Πανεπιστήμιο Δυτικής Αττικής

Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών

Σύνθεση Εικόνας από Κείμενο με Χρήση Γεννητικών Ανταγωνιστικών Δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΔΗΜΗΤΡΙΟΥ ΗΛΙΑΣ

Επιβλέπων: Αθανάσιος Βουλόδημος

Επ. Καθηγητής ΠΑΔΑ

Αθήνα, Ιούλιος 2021

Πανεπιστήμιο Δυτικής Αττικής

Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών

Σύνθεση Εικόνας από Κείμενο με Χρήση Γεννητικών Ανταγωνιστικών Δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΔΗΜΗΤΡΙΟΥ ΗΛΙΑΣ

Επιβλέπων: Αθανάσιος Βουλόδημος

Επ. Καθηγητής ΠΑΔΑ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Ιουλίου 2021.

Athanasios Voulodimos Digitally signed by Athanasios Voulodimos Date: 2021.07.20 19:43:43 +03'00'	Georgios Bardis Digitally signed by Georgios Bardis Date: 2021.07.20 20:25:51 +03'00'	Panagiota Tselenti Digitally signed by Panagiota Tselenti Date: 2021.07.21 11:59:13 +03'00'
Αθανάσιος Βουλόδημος Επ. Καθηγητής ΠΑΔΑ	Γεώργιος Μπαρδής Επ. Καθηγητής ΠΑΔΑ	Παναγιώτα Τσελέντη ΕΔΙΠ

Αθήνα, Ιούλιος 2021

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος **Ηλίας Δημητρίου** του **Ιωάννη**, