχεια, ο χάρτης χαρακτηριστικών συνενώνεται με τη διανυσματική αναπαράσταση κειμένου ϕ_t , η οποία πρώτα επαναλαμβάνεται χωρικά για να αποκτήσει χωρικές διαστάσεις 4×4 . Τέλος, με τη χρήση του τανυστή που προέκυψε από τη συνένωση, παράγεται η πρόβλεψη για την αντιστοιχία ή μη της εικόνας με την λεκτική περιγραφή της οποίας η διανυσματική αναπαράσταση δόθηκε ως είσοδος στον διευκρινιστή.

4.2.2 Λεπτομέρειες Υλοποίησης

Τα μπλοκ αύξησης διαστάσεων υλοποιούνται με χρήση τεχνικής πλησιέστερου γείτονα (nearest-neighbor), 3×3 συνέλιξη με μοναδιαίο γέμισμα (padding=1), κανονικοποίηση δέσμης και συνάρτηση ενεργοποίησης διορθωμένης γραμμικής μονάδας (rectified linear unit - ReLU). Η τελική εικόνα παράγεται με τη χρήση συνελκτικού επιπέδου 3×3 και υπερβολική εφαπτομένη ως συνάρτηση ενεργοποίησης.

Στα επίπεδα μείωσης διαστάσεων, η χωρική διάσταση της εισόδου υποδιπλασιάζεται. Για τον σκοπό αυτόν χρησιμοποιούνται συνελκτικά επίπεδα 4×4 με βήμα 2 (stride=2) και μοναδιαίο γέμισμα. Κάθε μία από αυτές τις συνέλιξεις ακολουθείται από κανονικοποίηση δέσμης και συνάρτηση ενεργοποίησης Leaky ReLU με αρνητική κλίση 0.2 (negative slope=0.2), με εξαίρεση το πρώτο μπλοκ που δεν περιέχει κανονικοποίηση δέσμης. Επιπλέον, χρησιμοποιούνται συνελκτικά επίπεδα 3×3 με μοναδιαίο γέμισμα και 1×1 δίχως γέμισμα για μετασχηματισμούς των χαρτών χαρακτηριστικών. Τέλος, σιγμοειδής συνάρτηση ενεργοποίησης (sigmoid) χρησιμοποιείται για την παραγωγή των πιθανοτήτων πρόβλεψης από τον διευκρινιστή D_1 .

4.3 Στάδιο-2

Το CGAN του δευτέρου σταδίου είναι υπεύθυνο για τον επιχρωματισμό των συνθετικών εικόνων που παράγονται από το πρώτο στάδιο. Πιο συγκεκριμένα, ο γεννήτορας G_2 του σταδίου αυτού λαμβάνει ως είσοδο τη συνθετική ασπρόμαυρη εικόνα που παράγει ο γεννήτορας G_1 διαστάσεων 128×128 και προσθέτει την πληροφορία του χρώματος, χρησιμοποιώντας και πάλι τη διανυσματική αναπαράσταση της λεκτικής περιγραφής. Συνεπώς, οι εικόνες που δίνει ως έξοδο ο G_2 διατηρούν τις διαστάσεις 128×128 , όντας πλέον έγχρωμες. Ο διευκρινιστής D_2 λαμβάνει ως είσοδο τις συνθετικές εικόνες του G_2 και τις πραγματικές εικόνες σε διαστάσεις 128×128 , μαζί με τη διανυσματική αναπαράσταση της εκάστοτε λεκτικής περιγραφής και εξετάζει και πάλι τη γνησιότητα της εικόνας, συνθετική ή πραγματική, καθώς και την αντιστοιχία μεταξύ της λεκτικής περιγραφής και της παραγόμενης εικόνας. Συνεπώς, δεν προστίθεται συγκεκριμένο σφάλμα για τα χρώματα στο βασικό μοντέλο και οι συναρτήσεις βελτιστοποίησης του διευκρινιστή D_2 (Εξίσωση 4.4) και του γεννήτορα G_2 (Εξίσωση 4.5) είναι όμοιες με τις Εξισώσεις 4.1 και 4.2 αντίστοιχα, με τη διαφορά ότι ο G_2 δε λαμβάνει ως είσοδο

διάνυσμα θορύβου, αλλά τη συνθετική εικόνα του πρώτου σταδίου s_1 , δηλαδή $s_2 = G_2(s_1, \hat{c}_2)$

$$\mathcal{L}_{D_2} = \underbrace{-\mathbb{E}_{I_2 \sim p_{data}} [\log D_2(I_2)] - \mathbb{E}_{s_2 \sim p_{G_2}} [\log (1 - D_2(s_2))]}_{\text{unconditional loss}} + \underbrace{-\mathbb{E}_{I_2 \sim p_{data}} [\log (D_2(I_2, \hat{c}_2))] - \mathbb{E}_{s_2 \sim p_{G_2}} [\log (1 - D_0(s_2, \hat{c}_2))]}_{\text{conditional loss}} \quad (4.4)$$

$$\mathcal{L}_{G_2} = \underbrace{-\mathbb{E}_{s_2 \sim p_{G_2}} [\log D_2(s_2)]}_{\text{unconditional loss}} + \underbrace{-\mathbb{E}_{s_2 \sim p_{G_2}} [\log D_2(s_2, \hat{c}_2)]}_{\text{conditional loss}} + \lambda D_{KL}(\mathcal{N}(\mu_2(\phi_t), \Sigma_2(\phi_t)) || \mathcal{N}(0, 1)) \quad (4.5)$$

όπου I_2 η εικόνα που προέρχεται από την κατανομή των πραγματικών δεδομένων p_{data} και $\hat{c}_2$ οι μεταβλητές συνθήκης που προκύπτουν από την Γκαουσιανή κατανομή $\mathcal{N}(\mu_2(\phi_t), \Sigma_2(\phi_t))$ της τεχνικής CA. Τα $\mu_2(\phi_t)$ και $\Sigma_2(\phi_t)$ είναι διαφορετικά από αυτά του πρώτου σταδίου, ενώ μαθαίνονται επίσης με την τεχνική της επαναπαραμετροποίησης σε συνδυασμό με το υπόλοιπο δίκτυο.

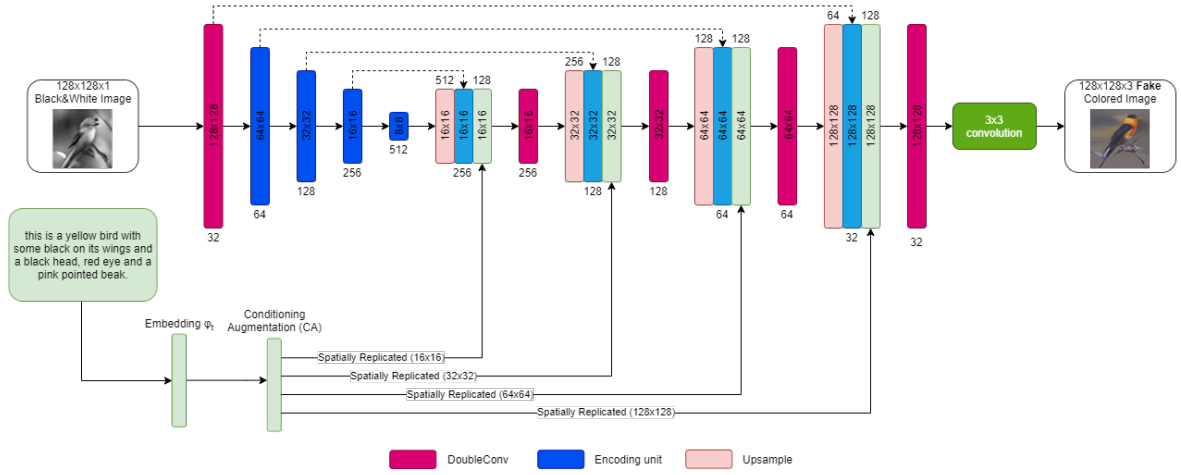
4.3.1 Αρχιτεκτονική

Για τον γεννήτορα G_2 χρησιμοποιείται η γνωστή αρχιτεκτονική U-Net [Ronn15]. Όμοια με την αρχιτεκτονική που προτάθηκε στο [Naze18], η αρχιτεκτονική του μοντέλου μας είναι συμμετρική, με 4 μονάδες κωδικοποίησης (encoding units) και 4 μονάδες αποκωδικοποίησης (decoding units). Για τις μονάδες κωδικοποίησης χρησιμοποιούνται *συγκεντρωτικά επίπεδα μεγίστου* (MaxPooling) ακολουθούμενα από μπλοκ διπλών *συνελκτικών επιπέδων* (double convolutions). Για τις μονάδες αποκωδικοποίησης χρησιμοποιούνται *επίπεδα αύξησης διαστάσεων* (upsample) με χρήση της *διγραμμικής* (bilinear) τεχνικής ακολουθούμενα και πάλι από μπλοκ διπλών *συνελκτικών επιπέδων*. Επίσης, στην αρχιτεκτονική U-Net χρησιμοποιούνται *συνδέσεις παράκαμψης* (skip-connections) μεταξύ των αντίστοιχων επιπέδων στη συμμετρική αρχιτεκτονική. Τέλος, επειδή ο επιχρωματισμός της εικόνας στο πρόβλημά μας πρέπει να γίνει βάσει της λεκτικής περιγραφής, σε κάθε μονάδα αποκωδικοποίησης, η μεταβλητή συνθήκης $\hat{c}_2$ επαναλαμβάνεται στην αντίστοιχη διάσταση βάθους και συνενώνεται με τον χάρτη χαρακτηριστικών της μονάδας.

Η αρχιτεκτονική του διευκρινιστή D_2 είναι παρόμοια με αυτή του πρώτου σταδίου. Η έγχρωμη εικόνα που λαμβάνει ως είσοδο, πραγματική ή συνθετική, περνά μέσα από μία σειρά από επίπεδα μείωσης διαστάσεων της εισόδου μέχρι να καταλήξει σε χωρική διάσταση 4x4. Από τον χάρτη χαρακτηριστικών που προκύπτει θα υπολογιστεί η πιθανότητα πρόβλεψης ως προς τη ρεαλιστικότητα της εικόνας με τη χρήση ενός *συνελκτικού επιπέδου* 4x4 και *σιγμοειδούς συνάρτησης ενεργοποίησης*. Ταυτόχρονα, ο χάρτης αυτός θα συνενωθεί με τη διανυσματική αναπαράσταση κειμένου ϕ_t , η οποία πρώτα επαναλαμβάνεται χωρικά για να αποκτήσει διαστάσεις 4x4. Ο συνενωμένος τανυστής περνά και αυτός μέσα από *συνελκτικό επίπεδο* 4x4 και *σιγμοειδή συνάρτηση ενεργοποίησης*, προκειμένου να υπολογιστεί η πιθανότητα πρόβλεψης ως προς την αντιστοιχία της εικόνας με τη λεκτική περιγραφή.

4.3.2 Λεπτομέρειες Υλοποίησης

Τα μπλοκ διπλών *συνελκτικών επιπέδων* που περιέχονται στις μονάδες κωδικοποίησης και αποκωδικοποίησης του γεννήτορα G_2 αποτελούνται από *συνελκτικό επίπεδο* 3x3 με μοναδιαίο γέμισμα, κανονικοποίηση δέσμης και *συνάρτηση ενεργοποίησης* Leaky ReLU με αρνητική κλίση 0.2, τα οποία επαναλαμβάνονται ακολουθιακά δύο φορές. Επίσης, τα *συγκεντρωτικά επίπεδα μεγίστου* έχουν μέγεθος πυρήνα 2x2 και τα *επίπεδα αύξησης διαστάσεων* με *διγραμμική τεχνική* χρησιμοποιούνται με *κλίμακα* (scale) ίση με 2 και *ευθυγράμμιση στις γωνίες* (align corners). Η τελική εικόνα παράγεται με τη χρήση *συνελκτικού επιπέδου* 3x3 και *υπερβολική εφαιπτομένη* ως *συνάρτηση ενεργοποίησης*.



Σχήμα 4.2: Απεικόνιση της αρχιτεκτονικής του δικτύου του δεύτερου σταδίου του TeleGAN. Ο γεννήτορας του δεύτερου σταδίου συνθέτει μία έγχρωμη εικόνα διαστάσεων 128x128, έχοντας λάβει ως είσοδο τη λεκτική περιγραφή και την ασπρόμαυρη εικόνα του πρώτου σταδίου.

Όμοια με το πρώτο στάδιο, τα επίπεδα μείωσης διαστάσεων αποτελούνται από συνελκτικά επίπεδα 4x4 με βήμα 2 και μοναδιαίο γέμισμα. Κάθε μία από αυτές τις συνελίξεις, που υποδιπλασιάζουν τις διαστάσεις εισόδου των καναλιών, ακολουθείται από κανονικοποίηση δέσμης και συνάρτηση ενεργοποίησης Leaky ReLU με αρνητική κλίση 0.2. Επίσης χρησιμοποιείται συνελκτικό επίπεδο 1x1 για μετασχηματισμό του χάρτη χαρακτηριστικών μετά από συνένωση με τη διανυσματική αναπαράσταση του κειμένου. Τέλος, σιγμοειδής συνάρτηση ενεργοποίησης χρησιμοποιείται για την παραγωγή των πιθανοτήτων πρόβλεψης από τον διευκρινιστή D_2 .

4.4 Στάδιο-3

Για το τρίτο και τελευταίο στάδιο χρησιμοποιείται ένα δίκτυο παρόμοιο με το *GAN υψηλής ανάλυσης* (Super Resolution GAN - SRGAN) [Led17]. Το GAN του τρίτου σταδίου είναι υπεύθυνο για τη μεγέθυνση των έγχρωμων συνθετικών εικόνων που παράγονται στο δεύτερο στάδιο. Ο γεννήτορας G_3 του σταδίου αυτού λαμβάνει ως είσοδο έγχρωμες εικόνες διαστάσεων 128x128, και δίνει ως έξοδο την ίδιας εικόνας σε υψηλότερες διαστάσεις (256x256). Η βασική διαφορά που παρουσιάζει ο G_3 , σε σύγκριση με τους γεννήτορες G_1 και G_2 , είναι ότι δε χρησιμοποιεί τις μεταβλητές συνθήκης κειμένου, δηλαδή τη λεκτική περιγραφή, για την παραγωγή της τελικής εικόνας. Μία ακόμη σημαντική διαφορά εντοπίζεται στη διαδικασία εκπαίδευσης του G_3 , κατά την οποία χρησιμοποιούνται ζεύγη εικόνων που προέρχονται από την κατανομή των πραγματικών δεδομένων. Περισσότερες λεπτομέρειες σχετικά με την εκπαίδευση παρουσιάζονται στην Ενότητα 4.5.

Ο διευκρινιστής D_3 είναι υπεύθυνος για την έλεγχο της ποιότητας των εικόνων, δηλαδή εξετάζει αν η εικόνα υψηλής ανάλυσης που λαμβάνει ως είσοδο προέρχεται από την πραγματική κατανομή δεδομένων ή έχει παραχθεί από τον G_3 . Ως συναρτήσεις σφάλματος για το GAN του τρίτου σταδίου χρησιμοποιήθηκαν αυτές που πρότειναν οι συγγραφείς στο [Led17]. Πιο συγκεκριμένα, στην Εξίσωση 4.6 απεικονίζεται το πρότυπο *σφάλμα αντίληψης* (perceptual loss) για τον γεννήτορα G_3 , το οποίο αποτελείται από το σταθμισμένο άθροισμα του *σφάλματος περιεχομένου* (content loss) της Εξίσωσης 4.7 και του *σφάλματος ανταγωνισμού* (adversarial loss) της Εξίσωσης 4.8

$$l^{s3} = \underbrace{l_{VGG}^{s3}}_{\text{σφάλμα περιεχομένου}} + \underbrace{10^{-3}l_{G_3}^{s3}}_{\text{σφάλμα ανταγωνισμού}} \quad (4.6)$$

όπου $s_3 = G_3(I_2)$ η εικόνα υψηλών διαστάσεων (256x256) που παράγει ο γεννήτορας G_3 δοθείσης εικόνας $I_2 \sim p_{data}$ της κατανομής των πραγματικών δεδομένων σε διαστάσεις 128x128.

Για το σφάλμα περιεχομένου, οι συγγραφείς του [Ledi17] προτείνουν το *σφάλμα VGG* (VGG loss), το οποίο ονομάστηκε έτσι επειδή χρησιμοποιεί το γνωστό δίκτυο VGG-19 [Simo15] προκειμένου να παραχθούν οι χάρτες χαρακτηριστικών για τις πραγματικές και τις συνθετικές εικόνες. Πιο συγκεκριμένα, ορίζοντας ως $\phi_{i,j}$ τον χάρτη χαρακτηριστικών που λαμβάνουμε από το j -οστό συνελκτικό επίπεδο (μετά τη συνάρτηση ενεργοποίησης) και πριν το i -οστό επίπεδο συγκέντρωσης μεγίστου από το προεκπαιδευμένο δίκτυο VGG-19, υπολογίζεται το σφάλμα VGG ως η ευκλείδεια απόσταση μεταξύ του χάρτη χαρακτηριστικών $\phi_{i,j}(s_3)$ της συνθετικής εικόνας s_3 και του χάρτη χαρακτηριστικών $\phi_{i,j}(I_3)$ της εικόνας $I_3 \sim p_{data}$, η οποία προέρχεται από την κατανομή των πραγματικών δεδομένων p_{data} και έχει διαστάσεις 256x256. Στην Εξίσωση 4.7 παρουσιάζεται ο τύπος υπολογισμού του σφάλματος περιεχομένου

$$l_{VGG/i,j}^{s_3} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I_3)_{x,y} - \phi_{i,j}(s_3)_{x,y})^2 \quad (4.7)$$

όπου $W_{i,j}$ και $H_{i,j}$ οι διαστάσεις του αντίστοιχου χάρτη ενεργοποίησης στο δίκτυο VGG-19.

Το σφάλμα ανταγωνισμού (Εξίσωση 4.8) υπολογίζεται από το άθροισμα των πιθανοτήτων του διευκρινιστή D_3 στα δείγματα της τρέχουσας δέσμης δεδομένων

$$l_{G_3}^{s_3} = \sum_{n=1}^N -\log D_3(s_3) \quad (4.8)$$

όπου s_3 η αντίστοιχη συνθετική εικόνα που παράγεται από τον G_3 .

Ο διευκρινιστής D_3 εκπαιδεύεται για να μεγιστοποιηθεί το $\mathcal{L}_{D_3}$ στην Εξίσωση 4.9

$$\mathcal{L}_{D_3} = -\mathbb{E}_{I_3 \sim p_{data}} [\log D_3(I_3)] - \mathbb{E}_{s_3 \sim p_{G_3}} [\log (1 - D_3(s_3))] \quad (4.9)$$

όπου $s_3 = G_3(I_2)$ η εικόνα υψηλών διαστάσεων (256x256) που παράγει ο γεννήτορας G_3 δοθείσας μίας εικόνας $I_2 \sim p_{data}$ της κατανομής των πραγματικών δεδομένων σε διαστάσεις 128x128 και $I_3 \sim p_{data}$ η αντίστοιχη εικόνα της κατανομής των πραγματικών δεδομένων σε διαστάσεις 256x256.

Κατά την εκπαίδευση, στο σταθμισμένο άθροισμα της Εξίσωσης 4.6 προστέθηκαν δύο ακόμα όροι που αναφέρονται στο [Ledi17]. Ο πρώτος όρος είναι το *μέσο τετραγωνικό σφάλμα* (mean squared error - MSE) σε επίπεδο εικονοστοιχείων (pixel-wise). Ο τύπος υπολογισμού του MSE παρουσιάζεται στην Εξίσωση 4.10 και αποτελεί την πιο δημοφιλή μετρική σφάλματος σε δίκτυα υπερ-ανάλυσης [Dong16, Shi16].

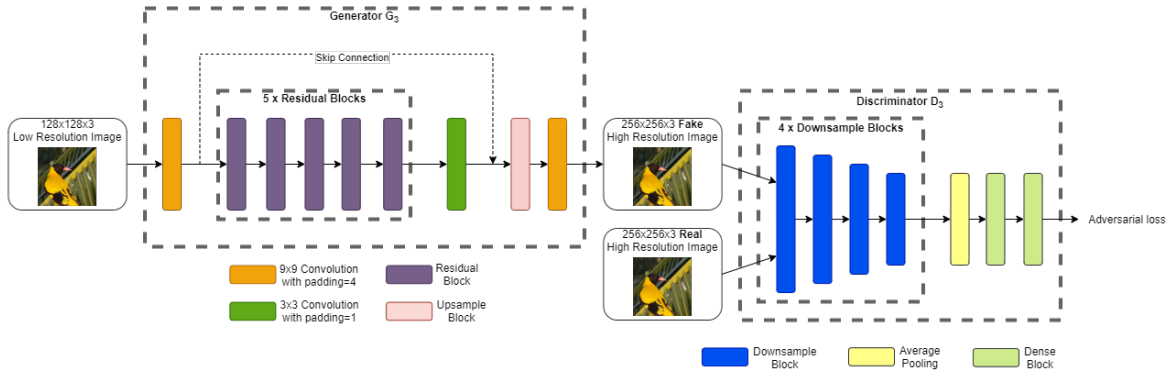
$$l_{MSE}^{s_3} = \frac{1}{r^2WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} ((I_3)_{x,y} - (s_3)_{x,y})^2 \quad (4.10)$$

Ο δεύτερος όρος που προστίθεται είναι το *σφάλμα ολικής διακύμανσης* (total variation loss - TV loss) [Aly05, John16] της σύνθετης εικόνας s_3 του G_3 . Ο υπολογισμός του TV loss μπορεί να γίνει αθροίζοντας τα τετράγωνα των διαφορών των εικονοστοιχείων για όλα τα γειτονικά εικονοστοιχεία, οριζόντια και κάθετα. Συνεπώς, χρησιμοποιούμε τον τύπο υπολογισμού της Εξίσωσης 4.11.

$$l_{TV}^{s_3} = \sum_{c=1}^3 \sum_{i=1}^{H-1} \sum_{j=1}^W ((s_3)_{c,i+1,j} - (s_3)_{c,i,j})^2 + \sum_{c=1}^3 \sum_{i=1}^H \sum_{j=1}^{W-1} ((s_3)_{c,i,j+1} - (s_3)_{c,i,j})^2 \quad (4.11)$$

Έτσι το σταθμισμένο άθροισμα της Εξίσωσης 4.6 έχει πλέον τη μορφή της Εξίσωσης 4.12.

$$l_{total}^{s_3} = \underbrace{l_{MSE}^{s_3}}_{\text{σφάλμα εικόνας}} + \underbrace{6 \cdot 10^{-3} l_{VGG}^{s_3}}_{\text{σφάλμα περιεχομένου}} + \underbrace{10^{-3} l_{G_3}^{s_3}}_{\text{σφάλμα ανταγωνισμού}} + \underbrace{2 \cdot 10^{-8} l_{TV}^{s_3}}_{\text{σφάλμα ολικής διακύμανσης}} \quad (4.12)$$



Σχήμα 4.3: Απεικόνιση της αρχιτεκτονικής του δικτύου του τρίτου σταδίου του TeleGAN. Ο Γεννήτορας του τρίτου σταδίου είναι υπεύθυνος για τη μεγέθυνση της συνθετικής εικόνας του δεύτερου σταδίου, την οποία λαμβάνει ως είσοδο, σε υψηλές διαστάσεις 256x256.

4.4.1 Αρχιτεκτονική

Για τα δίκτυα του γεννήτορα G_3 και του διευκρινιστή D_3 χρησιμοποιήθηκαν παρόμοιες αρχιτεκτονικές με αυτές του [Ledi17]. Ο βασικός κορμός του δικτύου του G_3 αποτελείται από πέντε residual blocks [He16], τα οποία είναι ίδια μεταξύ τους ως προς τη διάταξη των δομικών στοιχείων. Για την αύξηση της ανάλυσης της εικόνας εισόδου χρησιμοποιείται ένα μπλοκ διπλασιασμού διαστάσεων. Επίσης χρησιμοποιούνται δύο συνελκτικά επίπεδα για τον μετασχηματισμό διαστάσεων των χαρτών χαρακτηριστικών. Το ένα από τα δύο βρίσκεται στην αρχή του δικτύου για την κωδικοποίηση της εικόνας εισόδου και ακολουθείται από παραμετρική συνάρτηση διορθωμένης γραμμικής μονάδας (parametric rectified linear unit - ParametricReLU, PReLU) [He15], ενώ το άλλο είναι το τελευταίο συνελκτικό επίπεδο του δικτύου.

Στον διευκρινιστή D_3 , η μείωση των χωρικών διαστάσεων επιτυγχάνεται με τη χρήση μπλοκ μείωσης διαστάσεων. Στο δίκτυο χρησιμοποιούνται τέσσερα τέτοια μπλοκ, τα οποία ακολουθούνται από συγκεντρωτικό επίπεδο μέσης τιμής (Average Pooling), δύο πλήρως συνδεδεμένα επίπεδα, μεταξύ των οποίων χρησιμοποιείται συνάρτηση ενεργοποίησης Leaky ReLU με αρνητική κλίση 0.2 και την τελική σιγμοειδή συνάρτηση ενεργοποίησης που δίνει ως έξοδο την πιθανότητα πρόβλεψης του διευκρινιστή.

4.4.2 Λεπτομέρειες Υλοποίησης

Η διάταξη των residual blocks που χρησιμοποιήθηκε είναι όμοια με αυτή που προτάθηκε από τους Gross και Wilber [Gros16]. Πιο συγκεκριμένα, σε κάθε μπλοκ χρησιμοποιούνται δύο συνελκτικά επίπεδα με πυρήνα 3x3 και μοναδιαίο γέμισμα, το καθένα από τα οποία ακολουθείται από κανονικοποίηση δέσμης. Επίσης, χρησιμοποιούνται PReLU και ReLU ως συναρτήσεις ενεργοποίησης. Η αύξηση της ανάλυσης της εικόνας εισόδου επιτυγχάνεται με τη χρήση ενός εκπαιδευσιμου sub-pixel συνελκτικού επιπέδου [Shi16] στο μπλοκ διπλασιασμού διαστάσεων. Του επιπέδου αυτού προηγείται ένα συνελκτικό επίπεδο με πυρήνα 3x3 και μοναδιαίο γέμισμα, ενώ έπεται συνάρτηση ενεργοποίησης PReLU. Τα δύο συνελκτικά επίπεδα στην είσοδο και στην έξοδο έχουν πυρήνα 9x9 και γέμισμα ίσο με 4.

Τα μπλοκ μείωσης διαστάσεων του διευκρινιστή D_3 αποτελούνται από ένα συνελκτικό επίπεδο με πυρήνα 3x3 και μοναδιαίο γέμισμα και ένα συνελκτικό επίπεδο με πυρήνα 4x4, βήμα ίσο με 2 και μοναδιαίο γέμισμα, το οποίο υποδιπλασιάζει τις χωρικές διαστάσεις. Κάθε ένα από τα δύο συνελκτικά επίπεδα του μπλοκ ακολουθείται από κανονικοποίηση δέσμης και συνάρτηση ενεργοποίησης Leaky ReLU με αρνητική κλίση ίση με 0.2.

4.5 Εκπαίδευση

Για όλα τα σύνολα δεδομένων και σε όλα τα στάδια εκπαίδευσης χρησιμοποιήσαμε αριθμό φίλτρων γεννήτορα $N_g = 128$ και αριθμό φίλτρων διευκρινιστή $N_d = 64$. Στα πρώτα δύο στάδια η διάσταση διανύσματος θορύβου είναι $N_z = 100$ και διάσταση προβαλλόμενου διανύσματος της αναπαράστασης κειμένου $N_t = 128$. Επίσης, εκπαιδεύσαμε τα δίκτυα για 600 εποχές, με χρήση του αλγορίθμου βελτιστοποίησης Adam [King14] με παράμετρο $\beta_1 = 0.5$ και ρυθμό μάθησης (learning rate) 0.0002, τόσο για τον γεννήτορα, όσο και για τον διευκρινιστή. Η παράμετρος κανονικοποίησης λ των Εξισώσεων 4.2 και 4.5 τέθηκε ίση με 1. Ακόμα, χρησιμοποιήθηκαν οι τεχνικές που περιγράψαμε στην Ενότητα 2.5 για βελτιστοποίηση της διαδικασίας εκπαίδευσης. Πιο συγκεκριμένα, χρησιμοποιήθηκε μειούμενος προσθετικός θόρυβος στις εικόνες από την κατανομή των πραγματικών δεδομένων, για να αποφευχθεί η επικράτηση του διευκρινιστή στις πρώτες εποχές εκπαίδευσης. Επίσης εφαρμόστηκε εναλλαγή των ετικετών εκπαίδευσης κάθε εκατό δέσμες δεδομένων, έπειτα από τη δέκατη εποχή εκπαίδευσης.

Στο τρίτο στάδιο η διαδικασία εκπαίδευσης διαφέρει. Για την εκπαίδευση του GAN αύξησης ανάλυσης της εικόνας δε χρησιμοποιήθηκαν διανύσματα θορύβου και μεταβλητές συνθήκης, ενώ χρησιμοποιήθηκε ειδικό κριτήριο βελτιστοποίησης [Led17]. Παράλληλα, η εκπαίδευση του δικτύου αυτού έγινε με τη χρήση ζευγών πραγματικών εικόνων σε διαστάσεις 128x128 και 256x256. Ο γεννήτορας G_3 παίρνει ως είσοδο την πραγματική εικόνα χαμηλών διαστάσεων και παράγει συνθετική υψηλών διαστάσεων, η οποία στη συνέχεια ελέγχεται από την διευκρινιστή μαζί με την πραγματική. Η επιλογή αυτή έγινε επειδή η επιθυμητή ποικιλομορφία στις εικόνες που παράγει ο γεννήτορας G_2 οδηγεί σε αταίριαστα ζεύγη σύνθετων και πραγματικών εικόνων. Συνεπώς, είναι αδύνατο να εκπαιδευτεί το δίκτυο του τρίτου σταδίου με τη χρήση των εικόνων που παράγει ο G_2 . Στην πράξη παρατηρήθηκε ότι ο G_3 επιτυγχάνει στην αύξηση της ανάλυσης των σύνθετων εικόνων του G_2 παρά το γεγονός ότι έχει εκπαιδευτεί σε πραγματικά δεδομένα. Η παρατήρηση αυτή, οδηγεί στο συμπέρασμα ότι οι συνθετικές εικόνες του G_2 προσεγγίζουν την κατανομή των πραγματικών εικόνων.

Ο κώδικας του TeleGAN, μαζί με περισσότερες πληροφορίες για την εκπαίδευση και την αρχιτεκτονική των δικτύων, είναι δημόσια διαθέσιμος στο αποθετήριο <https://github.com/ails-lab/teleGAN>.

Κεφάλαιο 5

Πειραματική Διαδικασία και Αποτελέσματα

Προκειμένου να αξιολογήσουμε το προτεινόμενο μοντέλο, πραγματοποιήσαμε εκτεταμένα πειράματα. Σε πρώτο στάδιο επιδιώξαμε να αναπαράξουμε τα αποτελέσματα των μοντέλων GAN-INT-CLS [Reed16b], StackGAN [Zhan16] και StackGAN++ [Zhan17] για τα σύνολα δεδομένων Oxford-102 [Nils08] και CUB [Wah11]. Στη συνέχεια, εξετάσαμε την επίδραση της χρησιμοποιούμενης διανυσματικής αναπαράστασης κειμένων στην απόδοση των προαναφερθέντων μοντέλων. Πιο συγκεκριμένα, εκτός από τις διανυσματικές αναπαραστάσεις που δημιουργήθηκαν με τη χρήση του char-CNN-RNN [Reed16a], χρησιμοποιήθηκαν διανυσματικές αναπαραστάσεις που παράξαμε με τη χρήση των μοντέλων RoBERTa [Liu19b] και GPT-2 [Radf19]. Έπειτα, πραγματοποιήθηκε η εκπαίδευση του προτεινόμενου μοντέλου TeleGAN στα παραπάνω σύνολα δεδομένων, με σκοπό τη σύγκρισή του με τα υπόλοιπα μοντέλα που εξετάσαμε. Για τη σύγκριση αυτή χρησιμοποιήθηκαν οι μετρικές αξιολόγησης Inception Score (IS) [Sali16] και Fréchet Inception Distance (FID) [Heus18]. Τέλος, εκπαιδεύτηκαν όλα τα μοντέλα που εξετάσαμε στο πιο απαιτητικό σύνολο δεδομένων Flickr8k [Hodo13] και αξιολογήθηκε η επίδοσή τους.

5.1 Σύνολα Δεδομένων

Αξιολογήσαμε την επίδοση του TeleGAN στο πρόβλημα της σύνθεσης εικόνας από κείμενο στα σύνολα δεδομένων Oxford-102 [Nils08] και CUB [Wah11] και Flickr8k [Hodo13]. Το Oxford-102 [Nils08] περιέχει συνολικά 8,189 εικόνες από άνθη που ανήκουν σε 102 διαφορετικές κατηγορίες. Το CUB [Wah11] περιέχει 200 διαφορετικά είδη πτηνών που απεικονίζονται σε 11,788 φωτογραφίες. Επίσης, επειδή το 80% των πτηνών σε αυτό το σύνολο δεδομένων έχουν αναλογία αντικειμένου-εικόνας μικρότερη του 0.5, οι εικόνες περικλύπτονται, σύμφωνα με κουτιά οριοθέτησης που έχουν προκαθοριστεί, αποκτώντας έτσι αναλογία μεγαλύτερη του 0.75. Τέλος, χρησιμοποιείται το πλέον απαιτητικό σύνολο δεδομένων Flickr8k [Hodo13] που περιέχει 8,000 εικόνες επιλεγμένες από 6 ομάδες του Flickr. Η επιλογή των εικόνων έγινε πραγματοποιήθηκε με τέτοιο τρόπο ώστε οι φωτογραφίες να απεικονίζουν μία μεγάλη ποικιλία σκηνών και καταστάσεων και να μην περιέχουν αναγνωρίσιμα πρόσωπα ή τοποθεσίες. Ο Πίνακας 5.1 παρουσιάζει συγκεντρωμένα τα παραπάνω στατιστικά στοιχεία.

Για κάθε εικόνα στα Oxford-102 και CUB περιέχονται 10 λεκτικές περιγραφές, ενώ 5 λεκτικές περιγραφές για κάθε εικόνα υπάρχουν στο Flickr8k. Επίσης, τα δεδομένα που είχαν οριστεί ως δεδομένα επικύρωσης από τους δημιουργούς των συνόλων δεδομένων, χρησιμοποιούνται ως δεδομένα εκπαίδευσης.

Dataset	Oxford-102		CUB		Flickr8k	
Samples	Train	Test	Train	Test	Train	Test
	7,034	1,155	8,855	2,933	6,000	2,000

Πίνακας 5.1: Στατιστικά των συνόλων δεδομένων

5.2 Μετρικές Αξιολόγησης

Ένα γνωστό πρόβλημα που εμφανίζεται κατά την εκπαίδευση των γεννητικών μοντέλων είναι η δυσκολία αξιολόγησης των αποτελεσμάτων τους. Βασικός στόχος είναι τα συνθετικά δεδομένα που παράγονται να είναι ρεαλιστικά, πράγμα το οποίο είναι αδύνατο να αξιολογηθεί δίχως τη συμμετοχή ανθρώπινου παράγοντα. Δύο μετρικές που χρησιμοποιούνται για την εκτίμηση των αποτελεσμάτων των γεννητικών ανταγωνιστικών δικτύων είναι Inception Score (IS) [Sali16] και Fréchet Inception Distance (FID) [Heus18].

5.2.1 Inception Score

Το Inception Score (IS) [Sali16] αποτελεί μία ευρέως χρησιμοποιούμενη μετρική αξιολόγησης των γεννητικών ανταγωνιστικών δικτύων. Η εκτίμηση της ποιότητας των συνθετικών εικόνων βασίζεται στο πόσο καλά μπορεί να τις ταξινομήσει το μοντέλο αναφοράς Inception V3 [Szeg15] στις προκαθορισμένες κλάσεις. Η μετρική αυτή εξετάζει ταυτόχρονα την ποιότητα και την ποικιλία των συνθετικών εικόνων. Για την εκτίμηση της ποιότητας, δημιουργείται για κάθε δείγμα η κατανομή των βαθμολογιών πρόβλεψης (prediction scores) για κάθε κλάση από το Inception V3. Ιδανικά, η κατανομή αυτή θα πρέπει να είναι στενή (narrow distribution), δηλαδή θα πρέπει ο ταξινομητής να μπορεί να αναγνωρίζει με σιγουριά σε ποια κλάση ανήκει το κάθε δείγμα. Για την εκτίμηση της ποικιλίας των συνθετικών εικόνων υπολογίζεται η οριακή κατανομή, αθροίζοντας τις κατανομές των ετικετών που προέκυψαν από την εξέταση της ποιότητας των εικόνων. Αν το σύστημα είναι ικανό να παράξει ένα ευρύ φάσμα από εικόνες, η κατανομή αυτή θα πρέπει να μοιάζει με την ομοιόμορφη. Συνεπώς, συγκρίνοντας την κατανομή ετικετών της κάθε εικόνας με την οριακή κατανομή, υπολογίζεται η εκτίμηση για το Inception Score (IS) του συνόλου δεδομένων. Η σύγκριση μεταξύ των κατανομών πραγματοποιείται με τη χρήση της απόκλισης Kullback–Leibler. Όσο μεγαλύτερη είναι η απόκλιση, τόσο μεγαλύτερο θα είναι το Inception Score. Ο τύπος εκφράζεται μαθηματικά από την Εξίσωση 5.1

$$IS = \exp(\mathbb{E}_x D_{KL}(p(y|x) || p(y))) \quad (5.1)$$

όπου x είναι το συνθετικό δείγμα, y η ετικέτα που προβλέφθηκε από τον ταξινομητή, $p(y)$ η οριακή κατανομή και $p(y|x)$ είναι η κατανομή της ετικέτας του δείγματος x . Για τον υπολογισμό του Inception Score χρησιμοποιήθηκαν 10 σύνολα από 30,000 τυχαίες συνθετικές εικόνες που παρήχθησαν από τον γεννήτορα. Τα αποτελέσματα που παρουσιάζονται στη συνέχεια, αποτελούν το μέσο όρο των τιμών που υπολογίστηκαν για τα προαναφερθέντα 10 σύνολα. Επίσης, προκειμένου τα αποτελέσματα να είναι συγκρίσιμα, χρησιμοποιήθηκαν τα προεκπαιδευμένα μοντέλα InceptionV3 του [Zhan16].

5.2.2 Fréchet Inception Distance

Η Fréchet Inception Distance (FID) [Heus18] είναι μία μετρική που εκτιμά το κατά πόσο οι συνθετικές εικόνες μοιάζουν να είναι πραγματικές. Όπως αναφέρθηκε στο Κεφάλαιο 2, σκοπός του γεννήτορα είναι να μάθει μία κατανομή χαρακτηριστικών βάσει της οποίας να μπορεί παράγει ρεαλιστικά συνθετικά δείγματα. Η FID ουσιαστικά υπολογίζει την απόσταση μεταξύ της κατανομής αυτής του γεννήτορα $p(\cdot)$, και της κατανομής των πραγματικών δεδομένων $p_r(\cdot)$. Υποθέτοντας ότι τα διανύσματα χαρακτηριστικών ακολουθούν πολυδιάστατη γκαουσιανή κατανομή, τότε γκαουσιανή κατανομή με μέση τιμή και συνδιακύμανση (m, C) λαμβάνεται από την $p(\cdot)$ και γκαουσιανή κατανομή με μέση τιμή και συνδιακύμανση (m_r, C_r) λαμβάνεται από την $p_r(\cdot)$. Τότε η απόσταση μεταξύ της συνθετικής και της πραγματικής γκαουσιανής κατανομής υπολογίζεται με τη χρήση της απόστασης Fréchet [Frec57], γνωστή και ως απόσταση Wasserstein-2 [Vase69] (Εξίσωση 5.2)

$$FID = d^2((m, C), (m_r, C_r)) = \|m - m_r\|_2^2 + \text{Tr}(C + C_r - 2(CC_r)^{1/2}) \quad (5.2)$$

όπου Tr το ίχνος (trace) του πίνακα, δηλαδή το άθροισμα των στοιχείων της διαγωνίου του.

Στην πράξη για να υπολογιστεί η FID χρησιμοποιήθηκαν τα διανύσματα χαρακτηριστικών του τελευταίου κρυφού επιπέδου του, προεκπαιδευμένου στο σύνολο δεδομένων ImageNet [Deng09],

μοντέλου InceptionV3. Επίσης, για τον υπολογισμό της FID παρήχθησαν 10 σύνολα από 30,000 τυχαίες συνθετικές εικόνες από τον γεννήτορα. Τέλος, ο υπολογισμός του FID πραγματοποιήθηκε με τη χρήση του επίσημου κώδικα [IoBi17] του [Heus18], προκειμένου τα αποτελέσματά μας να είναι συγκρίσιμα.

5.3 Σύγκριση διανυσματικών αναπαραστάσεων δεδομένων

Στο [Reed16a] γίνεται σύγκριση της επίδοσης του char-CNN-RNN με άλλες τεχνικές μοντελοποίησης δεδομένων στο πρόβλημα της *αναγνώρισης με μηδενική-λήψη* (zero-shot recognition) για τα σύνολα δεδομένων CUB [Wah11] και Oxford-102 [Nils08]. Στην παρούσα διατριβή συγκρίναμε εναλλακτικές μεθόδους δημιουργίας διανυσματικών αναπαραστάσεων των δεδομένων, ως προς την επίδρασή τους στην απόδοση των μοντέλων GAN-INT-CLS [Reed16b] και StackGAN [Zhan16], στο πρόβλημα της σύνθεσης εικόνας από λεκτική περιγραφή. Ως μετρική αξιολόγησης στα συγκεκριμένα πειράματα χρησιμοποιήθηκε το Inception Score [Sal16].

Τα μοντέλα που επιλέχθηκαν για την παραγωγή διανυσματικών αναπαραστάσεων για τις λεκτικές περιγραφές είναι τα RoBERTa [Liu19b] και GPT-2 [Radf19]. Τα δύο αυτά μοντέλα έχουν επιτύχει βέλτιστα αποτελέσματα σε ένα μεγάλο εύρος προβλημάτων του τομέα της επεξεργασίας φυσικής γλώσσας και για το λόγο αυτό επιλέχθηκαν να χρησιμοποιηθούν στο πρόβλημά μας. Για την εξαγωγή των διανυσματικών αναπαραστάσεων πραγματοποιήσαμε προσαρμογή (finetuning) των προεκπαιδευμένων μοντέλων των RoBERTa και GPT-2, που παρέχονται στο [Face18], στις λεκτικές περιγραφές των συνόλων δεδομένων. Έπειτα, προκειμένου να εξάγουμε διανυσματικές αναπαραστάσεις δεδομένων από τα μοντέλα αυτά, ακολουθήσαμε τη διαδικασία που περιγράφεται ως η πιο αποδοτική στο [Pogi19], δηλαδή συνενώσαμε τους χάρτες χαρακτηριστικών των τεσσάρων τελευταίων κρυφών επιπέδων. Η τεχνική αυτή χρησιμοποιείται διότι, όπως αναφέρεται στο [Face18], τα μοντέλα αυτά έχουν προεκπαιδευτεί για συγκεκριμένα προβλήματα, όπως *πρόβλεψη επόμενης πρότασης* (next sentence prediction) και συνεπώς υπάρχει πιθανότητα να είναι προσαρμοσμένα σε αυτούς τους στόχους. Τέλος, κάθε διανυσματική αναπαράσταση που παράγεται από τα παραπάνω δύο μοντέλα αποτελείται από 768 χαρακτηριστικά.

Στον Πίνακα 5.2 απεικονίζονται τα Inception Scores, όπως αυτά υπολογίστηκαν για τις υλοποιήσεις μας των μοντέλων GAN-INT-CLS και StackGAN, που εκπαιδεύτηκαν με τις διανυσματικές αναπαραστάσεις των τριών διαφορετικών μοντέλων: (i) Char-CNN-RNN, (ii) RoBERTa και (iii) GPT-2. Ο Πίνακας 5.3 παρουσιάζει τα αποτελέσματα για την ίδια μετρική, όπως αυτά υπολογίστηκαν στο [Zhan16].

Language Model	Oxford-102		CUB	
	GAN-INT-CLS*	StackGAN*	GAN-INT-CLS*	StackGAN*
Char-CNN-RNN	2.97 ± .03	3.11 ± .02	2.86 ± .04	3.53 ± 0.04
RoBERTa	2.66 ± .03	2.81 ± .03	1.92 ± .02	3.32 ± 0.03
GPT-2	2.43 ± .02	2.78 ± .02	1.74 ± .02	3.48 ± 0.04

Πίνακας 5.2: Inception Scores (IS) των μοντέλων GAN-INT-CLS [Reed16b] και StackGAN [Zhan16] με τη χρήση διαφορετικών διανυσματικών αναπαραστάσεων δεδομένων. (Το * σημαίνει ότι τα αποτελέσματα αυτά υπολογίστηκαν για τα μοντέλα που υλοποιήσαμε και εκπαιδεύσαμε εμείς, με σκοπό την αναπαραγωγή των αποτελεσμάτων των αντίστοιχων άρθρων)

Official Scores	Oxford-102		CUB	
	GAN-INT-CLS	StackGAN	GAN-INT-CLS	StackGAN
Char-CNN-RNN	2.66 ± .03	3.20 ± .01	2.88 ± .04	3.70 ± 0.04

Πίνακας 5.3: Αποτελέσματα Inception Score (IS) για GAN-INT-CLS και StackGAN (οι τιμές που υπολογίστηκαν στο [Zhan16])

Η πρώτη παρατήρηση από τους Πίνακες 5.2 και 5.3 αφορά την αναπαραγωγή αποτελεσμάτων στο μοντέλο StackGAN. Πιο συγκεκριμένα, η υλοποίησή που πραγματοποιήσαμε για τα μοντέλα αυτά, επιτυγχάνει χαμηλότερο IS σε σύγκριση με τα πειράματα των συγγραφέων του αντίστοιχου άρθρου. Συνεπώς, κρίνεται σκόπιμο στη συνέχεια να αναφερόμαστε και στα επίσημα αποτελέσματα, ενώ οι εικόνες/παραδείγματα που παρουσιάζονται έχουν παραχθεί από τη δική μας υλοποίηση. Αντιθέτως, η αναπαραγωγή των αποτελεσμάτων για το GAN-INT-CLS ήταν επιτυχής και μάλιστα με καλύτερη επίδοση, ως προς το IS, για το σύνολο δεδομένων Oxford-102, σε σύγκριση με αυτή που υπολογίστηκε στο [Zhan16]. Επομένως δεν κρίνεται απαραίτητη σε αυτό το μοντέλο η παραπάνω διάκριση των αποτελεσμάτων. Αυτό σημαίνει ότι οι τιμές των μετρικών, καθώς και τα παραδείγματα συνθετικών εικόνων που θα παρουσιαστούν στη συνέχεια, προέκυψαν από τη δική μας υλοποίηση του GAN-INT-CLS.

Σχετικά με τις διανυσματικές αναπαραστάσεις που παράχθηκαν από το μοντέλο RoBERTa παρατηρούμε από τον Πίνακα 5.2 ότι οδηγούν σε μείωση της απόδοσης των μοντέλων GAN-INT-CLS και StackGAN ως προς το IS. Παρά το γεγονός ότι τα συνθετικά δείγματα που παράγονται είναι εξίσου αληθοφανή με αυτά που παράγονται με τη χρήση των διανυσματικών αναπαραστάσεων του Char-CNN-RNN, η ποικιλομορφία των εικόνων μειώνεται, με αποτέλεσμα τη μείωση του Inception Score. Επίσης, παρατηρήσαμε το γεγονός ότι με τη χρήση των αναπαραστάσεων αυτών, μειώθηκε και η συσχέτιση μεταξύ των εικόνων και των λεκτικών περιγραφών, κυρίως ως προς την πληροφορία του χρώματος. Το ίδιο πρόβλημα εντοπίζεται και στην περίπτωση των διανυσματικών αναπαραστάσεων του μοντέλου GPT-2. Συνεπώς, καταλήγουμε στο συμπέρασμα ότι οι διανυσματικές αναπαραστάσεις που παράγονται από το μοντέλο Char-CNN-RNN, είναι οι πιο κατάλληλες για το πρόβλημα της σύνθεσης εικόνας από λεκτική περιγραφή. Ωστόσο, πρέπει να τονίσουμε ότι τα μοντέλα των RoBERTa και GPT-2 έχουν αναπτυχθεί για διαφορετικού είδους προβλήματα του τομέα NLP και ότι παράγουν διανυσματικές αναπαραστάσεις με 768 διαστάσεις χαρακτηριστικών, σε αντίθεση με τις 1024 που παράγει το Char-CNN-RNN.

5.4 Σύγκριση με άλλα δίκτυα

Στον Πίνακα 5.4 συγκρίνουμε αρχικά το TeleGAN, για το πρόβλημα της σύνθεσης εικόνας από λεκτική περιγραφή, με τα μοντέλα GAN-INT-CLS και StackGAN, στα σύνολα δεδομένων CUB και Oxford-102 και για τις μετρικές IS και FID. Πρέπει να τονιστεί πως ο υπολογισμός των δύο μετρικών έγινε με τις συνθετικές εικόνες στις κανονικές τους διαστάσεις. Συνεπώς, στην περίπτωση του GAN-INT-CLS οι μετρικές υπολογίζονται σε εικόνες διαστάσεων 64x64, ενώ σε όλα τα υπόλοιπα δίκτυα 256x256.

Στην περίπτωση του συνόλου δεδομένων Oxford-102 παρατηρούμε ότι το δικό μας μοντέλο, TeleGAN, επιτυγχάνει την καλύτερη απόδοση ως προς τη μετρική IS, η οποία εκτιμά την ποικιλομορφία των συνθετικών δειγμάτων. Πιο συγκεκριμένα, από τον Πίνακα 5.4 παρατηρούμε μία αύξηση 1.25% στο IS για το Oxford-102 σύνολο δεδομένων (από 3.20 σε 3.24). Ταυτόχρονα, όμως, παρατηρείται και αύξηση κατά 5.03 στη μετρική FID, η οποία δεν είναι επιθυμητή καθώς όσο μεγαλύτερη είναι αυτή η τιμή, τόσο μεγαλύτερη είναι και η απόσταση μεταξύ των κατανομών των συνθετικών και των πραγματικών δεδομένων. Στην πράξη οι παρατηρήσεις μας συμβαδίζουν με τις τιμές των μετρικών, καθώς τα συνθετικά δείγματα του μοντέλου μας παρουσιάζουν ποικιλομορφία ως προς τα σχήματα και τα χρώματα, ωστόσο οι αποχρώσεις των εικόνων εμφανίζονται μερικώς ξεθωριασμένες σε ορισμένα παραδείγματα και κατά συνέπεια μεγαλώνουν την απόσταση της κατανομής των συνθετικών δειγμάτων από την πραγματική.

Ως προς το σύνολο δεδομένων CUB, παρατηρήσαμε ότι τα συνθετικά δείγματα του TeleGAN είναι αρκετά αληθοφανή τόσο ως προς τα σχήματα και τα χρώματα των πτηνών, όσο και ως προς τον περιβάλλοντα χώρο που απεικονίζουν. Ωστόσο, υπάρχουν συγκεκριμένες περιπτώσεις κλάσεων στις οποίες το δίκτυό μας δεν παρουσιάζει την ανάλογη ποικιλομορφία με το StackGAN. Επίσης, εντοπίζονται συγκεκριμένα δείγματα με αποτυχημένη αποτύπωση της λεκτικής περιγραφής. Οι παρατηρήσεις αυτές αντικατοπτρίζονται στις τιμές των μετρικών IS και FID του Πίνακα 5.4. Πιο συγκεκριμένα, ενώ

για το δίκτυό μας παρατηρείται μείωση (43.41%) του FID σε λίγο πάνω από το μισό της αντίστοιχης τιμής του StackGAN, υπάρχει ταυτόχρονα μείωση και του IS.

Model	Oxford-102		CUB	
	IS $\uparrow$	FID $\downarrow$	IS $\uparrow$	FID $\downarrow$
GAN-INT-CLS*	$2.97 \pm .03$	61.02	$2.86 \pm .04$	69.34
StackGAN*	$3.11 \pm .02$	65.01	$3.53 \pm .02$	60.34
StackGAN	$3.20 \pm .01$	55.28	$3.70 \pm .04$	51.89
TeleGAN	$3.24 \pm .03$	60.31	$3.47 \pm .03$	29.36

Πίνακας 5.4: Inception Scores (IS) και Fréchet Inception Distance (FID) των δικτύων GAN-INT-CLS, StackGAN και TeleGAN στα σύνολα δεδομένων Oxford-102 και CUB. (Το * σημαίνει ότι τα αποτελέσματα αυτά υπολογίστηκαν για τα μοντέλα που υλοποιήσαμε και εκπαιδεύσαμε εμείς, με σκοπό την αναπαραγωγή των αποτελεσμάτων των αντίστοιχων άρθρων)

Model	IS $\uparrow$
StackGAN++ [Zhan17]	$4.04 \pm .06$
AttnGAN [Xu18]	$4.36 \pm .03$
MirrorGAN [Qiao19]	$4.56 \pm .05$
DM-GAN [Zhu19]	$4.75 \pm .07$

Πίνακας 5.5: Inception Scores ορισμένων από τα state-of-the-art μοντέλα για το σύνολο δεδομένων CUB.

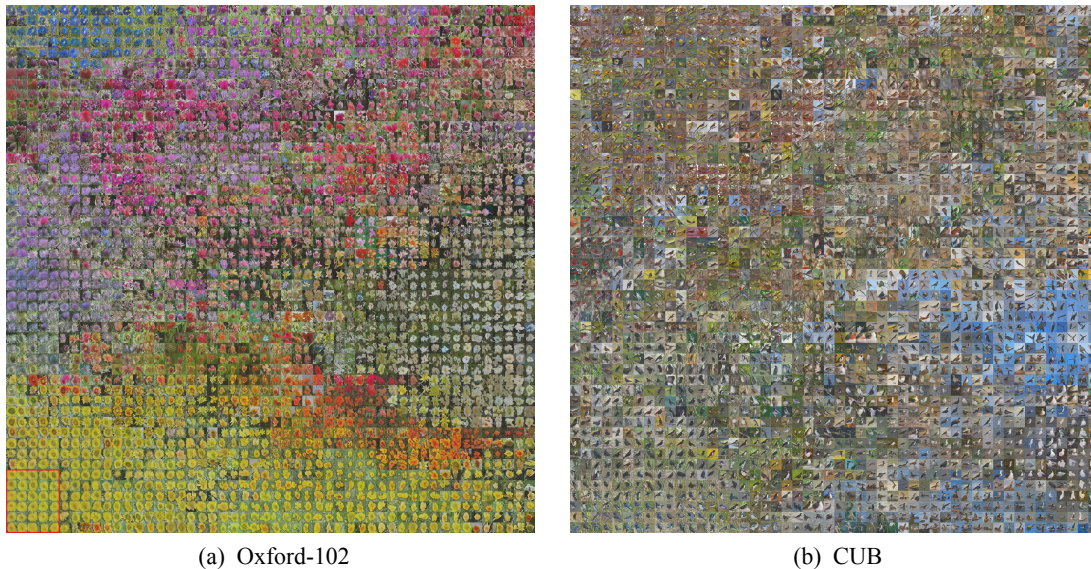
Τα δίκτυα με τα οποία πραγματοποιήσαμε τις παραπάνω συγκρίσεις επιλέχθηκαν επειδή αποτελούν εξίσου απλές υλοποιήσεις, πάνω στις οποίες βασίστηκαν για τη βελτιστοποίηση του προβλήματος τα state-of-the-art μοντέλα [Zhan17, Xu18, Qiao19, Zhu19]. Πιο συγκεκριμένα, στο StackGAN++ [Zhan17] υπάρχουν τρεις στοιβαγμένοι γεννήτορες για τη σύνθεση της τελικής εικόνας, οι οποίοι εκπαιδεύονται ταυτόχρονα με τη χρήση δενδρικής δομής, ενώ στο AttnGAN [Xu18] εφαρμόζεται και η προσθήκη μηχανισμών εστίασης. Επίσης, στο MirrorGAN [Qiao19] χρησιμοποιούνται και πάλι στοιβαγμένοι γεννήτορες, ενώ χρησιμοποιείται ένα πρότυπο σφάλμα που προκύπτει από την αναδημιουργία της λεκτικής περιγραφής από τη σύνθετη εικόνα. Τέλος και στο DM-GAN [Zhu19] χρησιμοποιούνται στοιβαγμένοι γεννήτορες με την προσθήκη πυλών μνήμης για τη διόρθωση σημείων της εικόνας που δεν έχουν αποτυπωθεί σωστά στο προηγούμενο στάδιο. Όπως αναφέρουμε και στην Ενότητα 6, θα είχε ενδιαφέρον η αξιοποίηση των τεχνικών αυτών και στο TeleGAN. Ενδεικτικά στον Πίνακα 5.5 παρουσιάζονται τα Inception Scores των πολύπλοκων αυτών μοντέλων στο σύνολο δεδομένων CUB.

5.5 Έλεγχος για κατάρρευση συστήματος

Για τον έλεγχο πιθανής κατάρρευσης του συστήματος χρησιμοποιήσαμε, σε αντιστοιχία με το [Zhan17], τον αλγόριθμο t-SNE [vdMa08], ο οποίος αποτελεί ένα ιδανικό εργαλείο για την οπτικοποίηση δεδομένων υψηλών διαστάσεων σε ένα δισδιάστατο χάρτη. Στο Σχήμα 5.1 απεικονίζονται δύο πλέγματα, για τα σύνολα δεδομένων Oxford-102 (Σχήμα 5.1a) και CUB (Σχήμα 5.1b), από 50x50 συνθετικές εικόνες που έχουν παραχθεί από το TeleGAN. Οι εικόνες αυτές έχουν συμπίεστεί σε διαστάσεις 50x50 και έχουν τοποθετηθεί στο πλέγμα με τρόπο τέτοιο, ώστε όμοιες μεταξύ τους εικόνες να παρουσιάζονται σε γειτονικές θέσεις.

Προκειμένου να υλοποιήσουμε την παραπάνω διαδικασία, παράξαμε ένα σύνολο από 30,000 συνθετικές εικόνες από το TeleGAN. Στη συνέχεια χρησιμοποιήσαμε το προεκπαιδευμένο στο ImageNet δίκτυο VGG-16 [Simo15], προκειμένου να εξάγουμε χάρτες χαρακτηριστικών για τις εικόνες αυτές. Πιο συγκεκριμένα, χρησιμοποιήθηκαν χάρτες χαρακτηριστικών διάστασης 4,096, που παράγονται από το τελευταίο πλήρως συνδεδεμένο κρυφό επίπεδο («fc2») του δικτύου. Έπειτα, προκειμένου να

μειώσουμε τις διαστάσεις των διανυσμάτων αυτών, χρησιμοποιήσαμε τον αλγόριθμο *Ανάλυσης Κύριων Συνιστωσών* (principal components analysis - PCA), η χρήση του οποίου δεν ήταν αναγκαία, καθώς θα μπορούσαμε να χρησιμοποιήσουμε απευθείας τον αλγόριθμο t-SNE. Ωστόσο, χρησιμοποιήσαμε τον PCA, για γραμμική προβολή των δεδομένων σε ενδιάμεσες διαστάσεις (300), διότι είναι περισσότερο αποδοτικός από άποψη χρόνου εκτέλεσης. Τέλος, τα διανύσματα αυτά, προβλήθηκαν με τη χρήση του t-SNE σε δισδιάστατο χώρο και επιλέχθηκαν τυχαία 2,500 συνθετικές εικόνες για να προβληθούν στο τελικό πλέγμα που παρουσιάζεται στο Σχήμα 5.1. Περισσότερες πληροφορίες και ένα παράδειγμα χρήσης του αλγορίθμου t-SNE για ένα σύνολο από εικόνες μπορούν να βρεθούν στο [Domb19].



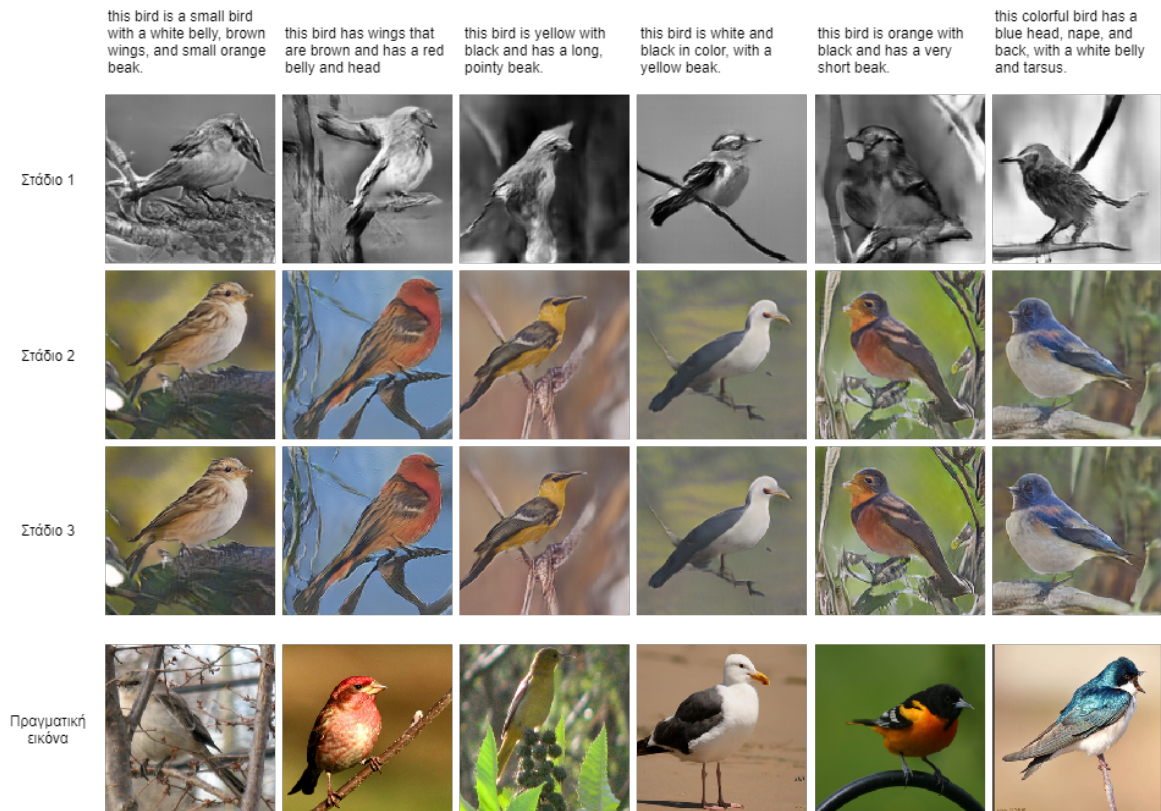
Σχήμα 5.1: Έλεγχος για κατάρρευση συστήματος με τη χρήση του αλγορίθμου t-SNE.

Από το Σχήμα 5.1 παρατηρούμε ότι το μοντέλο μας δεν παρουσιάζει σε κανένα από τα δύο σύνολα δεδομένων το πρόβλημα της κατάρρευσης συστήματος, ως προς την παραγωγή εικόνων δίχως οπτική ουσία. Ωστόσο, στην περίπτωση του Oxford-102 (Σχήμα 5.1a) παρατηρούμε μία ενδεχόμενη κατάρρευση συστήματος ως προς την έλλειψη ποικιλομορφίας εικόνων. Πιο συγκεκριμένα, στο κόκκινο παραλληλόγραμμο εντοπίζονται συνθετικές εικόνες από *ηλιοτρόπια*, τα οποία παρουσιάζουν αρκετές ομοιότητες μεταξύ τους. Η έλλειψη της σχηματικής ποικιλομορφίας που παρουσιάζεται στη συγκεκριμένη κατηγορία λουλουδιών, επηρεάζει αρνητικά το IS του TeleGAN.

Στο Παράρτημα Β.1 παρουσιάζονται επίσης οι χάρτες που παρήχθησαν από τον αλγόριθμο t-SNE για τα συνθετικά δείγματα των δύο συνόλων δεδομένων.

5.6 Παρουσίαση αποτελεσμάτων σε CUB και Oxford-102

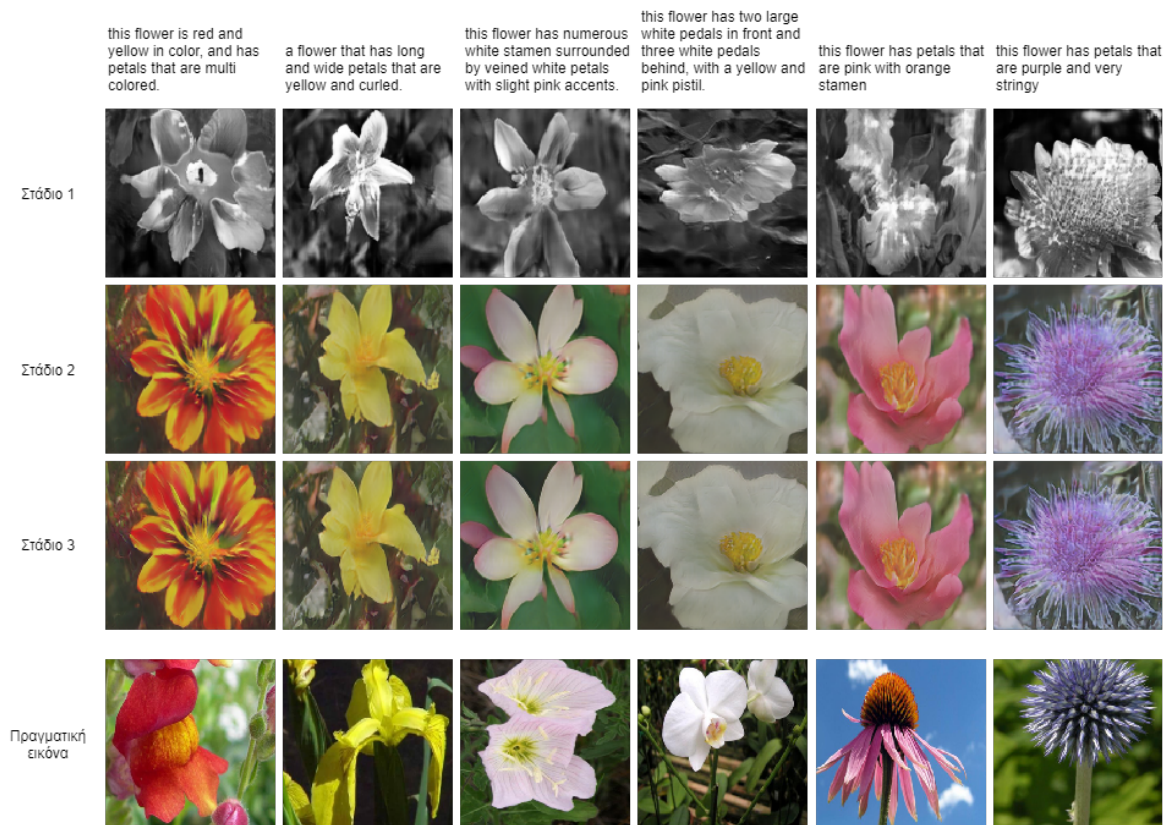
Στη συνέχεια θα παρουσιαστούν ορισμένα παραδείγματα συνθετικών εικόνων του TeleGAN. Πιο συγκεκριμένα, στο Σχήμα 5.2 μπορούμε να δούμε τόσο τις τελικές συνθετικές εικόνες του σταδίου 3, όσο και τις ενδιάμεσες εικόνες που παράγονται από το δίκτυό μας στα στάδια 1 και 2, για έξι δείγματα του συνόλου δεδομένων CUB. Αρχικά, σημασία έχει να σχολιάσουμε την ποιότητα των τελικών συνθετικών εικόνων. Παρατηρούμε ότι οι εικόνες του σταδίου 3 είναι αρκετά αληθοφανείς, πράγμα το οποίο σημαίνει ότι το μοντέλο μας έχει κατορθώσει να προσεγγίσει την κατανομή των πραγματικών δεδομένων. Η παρατήρηση αυτή αποτυπώνεται και στο χαμηλό FID που υπολογίστηκε για το συγκεκριμένο σύνολο δεδομένων, όπως παρουσιάσαμε στην Ενότητα 5.4.



Σχήμα 5.2: Συνθετικές εικόνες του TeleGAN στο σύνολο δεδομένων CUB.

Επιπροσθέτως, έχει σημασία να τονίσουμε την αντιστοιχία μεταξύ των συνθετικών εικόνων και των λεκτικών περιγραφών εισόδου στο TeleGAN, οι οποίες παρουσιάζονται στο πάνω μέρος του σχήματος. Πιο συγκεκριμένα παρατηρούμε ότι πληροφορίες των περιγραφών σχετικά με τα ράμφη (beaks) των πτηνών, όπως «small orange», «long, pointy», «yellow» και «very short» στις εικόνες των στηλών 1, 3, 4 και 5 αντίστοιχα, έχουν αποτυπωθεί στις συνθετικές εικόνες. Το ίδιο συμβαίνει και με τις πληροφορίες των περιγραφών σχετικά με τις αποχρώσεις διαφόρων σημείων του σώματος των πτηνών. Για παράδειγμα στις συνθετικές εικόνες της τελευταίας στήλης παρατηρούμε ότι ο γεννήτορας του δευτέρου σταδίου είναι σε θέση να διακρίνει το κεφάλι (head), τον αυχένα (nape) και την πλάτη (back) από την κοιλιακή χώρα (belly) και το κάτω μέρος (tarsus) του πτηνού κατά τον επιχρωματισμό. Εκτός αυτού, είναι εμφανές ότι κατορθώνει να επιδιορθώσει ορισμένες σχηματικές ατέλειες του γεννήτορα του πρώτου σταδίου.

Επιπλέον, μπορούμε να διακρίνουμε ότι ο γεννήτορας του τελευταίου σταδίου είναι ικανός να μεγεθύνει τη συνθετική εικόνα του δευτέρου σταδίου, δίχως σημαντικές απώλειες στην ποιότητα. Σημασία έχει επίσης να επισημάνουμε ότι οι συνθετικές εικόνες, δεν πρέπει να ταυτίζονται με τις πραγματικές. Πιο συγκεκριμένα, ο γενικός στόχος του συστήματος είναι να είναι σε θέση να παράγει συνθετικές εικόνες που να αντιστοιχούν με τις λεκτικές περιγραφές. Για παράδειγμα, στην τρίτη στήλη εύκολα διακρίνουμε ότι το μοντέλο μας έχει παράξει μία συνθετική εικόνα στην οποία το είδος του πτηνού δεν αντιστοιχεί με αυτό της πραγματικής. Ωστόσο, η συνθετική εικόνα αντιστοιχεί στην λεκτική περιγραφή, η οποία δεν παρέχει τις περαιτέρω απαραίτητες πληροφορίες που απαιτούνται για να παραχθεί πτηνό αντίστοιχου είδους με την πραγματική εικόνα. Τέλος, ένα πρόβλημα των συνθετικών εικόνων που μπορούμε να παρατηρήσουμε είναι ότι τα χρώματα δεν είναι τόσο έντονα όσο στις πραγματικές εικόνες.



Σχήμα 5.3: Συνθετικές εικόνες του TeleGAN στο σύνολο δεδομένων Oxford-102.

Σχετικά με το σύνολο δεδομένων Oxford-102, στο Σχήμα 5.3 παρατηρούμε ότι το TeleGAN είναι ικανό να παράξει ρεαλιστικές εικόνες από λουλούδια, με ποικίλους σχηματισμούς. Ταυτόχρονα, παρατηρούμε ότι η επίδοση του μοντέλου σχετικά με την αντιστοιχία λεκτικών περιγραφών και συνθετικών εικόνων είναι αρκετά υψηλή. Πιο συγκεκριμένα, το μοντέλο μας έχει κατορθώσει να αποτυπώσει πληροφορίες όπως την εμφάνιση χορδής (stringy) των φύλλων της εικόνας στην τελευταία στήλη, καθώς και την πολυχρωμία κίτρινου και κόκκινου στα φύλλα της εικόνας στην πρώτη στήλη. Εντυπωσιακή είναι επίσης η αποτύπωση ιδιομορφιών στους σχηματισμούς των φύλλων. Για παράδειγμα, στη δεύτερη στήλη τα φύλλα απεικονίζονται γυρισμένα (curled) και στην εικόνα της τέταρτης στήλης φαίνεται να έχει αποτυπωθεί η αριθμητική πληροφορία για τα δύο μπροστινά φύλλα και τα τρία στο πίσω μέρος, όπως αναφέρουν οι αντίστοιχες περιγραφές.

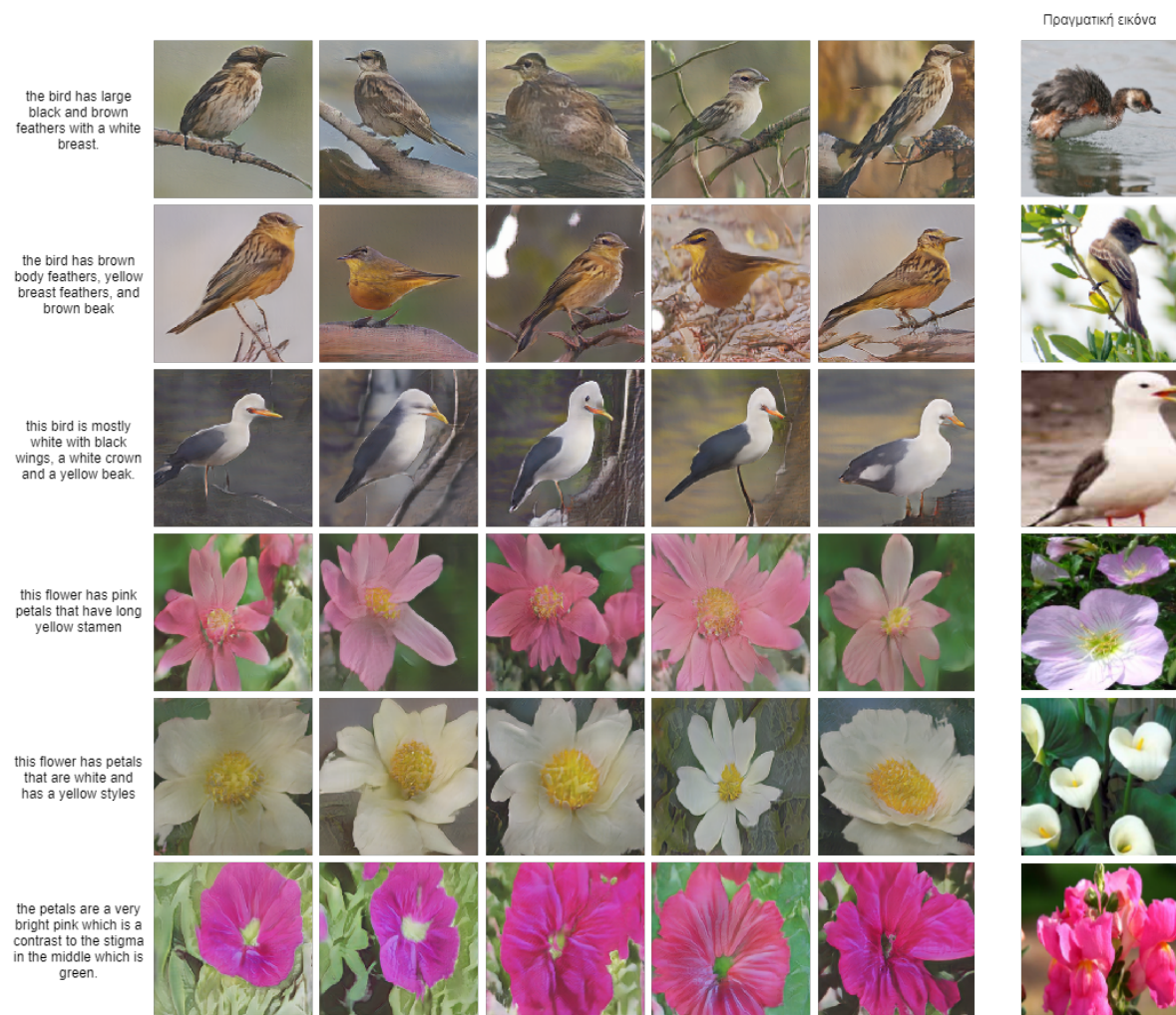
Ακόμα, όμοια με την περίπτωση του συνόλου δεδομένων CUB, παρατηρούμε ότι ο γεννήτορας του τρίτου σταδίου είναι σε θέση να μεγεθύνει τις συνθετικές εικόνες των λουλουδιών, σε διαστάσεις 256x256, δίχως σημαντικές απώλειες. Επιπροσθέτως, παρατηρείται και σε αυτή την περίπτωση το φαινόμενο της διαφοροποίησης μεταξύ του συγκεκριμένου είδους των αντικειμένων που απεικονίζονται στις συνθετικές εικόνες, έναντι των πραγματικών. Στις στήλες 1 και 5, για παράδειγμα, παρατηρούμε ότι η μορφή των λουλουδιών στις συνθετικές εικόνες δεν αντιστοιχεί με αυτή στις πραγματικές. Ωστόσο, επειδή στις αντίστοιχες λεκτικές περιγραφές δεν παρέχονται οι απαραίτητες πληροφορίες για το σχηματισμό των λουλουδιών, οι απεικονίσεις του TeleGAN θεωρούνται επιτυχημένες, δεδομένου ότι έχουν κατορθώσει να αποτυπώσουν τις υπόλοιπες πληροφορίες. Τέλος, και στην περίπτωση του συνόλου δεδομένων Oxford-102 παρατηρούμε ότι οι συνθετικές εικόνες εμφανίζονται λίγο «ξεθωριασμένες» σε σύγκριση με τα έντονα χρώματα των πραγματικών εικόνων.

Γενικότερα, όπως αναφέραμε στην Ενότητα 2.1.2, σκοπός των CGANs είναι να έχουν τη δυνατότητα να παράξουν συνθετικά δείγματα, τα οποία να αντιστοιχούν στις μεταβλητές συνθήκες. Ωστόσο, πρέπει για κάθε διαφορετική τυχαία είσοδο διανύσματος θορύβου το συνθετικό δείγμα να διαφοροποιείται, διατηρώντας ταυτόχρονα την αντιστοιχία με τις μεταβλητές συνθήκες. Αντίστοιχα, το μοντέλο μας θα πρέπει να είναι σε θέση, διατηρώντας σταθερή τη λεκτική περιγραφή εισόδου,

να παράξει διαφορετικές συνθετικές εικόνες που να αντιστοιχούν στη λεκτική περιγραφή, για κάθε διαφορετική είσοδο τυχαίου Γκαουσιανού θορύβου.

Στο Σχήμα 5.4 μπορούμε να παρατηρήσουμε τη δυνατότητα αυτή του TeleGAN. Πιο συγκεκριμένα, στα συνθετικά δείγματα για το σύνολο δεδομένων CUB βλέπουμε ότι τα πτηνά έχουν διαφορετικό σχήμα, κατεύθυνση και περιβάλλοντα χώρο για κάθε διαφορετικό διάνυσμα θορύβου που δόθηκε στο μοντέλο. Ακόμα και στην τρίτη γραμμή του σχήματος, στην οποία εκ πρώτης όψεως τα πτηνά που απεικονίζονται είναι ίδια μεταξύ τους, μπορούμε να παρατηρήσουμε διαφορετικά είδη σωματοτύπων καθώς και διαφορετικό περιβάλλοντα χώρο. Επιπλέον, στο δείγμα της πρώτης γραμμής είναι εμφανές ότι επειδή η περιγραφή είναι γενική, το μοντέλο μας έχει μάθει ότι τα πτηνά που αντιστοιχούν σε αυτή την περιγραφή είναι πιθανό να ανήκουν σε διαφορετικά είδη, γεγονός το οποίο απεικονίζεται στις αντίστοιχες συνθετικές εικόνες.

Σχετικά με τις συνθετικές εικόνες που αντιστοιχούν στο σύνολο Oxford-102 (3 τελευταίες γραμμές του Σχήματος 5.4) παρατηρούμε ότι ενώ η σχηματική ποικιλομορφία είναι μικρότερη, τα συνθετικά δείγματα εξακολουθούν να διαφοροποιούνται μεταξύ τους, για κάθε διαφορετική είσοδο τυχαίου θορύβου, διατηρώντας ταυτόχρονα την αντιστοιχία τους με την εκάστοτε λεκτική περιγραφή. Περισσότερες συνθετικές εικόνες και αποτελέσματα του TeleGAN παρουσιάζονται στο Παράρτημα Β.

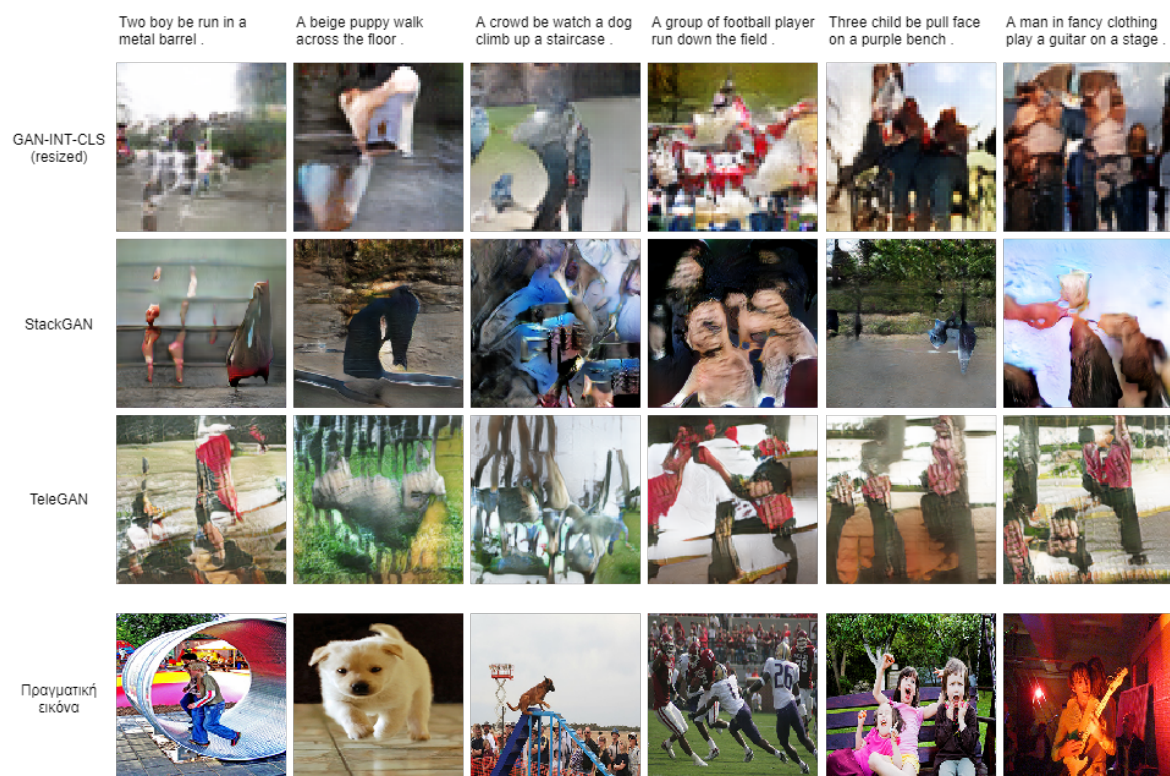


Σχήμα 5.4: Για κάθε μία από τις 6 λεκτικές περιγραφές, έχουν παραχθεί 5 εικόνες από το TeleGAN. Οι 3 πρώτες γραμμές αντιστοιχούν σε δείγματα του συνόλου δεδομένων CUB, ενώ οι 3 τελευταίες του Oxford-102.

5.7 Σύνολο δεδομένων Flickr8k

Όπως αναφέρθηκε στην αρχή του τρέχοντος Κεφαλαίου, το Flickr8k αποτελείται από 8,000 εικόνες που αναπαριστούν πληθώρα σκηνών και καταστάσεων. Η τεράστια αυτή ποικιλομορφία των δειγμάτων καθιστά ιδιαίτερα δύσκολο το έργο των μοντέλων σύνθεσης εικόνας από λεκτική περιγραφή. Επίσης, σε αντίθεση με τα σύνολα δεδομένων CUB και Oxford-102, στα οποία παρέχονται 10 λεκτικές περιγραφές ανά εικόνα, στην περίπτωση του Flickr8k παρέχονται 5. Προκειμένου να εκπαιδεύσουμε τα μοντέλα που εξετάσαμε στο Flickr8k, χρησιμοποιήσαμε το γλωσσικό μοντέλο RoBERTa για την εξαγωγή διανυσματικών αναπαραστάσεων για τις λεκτικές περιγραφές. Πιο συγκεκριμένα, πραγματοποιήσαμε και πάλι τη διαδικασία προσαρμογής (fine-tuning) του προεκπαιδευμένου μοντέλου του [Face18] στις λεκτικές περιγραφές του Flickr8k και εξάγαμε διανύσματα 768 χαρακτηριστικών για κάθε λεκτική περιγραφή.

Κατά την εκπαίδευση τόσο του TeleGAN, όσο και των GAN-INT-CLS και StackGAN στο συγκεκριμένο σύνολο, παρατηρήθηκε αδυναμία προσέγγισης της κατανομής των πραγματικών δεδομένων. Οι συνθετικές εικόνες και των τριών δικτύων δεν έχουν οπτική ουσία, καθώς παράγονται αυθαίρετα σχήματα δίχως κάποια συνοχή (Σχήμα 5.5). Το συμπέρασμά μας είναι ότι υπαίτιο για το αποτέλεσμα αυτό, είναι το μικρό σύνολο δεδομένων εκπαίδευσης που παρέχεται στο Flickr8k. Πιο συγκεκριμένα, σε αυτό περιέχονται μόνο 6,000 δείγματα εκπαίδευσης, έναντι των 7,034 δειγμάτων στο Oxford-102 και των 8,855 δειγμάτων στο CUB. Επιπροσθέτως, έχει σημασία να τονίσουμε ότι ενώ στα δύο τελευταία σύνολα δεδομένων τα δείγματα όχι μόνο μεταξύ των ίδιων κλάσεων, αλλά και συνολικά, παρουσιάζουν ομοιότητες στην κατανομή τους, στο σύνολο Flickr8k οι ομοιότητες αυτές περιορίζονται. Αυτό συμβαίνει διότι οι εικόνες απεικονίζουν τυχαία στιγμιότυπα της καθημερινότητας, τα οποία δεν περιέχουν εντοπισμένα θέματα, και συνεπώς δεν έχουν ενταχθεί σε κλάσεις. Λόγω των αποτελεσμάτων αυτών, δεν κρίναμε σκόπιμο τον υπολογισμό των μετρικών IS και FID γι' αυτό το σύνολο δεδομένων.



Σχήμα 5.5: Συνθετικές εικόνες των GAN-INT-CLS, StackGAN και TeleGAN στο σύνολο δεδομένων Flickr8k.

Κεφάλαιο 6

Συμπεράσματα και Μελλοντικές Κατευθύνσεις

Στα πλαίσια της συγκεκριμένης διατριβής εξετάσαμε διάφορες αρχιτεκτονικές σύνθεσης εικόνας από λεκτικές περιγραφές, καθώς και την επίδραση σύγχρονων μεθόδων εξαγωγής διανυσματικών αναπαραστάσεων κειμένου. Σχετικά με το δεύτερο ζήτημα, παρατηρήσαμε ότι οι διανυσματικές αναπαραστάσεις που παράξαμε με τη χρήση των γλωσσικών μοντέλων RoBERTa και GPT-2, οδηγούν σε χειρότερηση της επίδοσης των μοντέλων GAN-INT-CLS και StackGAN στα σύνολα δεδομένων Oxford-102 και CUB. Ωστόσο, οι διανυσματικές αναπαραστάσεις που χρησιμοποιήθηκαν από αυτά τα μοντέλα είχαν 768 διαστάσεις χαρακτηριστικών, έναντι των 1024 που παρήχθησαν από το μοντέλο Char-CNN-RNN. Συνεπώς, θα ήταν χρήσιμη η επανεκπαίδευση των προαναφερθέντων μοντέλων με σκοπό την παραγωγή διανυσματικών αναπαραστάσεων με περισσότερες διαστάσεις χαρακτηριστικών.

Σε ότι αφορά το TeleGAN, το μοντέλο που προτάθηκε στην παρούσα διατριβή, είναι εμφανές ότι η προσέγγιση του προβλήματος της σύνθεσης εικόνας από λεκτική περιγραφή με τη χρήση στοιβαγμένων δικτύων για επεξεργασία τριών σταδίων (ασπρόμαυρη απεικόνιση, επιχρωματισμός και μεγέθυνση) παρουσιάζει υποσχόμενα αποτελέσματα, παρά το γεγονός ότι βρίσκεται σε βασικό στάδιο υλοποίησης. Παράλληλα, η δυνατότητα ανεξάρτητης εκπαίδευσης των GANs των τριών σταδίων, διευκολύνει την επέμβαση στην αρχιτεκτονική του κάθε σταδίου ξεχωριστά. Πιο συγκεκριμένα, στη διάρκεια των πειραμάτων μας εντοπίσαμε ότι η βελτίωση ενός σταδίου, οδηγεί σε συνολική βελτίωση της τελικής συνθετικής εικόνας του μοντέλου μας.

Ένα πρόβλημα που παρατηρήθηκε στις συνθετικές εικόνες της Ενότητας 5.4 είναι τα ελαφρώς ξεθωριασμένα χρώματά τους σε σύγκριση με τις αντίστοιχες πραγματικές εικόνες. Στο στάδιο του επιχρωματισμού, το GAN που έχουμε υλοποιήσει επεξεργάζεται τις εικόνες στον RGB χρωματικό χώρο. Το μειονέκτημα αυτής της προσέγγισης είναι ότι το μοντέλο μας χρειάζεται να αντιστοιχίσει τρεις τιμές για κάθε εικονοστοιχείο της ασπρόμαυρης συνθετικής εικόνας του πρώτου σταδίου. Μια μελλοντική επέκταση θα μπορούσε να είναι η χρήση του χώρου χρωμάτων CIE-LAB. Το πλεονέκτημα αυτού του χώρου είναι ότι το ένα κανάλι χρώματος αντιστοιχεί στη φωτεινότητα (L) της ασπρόμαυρης εικόνας, η οποία διατηρείται και στην έγχρωμη. Συνεπώς, το GAN του δεύτερου σταδίου του δικτύου μας θα χρειάζεται να προβλέπει τιμές για δύο κανάλια, A και B, έναντι των τριών με τη χρήση του RGB.

Επιπροσθέτως, μία ακόμα προσθήκη που αξίζει να εξεταστεί είναι η χρήση μηχανισμών εστίασης. Στη μέχρι τώρα αρχιτεκτονική του TeleGAN, οι γεννήτορες των δύο πρώτων σταδίων δέχονται ως είσοδο τις διανυσματικές αναπαραστάσεις των λεκτικών περιγραφών των αντικειμένων. Ωστόσο, ανάλογα με το στάδιο, χρειάζεται να δοθεί περισσότερη έμφαση, ή αλλιώς να *εστιάσει* το δίκτυο, σε συγκεκριμένες πληροφορίες της λεκτικής περιγραφής. Πιο συγκεκριμένα, δεδομένου ότι ο γεννήτορας του πρώτου σταδίου θα παράξει ασπρόμαυρη εικόνα, οι πληροφορίες της διανυσματικής αναπαράστασης που αφορούν τα χρώματα δε θα έπρεπε να έχουν την ίδια αξία με τις πληροφορίες που αφορούν το σχήμα του αντικειμένου. Ταυτόχρονα, όμως, δε θα έπρεπε και να εξαλειφθούν, δεδομένου ότι η χρωματική διαφορά μεταξύ σημείων του αντικειμένου καθορίζει τη φωτεινότητα των σημείων της ασπρόμαυρης εικόνας. Η αντίστροφη λογική ισχύει στην περίπτωση του δεύτερου σταδίου που αφορά την επιχρωματισμό της εικόνας. Η πληροφορία του σχήματος στις διανυσματικές αναπαραστάσεις θα έπρεπε να έχει μικρότερη αξία σε σχέση με αυτή του χρώματος.

Επίσης, το TeleGAN θα μπορούσε να προσαρμοστεί με τέτοιο τρόπο, ώστε να αντικαταστήσει

τα στοιβαγμένα GANs πολύπλοκων δικτύων, όπως το MirrorGAN, προκειμένου να εξεταστεί η επίδραση της νέας αυτής προσέγγισης, ως προς τη σύνθεση της εικόνας σε state-of-the-art δίκτυα του συγκεκριμένου προβλήματος. Τέλος, δεδομένου ότι η προσπάθεια εκπαίδευσης του δικτύου μας στο πιο απαιτητικό σύνολο δεδομένων Flickr8K δεν ήταν επιτυχημένη, λόγω του μικρού αριθμού δειγμάτων, θα είχε ενδιαφέρον να εξεταστεί η ικανότητα γενίκευσης του TeleGAN στο σύνολο δεδομένων COCO [Lin15], το οποίο, όμοια με το Flickr8K, αποτελείται από πολλαπλές κατηγορίες αντικειμένων με πληθώρα αναπαραστάσεων. Η διαφορά τους, ωστόσο, έγκειται στο γεγονός ότι το σύνολο δεδομένων COCO περιέχει 80,000 δείγματα εκπαίδευσης, έναντι των 6,000 του Flickr8K.

Βιβλιογραφία

- [Aly05] H. A. Aly and E. Dubois, “Image up-sampling using total-variation regularization with a new observation model”, *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1647–1659, 2005.
- [Arjo17] Martin Arjovsky and Léon Bottou, “Towards Principled Methods for Training Generative Adversarial Networks”, 2017.
- [Assa16] Yannis M Assael, Brendan Shillingford, Shimon Whiteson and Nando De Freitas, “Lipnet: End-to-end sentence-level lipreading”, *arXiv preprint arXiv:1611.01599*, 2016.
- [Barn18] Samuel A. Barnett, “Convergence Problems with Generative Adversarial Networks (GANs)”, 2018.
- [Barr18] Shane Barratt and Rishi Sharma, “A Note on the Inception Score”, 2018.
- [Deng09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database”, in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.
- [Domb19] Micah P. Dombrowski, “Image t-SNE”, <https://nextjournal.com/ml4a/image-t-sne>, 2019. Accessed: 2020-10-22.
- [Dong16] C. Dong, C. C. Loy, K. He and X. Tang, “Image Super-Resolution Using Deep Convolutional Networks”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.
- [Dumo16] Vincent Dumoulin and Francesco Visin, “A guide to convolution arithmetic for deep learning”, 2016.
- [Face18] Hugging Face, “transformers”, <https://github.com/huggingface/transformers>, 2018.
- [Frec57] Maurice Fréchet, “Sur la distance de deux lois de probabilité”, *COMPTES RENDUS HEBDOMADAIRES DES SEANCES DE L ACADEMIE DES SCIENCES*, vol. 244, no. 6, pp. 689–692, 1957.
- [Glor11] Xavier Glorot, Antoine Bordes and Yoshua Bengio, “Deep sparse rectifier neural networks”, in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323, 2011.
- [Good14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio, “Generative Adversarial Networks”, 2014.
- [Gros16] Sam Gross and Michael Wilber, “Training and investigating Residual Nets”, <http://torch.ch/blog/2016/02/04/resnets.html>, 2016. Accessed: 2020-10-10.

- [Han20] Fangda Han, Ricardo Guerrero and Vladimir Pavlovic, “CookGAN: Meal Image Synthesis from Ingredients”, *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar 2020.
- [He15] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”, 2015.
- [He16] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, “Deep Residual Learning for Image Recognition”, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.
- [Heus18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler and Sepp Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium”, 2018.
- [Hinz19] Tobias Hinz, Stefan Heinrich and Stefan Wermter, “Semantic Object Accuracy for Generative Text-to-Image Synthesis”, 2019.
- [Hodo13] Micah Hodosh, Peter Young and Julia Hockenmaier, “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics”, *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 05 2013.
- [Hote33] Harold Hotelling, “Analysis of a complex of statistical variables into principal components.”, *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.
- [IoBi17] Johannes Kepler University Linz Institute of Bioinformatics, “TTUR”, <https://github.com/bioinf-jku/TTUR>, 2017.
- [Ioff15] Sergey Ioffe and Christian Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”, 2015.
- [Isol18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou and Alexei A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks”, 2018.
- [John16] Justin Johnson, Alexandre Alahi and Li Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution”, in *European conference on computer vision*, pp. 694–711, Springer, 2016.
- [Karr19a] Tero Karras, Samuli Laine and Timo Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks”, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.
- [Karr19b] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen and Timo Aila, “Analyzing and Improving the Image Quality of StyleGAN”, *CoRR*, vol. abs/1912.04958, 2019.
- [King13] Diederik P Kingma and Max Welling, “Auto-Encoding Variational Bayes”, 2013.
- [King14] Diederik P. Kingma and Jimmy Ba, “Adam: A Method for Stochastic Optimization”, 2014.
- [Ledi17] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang and Wenzhe Shi, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network”, 2017.

- [Lee09] Honglak Lee, Roger Grosse, Rajesh Ranganath and Andrew Y Ng, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations”, in *Proceedings of the 26th annual international conference on machine learning*, pp. 609–616, 2009.
- [Li19a] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu and Jianfeng Gao, “Object-Driven Text-To-Image Synthesis via Adversarial Training”, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.
- [Li19b] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson and Jianfeng Gao, “StoryGAN: A Sequential Conditional GAN for Story Visualization”, 2019.
- [Lin15] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick and Piotr Dollár, “Microsoft COCO: Common Objects in Context”, 2015.
- [Liu19a] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern et al., “A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis”, *The lancet digital health*, vol. 1, no. 6, pp. e271–e297, 2019.
- [Liu19b] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, 2019.
- [McCu43] Warren S McCulloch and Walter Pitts, “A logical calculus of the ideas immanent in nervous activity”, *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.
- [Mirz14] Mehdi Mirza and Simon Osindero, “Conditional Generative Adversarial Nets”, 2014.
- [Mull19] Rafael Müller, Simon Kornblith and Geoffrey E Hinton, “When does label smoothing help?”, in H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pp. 4694–4703, Curran Associates, Inc., 2019.
- [Naze18] Kamyar Nazeri, Eric Ng and Mehran Ebrahimi, “Image Colorization Using Generative Adversarial Networks”, *Lecture Notes in Computer Science*, p. 85–94, 2018.
- [Nils08] Maria-Elena Nilsback and Andrew Zisserman, “Automated Flower Classification over a Large Number of Classes”, in *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.
- [Pogi19] Andreas Pogiatis, “NLP: Contextualized word embeddings from BERT”, <https://towardsdatascience.com/nlp-extract-contextualized-word-embeddings-from-bert-keras-tf-67ef29f60a7b>, 2019. Accessed: 2020-08-17.
- [Qiao19] Tingting Qiao, Jing Zhang, Duanqing Xu and Dacheng Tao, “MirrorGAN: Learning Text-To-Image Generation by Redescription”, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.
- [Radf15] Alec Radford, Luke Metz and Soumith Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”, 2015.

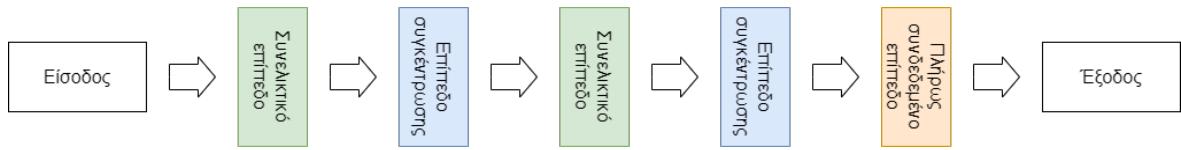
- [Radf19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever, “Language Models are Unsupervised Multitask Learners”, 2019.
- [Reed16a] Scott Reed, Zeynep Akata, Honglak Lee and Bernt Schiele, “Learning Deep Representations of Fine-Grained Visual Descriptions”, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.
- [Reed16b] Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele and Honglak Lee, “Generative Adversarial Text to Image Synthesis”, 2016.
- [Ronn15] Olaf Ronneberger, Philipp Fischer and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation”, in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.
- [Russ15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein et al., “Imagenet large scale visual recognition challenge”, *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [Sali16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen and Xi Chen, “Improved Techniques for Training GANs”, in D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pp. 2234–2242, Curran Associates, Inc., 2016.
- [Shi16] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert and Zehan Wang, “Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network”, 2016.
- [Simo15] Karen Simonyan and Andrew Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, 2015.
- [Suts09] Ilya Sutskever, Geoffrey E Hinton and Graham W Taylor, “The recurrent temporal restricted boltzmann machine”, in *Advances in neural information processing systems*, pp. 1601–1608, 2009.
- [Szeg15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens and Zbigniew Wojna, “Rethinking the Inception Architecture for Computer Vision”, 2015.
- [Turi09] Alan M Turing, “Computing machinery and intelligence”, in *Parsing the turing test*, pp. 23–65, Springer, 2009.
- [Vase69] Leonid Nisonovich Vaserstein, “Markov processes over denumerable products of spaces, describing large systems of automata”, *Problemy Peredachi Informatsii*, vol. 5, no. 3, pp. 64–72, 1969.
- [vdMa08] Laurens van der Maaten and Geoffrey Hinton, “Visualizing Data using t-SNE”, *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.
- [Wah11] C. Wah, S. Branson, P. Welinder, P. Perona and S. Belongie, “The Caltech-UCSD Birds-200-2011 Dataset”, Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
- [Xu18] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang and Xiaodong He, “AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks”, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

- [Zhan16] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang and Dimitris Metaxas, “StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks”, 2016.
- [Zhan17] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang and Dimitris Metaxas, “StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks”, 2017.
- [Zhu19] Minfeng Zhu, Pingbo Pan, Wei Chen and Yi Yang, “DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthesis”, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

Παράρτημα Α

Δομικά στοιχεία Συνελικτικών Νευρωνικών Δικτύων

Τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNNs) αποτελούν μία υποκατηγορία των νευρωνικών δικτύων. Στα δίκτυα αυτά χρησιμοποιείται η μαθηματική πράξη της συνέλιξης σε τουλάχιστον ένα από τα επίπεδα τους. Συνήθως αποτελούνται από το επίπεδο εισόδου, το επίπεδο εξόδου και πολλαπλά ενδιάμεσα κρυφά επίπεδα. Τα κρυφά επίπεδα αποτελούνται από μία σειρά από *συνελικτικά επίπεδα* (convolutional layers) τα οποία συχνά ακολουθούνται από *συγκεντρωτικά επίπεδα* (pooling layers) και *πλήρως συνδεδεμένα επίπεδα* (fully connected layers). Επίσης, είναι πιθανό να χρησιμοποιηθούν επίπεδα *κανονικοποίησης* (normalization). Τέλος, τα CNNs εκπαιδεύονται για να ελαχιστοποιήσουν μια συνάρτηση σφάλματος η οποία ορίζεται βάσει της εξόδου του δικτύου από το τελευταίο πλήρως συνδεδεμένο επίπεδο. Στο Σχήμα Α.1 παρουσιάζεται διαγραμματικά η δομή ενός CNN.



Σχήμα Α.1: Αρχιτεκτονική Συνελικτικού Νευρωνικού Δικτύου

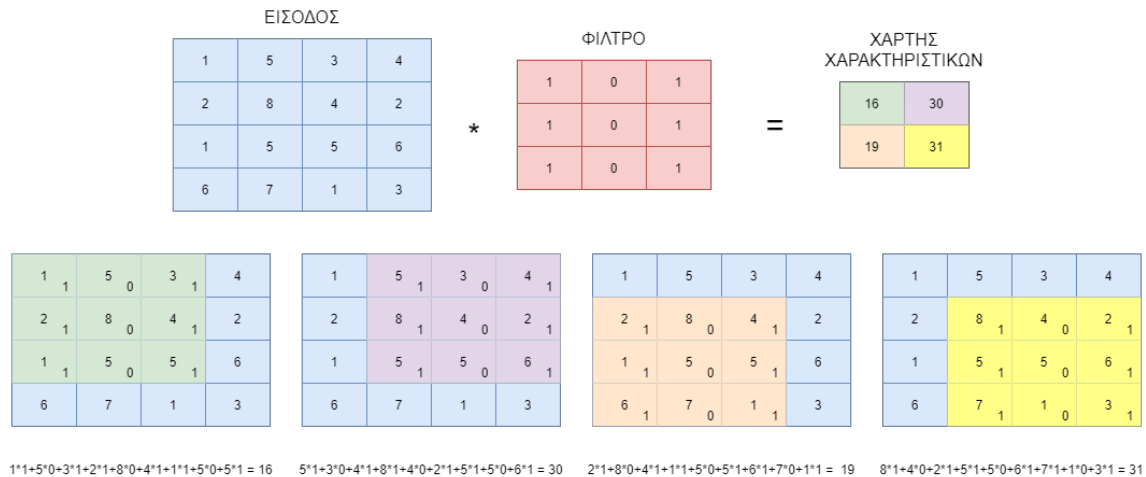
Α.1 Συνελικτικό επίπεδο

Το συνελικτικό επίπεδο είναι το βασικό δομικό στοιχείο της αρχιτεκτονικής των CNNs. Αποτελείται από έναν αριθμό φίλτρων με εκπαιδευσιμες παραμέτρους. Τα φίλτρα αυτά, τα οποία ονομάζονται και *πυρήνες* (kernels), έχουν διαστάσεις μικρότερες από αυτές της εκάστοτε εισόδου. Κατά την *πρόσθια διάδοση* (forward pass) πραγματοποιείται η συνέλιξη καθενός από αυτά τα φίλτρα με την είσοδό τους προκειμένου να παραχθούν *χάρτες χαρακτηριστικών* (feature maps). Το μέγεθος του χάρτη χαρακτηριστικών καθορίζεται από την Εξίσωση Α.1.

$$\frac{W - K + 2P}{S} + 1 \quad (\text{A.1})$$

όπου W είναι η διάσταση της εισόδου, K η διάσταση του φίλτρου, P το μέγεθος του μηδενικού γεμίματος και S το βήμα που χρησιμοποιείται κατά τη συνέλιξη. Το μέγεθος της εξόδου του συνελικτικού επιπέδου υπολογίζεται συνενώνοντας τους χάρτες χαρακτηριστικών όλων των φίλτρων στη διάσταση του βάθους.

Στο Σχήμα Α.2 δίνεται ένα παράδειγμα εφαρμογής της δισδιάστατης συνέλιξης ενός τυχαίου φίλτρου με μία τυχαία είσοδο. Στο συγκεκριμένο παράδειγμα η συνέλιξη πραγματοποιείται με μοναδιαίο βήμα και χωρίς μηδενικό γέμισμα. Η είσοδος έχει διαστάσεις 4×4 και το φίλτρο 3×3 . Συνεπώς με βάση τον παραπάνω τύπο αναμένουμε οι διαστάσεις του χάρτη χαρακτηριστικών να είναι 2×2 . $(\frac{4-3+2 \cdot 0}{1} + 1 = 2)$ [Dumol16].



Σχήμα Α.2: Παράδειγμα εφαρμογής συνέλιξης ενός τυχαίου φίλτρου με μία τυχαία είσοδο.

Α.2 Συγκεντρωτικό επίπεδο

Τα συγκεντρωτικά επίπεδα χρησιμοποιούνται μετά από συνελκτικά επίπεδα λόγω της ανάγκης μείωσης της διάστασης των χαρτών χαρακτηριστικών. Αυτό το επιτυγχάνουν εφαρμόζοντας μια μέθοδο συγκέντρωσης σε υποπεριοχές του χάρτη χαρακτηριστικών, με αποτέλεσμα αυτός να υποδειγματοληπτείται. Πιο συγκεκριμένα ένα κυλιόμενο παράθυρο κινείται πάνω στο χάρτη χαρακτηριστικών προκειμένου, βάσει της μεθόδου συγκέντρωσης που έχει επιλεγεί, να βρεθεί η ζητούμενη ενεργοποίηση της υποπεριοχής. Οι τρεις μέθοδοι συγκέντρωσης που χρησιμοποιούνται συχνότερα είναι:

1. *Συγκέντρωση Μεγίστου* (Max Pooling): Διατηρείται το μέγιστο στοιχείο της περιοχής του παραθύρου ως τιμή δειγματοληψίας.
2. *Συγκέντρωση Ελαχίστου* (Min Pooling): Διατηρείται το ελάχιστο στοιχείο της περιοχής του παραθύρου ως τιμή δειγματοληψίας.
3. *Συγκέντρωση Μέσης Τιμής* (Average Pooling): Υπολογίζεται και διατηρείται η μέση τιμή των στοιχείων της περιοχής του παραθύρου ως τιμή δειγματοληψίας.

Στο Σχήμα Α.3 υπάρχει ένα παράδειγμα εφαρμογής συγκεντρωτικού επιπέδου με τη χρήση της μεθόδου συγκέντρωσης μεγίστου και στο Σχήμα Α.4 υπάρχει το αντίστοιχο παράδειγμα αλλά αυτή τη φορά με τη χρήση της μεθόδου συγκέντρωσης μέσης τιμής [Dumol16].

1	5	3	4
2	8	4	2
1	5	5	6
6	7	1	3

1	5	3	4
2	8	4	2
1	5	5	6
6	7	1	3

8	4
7	6

1	5	3	4
2	8	4	2
1	5	5	6
6	7	1	3

1	5	3	4
2	8	4	2
1	5	5	6
6	7	1	3

Σχήμα Α.3: Παράδειγμα εφαρμογής συγκεντρωτικού επιπέδου με συγκέντρωση μεγίστου. Στο παράδειγμα αυτό το παράθυρο του συγκεντρωτικού επιπέδου είναι 2x2 και χρησιμοποιείται βήμα 2.

1	5	3	4
2	8	4	2
1	5	5	6
6	7	1	3

1	5	3	4
2	8	4	2
1	5	5	6
6	7	1	3

4.0	3.25
4.75	3.75

1	5	3	4
2	8	4	2
1	5	5	6
6	7	1	3

1	5	3	4
2	8	4	2
1	5	5	6
6	7	1	3

Σχήμα Α.4: Παράδειγμα εφαρμογής συγκεντρωτικού επιπέδου με συγκέντρωση μέσης τιμής. Στο παράδειγμα αυτό το παράθυρο του συγκεντρωτικού επιπέδου είναι 2x2 και χρησιμοποιείται βήμα 2.

Α.3 Πλήρως Συνδεδεμένο επίπεδο

Οι νευρώνες ενός πλήρους συνδεδεμένου επιπέδου συνδέονται με όλες τις ενεργοποιήσεις του προηγούμενου επιπέδου. Αυτό σημαίνει ότι κάθε νευρώνας του προηγούμενου επιπέδου συνδέεται με κάθε νευρώνα του επόμενου. Οι έξοδοι των συνελκτικών επιπέδων και των επιπέδων συγκέντρωσης αναπαριστούν χαρακτηριστικά υψηλού επιπέδου (high-level features). Ο σκοπός λοιπόν του πλήρους συνδεδεμένου επιπέδου, για παράδειγμα σε ένα πρόβλημα ταξινόμησης, είναι να χρησιμοποιήσει αυτά

τα χαρακτηριστικά προκειμένου να ταξινομήσει το δείγμα εισόδου σε διάφορες κλάσεις.

A.4 Συναρτήσεις Ενεργοποίησης

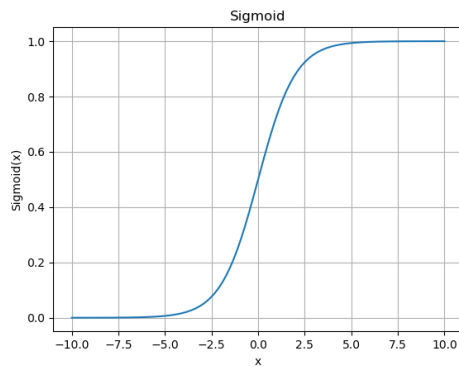
Για να έχουν τα νευρωνικά δίκτυα τη δυνατότητα να προσεγγίσουν οποιαδήποτε συνάρτηση, πρέπει οι *συναρτήσεις ενεργοποίησης* (activation functions) να εισάγουν τον παράγοντα της μη γραμμικότητας (non-linearity) στο δίκτυο.

A.4.1 Σιγμοειδής

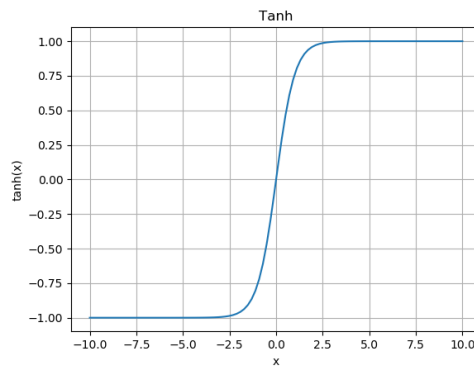
Η *σιγμοειδής συνάρτηση* (sigmoid) είναι μία μαθηματική συνάρτηση της οποίας η καμπύλη έχει τη χαρακτηριστική μορφή «S» και ορίζεται στην Εξίσωση A.2.

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (\text{A.2})$$

Το σύνολο τιμών της σιγμοειδούς συνάρτησης είναι το $(0, 1)$. Η μη-γραμμικότητά της είναι εμφανής στο Σχήμα A.5a, ενώ η κλίση της στις ουρές είναι σχεδόν μηδενική.



(a)



(b)

Σχήμα A.5: (a) Σιγμοειδής συνάρτηση, (b) Υπερβολική εφαπτομένη

A.4.2 Συνάρτηση υπερβολικής εφαπτομένης

Η μη γραμμικότητα της *συνάρτησης υπερβολικής εφαπτομένης* (tanh) ορίζεται στην Εξίσωση A.3

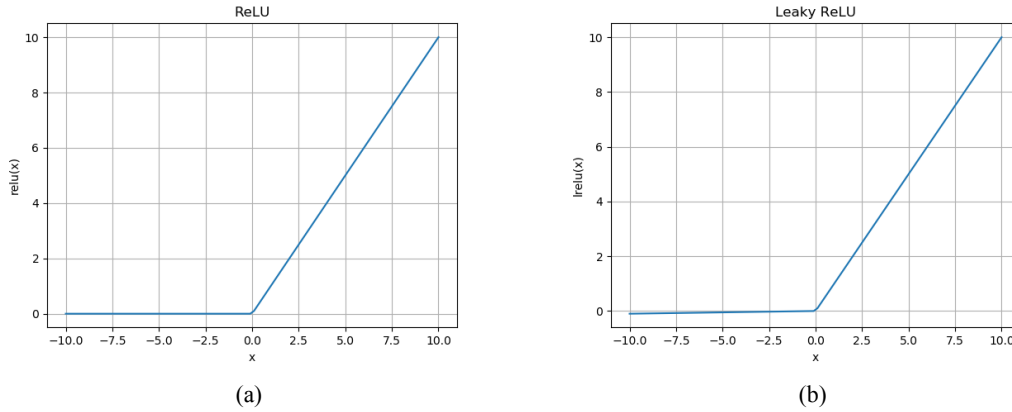
$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (\text{A.3})$$

Όπως παρατηρούμε και στο Σχήμα A.5b μπορεί να θεωρηθεί ως μια επανακλιμάκωση της σιγμοειδούς συνάρτησης ώστε το σύνολο τιμών της πλέον να είναι το $(-1, 1)$.

A.4.3 Συνάρτηση διορθωμένης γραμμικής μονάδας

Η *συνάρτηση διορθωμένης γραμμικής μονάδας* (rectified linear unit - ReLU) (Σχήμα A.6a) έχει γίνει ιδιαίτερα δημοφιλής καθώς έχει εντοπιστεί ότι μικραίνει τον κύκλο εκπαίδευσης επειδή δεν παρουσιάζει κορεσμό στη γραμμικότητά της. Επίσης στο [Glor11] αποδεικνύεται πειραματικά ότι έχει μεγαλύτερη ταχύτητα εκπαίδευσης από τη σιγμοειδή συνάρτηση και την υπερβολική εφαπτομένη. Ωστόσο στη συνάρτηση ReLU είναι πιθανό να εντοπιστεί το πρόβλημα του «θανάτου» κατά την εκπαίδευση. Σε αυτή την κατάσταση η συνάρτηση δίνει την ίδια έξοδο για οποιαδήποτε είσοδο. Ορίζεται από τη σχέση της Εξίσωσης A.4.

$$\text{relu}(x) = \max(0, x) \quad (\text{A.4})$$



Σχήμα A.6: (a) ReLU, (b) Leaky ReLU

A.4.4 Διαρρέουσα συνάρτηση διορθωμένης γραμμικής μονάδας

Η διαρρέουσα συνάρτηση διορθωμένης γραμμικής μονάδας (leaky rectified linear unit - Leaky ReLU) (Σχήμα A.6b) αποτελεί μία προσπάθεια διόρθωσης του προβλήματος «θανάτου» της ReLU. Για να το επιτύχει αυτό, η έξοδος της για τιμές εισόδου μικρότερες του μηδενός δεν είναι μηδενική. Έτσι η Leaky ReLU έχει μια μικρή θετική κλίση για τιμές μικρότερες του μηδενός, όπως φαίνεται στην Εξίσωση A.5.

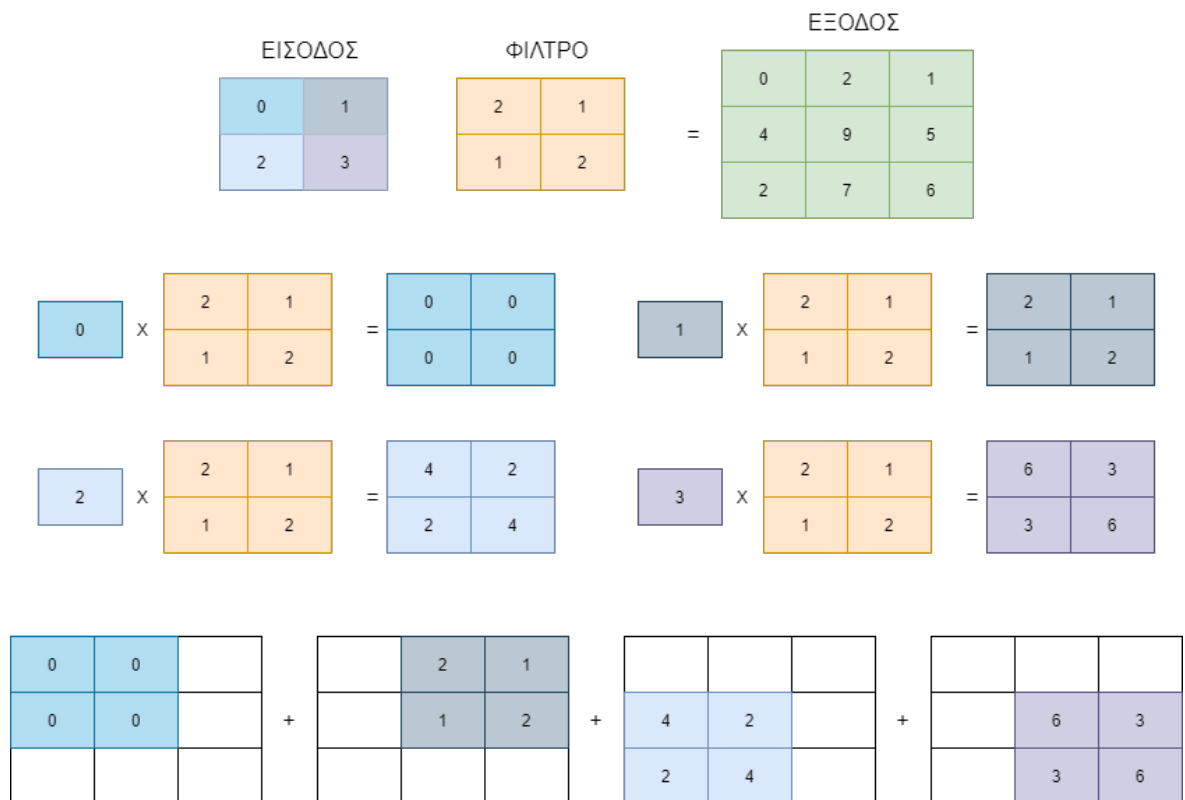
$$lrelu(x) = \begin{cases} x & x > 0, \\ 0.01x & x \leq 0 \end{cases} \quad (A.5)$$

A.5 Κανονικοποίηση δέσμης δεδομένων

Η κανονικοποίηση δέσμης δεδομένων (batch normalization) [Ioff15] προτάθηκε προκειμένου να καταπολεμήσει προβλήματα αστάθειας κατά την εκπαίδευση που προέκυπταν λόγω της αλλαγής της κατανομής των δεδομένων αφού περάσουν από ένα επίπεδο. Η κανονικοποίηση αυτή πραγματοποιείται για κάθε δέσμη δεδομένων προσθέτοντας, παραμέτρους κανονικοποίησης κατά την οπίσθια διάδοση του σφάλματος. Με τη χρήση της έχει αποδειχτεί ότι τα νευρωνικά δίκτυα επιτυγχάνουν γρηγορότερη σύγκλιση, με μεγαλύτερη ακρίβεια.

A.6 Αντιμεταθετικό συνελκτικό επίπεδο

Το αντιμεταθετικό συνελκτικό επίπεδο (transposed convolutional layer) προέκυψε από την ανάγκη ύπαρξης ενός επιπέδου που να πραγματοποιεί την αντίθετη διαδικασία από αυτή του απλού συνελκτικού επιπέδου. Στην πράξη χρησιμοποιείται προκειμένου να υπερδειγματοληπτήσει έναν χάρτη χαρακτηριστικών ώστε να προκύψει ένας νέος χάρτης με μεγαλύτερη χωρική διάσταση. Στο Σχήμα A.7 παρουσιάζεται ένα παράδειγμα βασικής λειτουργίας αυτού του επιπέδου [Dumo16].



Σχήμα Α.7: Παράδειγμα εφαρμογής αντιμεταθετικής συνέλιξης ενός τυχαίου φίλτρου με μία τυχαία είσοδο.

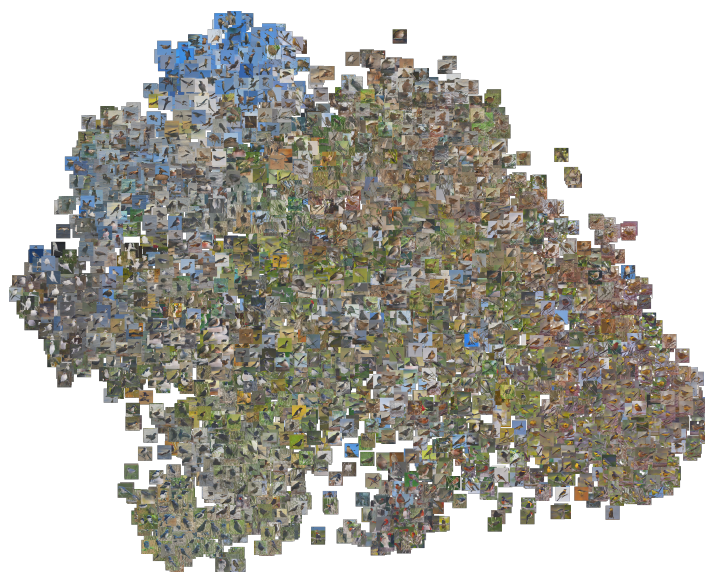
Παράρτημα Β

Επιπλέον αποτελέσματα

Β.1 Απεικόνιση συνθετικών εικόνων στον χώρο



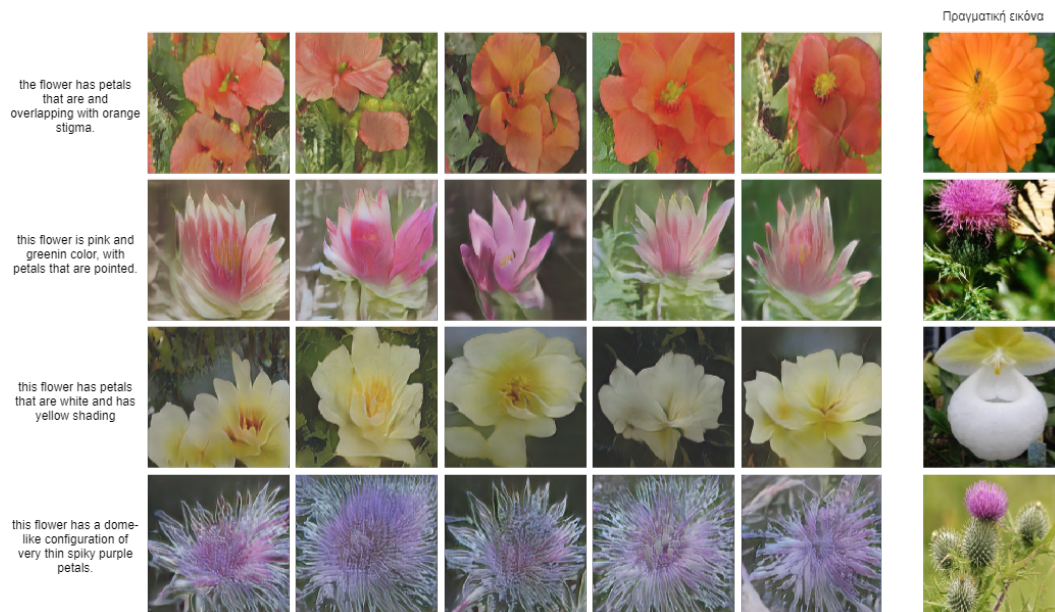
(a) Oxford-102



(b) CUB

Σχήμα Β.1: Απεικόνιση συνθετικών εικόνων του TeleGAN, στα σύνολα δεδομένων Oxford-102 και CUB, σε χάρτη δύο διαστάσεων με τη χρήση του αλγορίθμου t-SNE. Σε κάθε χάρτη απεικονίζονται 5,000 συνθετικές εικόνες.

B.2 Πρόσθετα αποτελέσματα στο Oxford-102



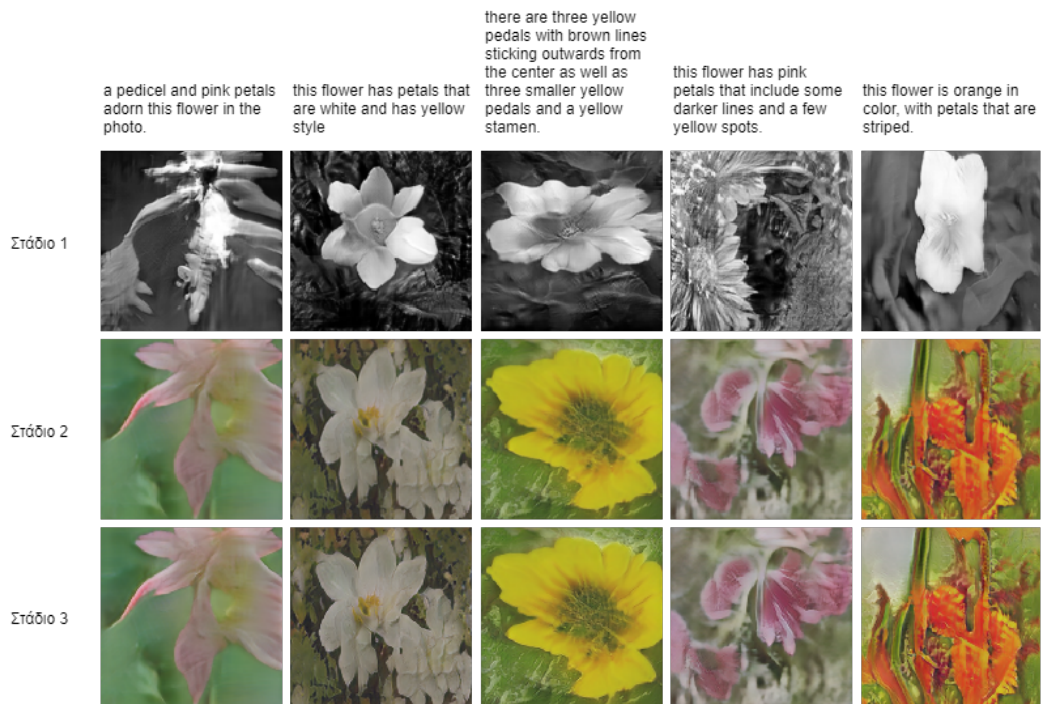
Σχήμα B.2: Συνθετικές εικόνες του TeleGAN στο σύνολο δεδομένων Oxford-102. Για κάθε μία από τις 4 λεκτικές περιγραφές έχουν παραχθεί 5 διαφορετικές συνθετικές εικόνες με τη χρήση διαφορετικού διανύσματος θορύβου στην είσοδο του GAN.

B.3 Πρόσθετα αποτελέσματα στο CUB

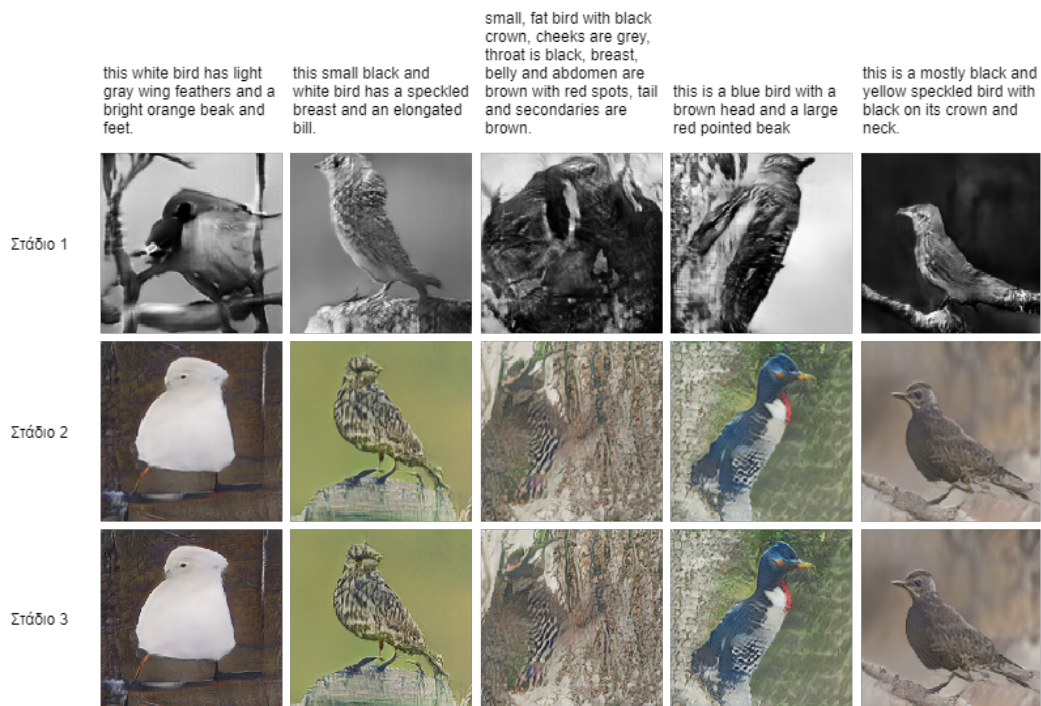


Σχήμα B.3: Συνθετικές εικόνες του TeleGAN στο σύνολο δεδομένων CUB. Για κάθε μία από τις 4 λεκτικές περιγραφές έχουν παραχθεί 5 διαφορετικές συνθετικές εικόνες με τη χρήση διαφορετικού διανύσματος θορύβου στην είσοδο του GAN.

B.4 Περιπτώσεις αποτυχημένης σύνθεσης εικόνας



Σχήμα Β.4: Αποτυχημένες συνθετικές εικόνες του TeleGAN στο σύνολο δεδομένων Oxford-102.



Σχήμα Β.5: Αποτυχημένες συνθετικές εικόνες του TeleGAN στο σύνολο δεδομένων CUB.

Παράρτημα C

Ευρετήριο Ακρωνυμίων και Συντμήσεων

BP: Backpropagation (Οπίσθια Διάδοση)
CA: Conditioning Augmentation (Επαύξηση Συνθήκης)
CGAN: Conditional GAN (GAN υπό συνθήκη)
CNN: Convolutional Neural Network (Συνελκτικό Νευρωνικό Δίκτυο)
CV: Computer Vision (Οραση Υπολογιστών)
DCGAN: Deep Convolutional GAN (Βαθύ Συνελκτικό GAN)
FC layer: Fully Connected layer (Πλήρως Συνδεδεμένο Επίπεδο)
FID: Fréchet Inception Distance
GAN: Generative Adversarial Network (Γεννητικό Ανταγωνιστικό Δίκτυο)
GD: Gradient Descent (Αλγόριθμος Κατάβασης Κλίσης)
IS: Inception Score
MLP: Multilayer Perceptron (Πολυεπίπεδο Perceptron)
MSE: Mean Squared Error (Μέσο Τετραγωνικό Σφάλμα)
NLP: Natural Language Processing (Επεξεργασία Φυσικής Γλώσσας)
PCA: Principal Components Analysis (Ανάλυση Κύριων Συνιστωσών)
PReLU: Parametric ReLU (Παραμετρική ReLU)
ReLU: Rectified Linear Unit (Συνάρτηση Διορθωμένης Γραμμικής Μονάδας)
RNN: Recurrent Neural Network (Ανατροφοδοτούμενο Νευρωνικό Δίκτυο)
SGD: Stochastic Gradient Descent (Αλγόριθμος Στοχαστικής Κατάβασης Κλίσης)
SRGAN: Super Resolution GAN (GAN Υψηλής Ανάλυσης)
t-SNE: t-distributed Stochastic Neighbor Embedding
TV loss: Total Variation loss (Σφάλμα Ολικής Διακύμανσης)
VAE: Variational Autoencoder (Μεταβλητοί Αυτοκωδικοποιητές)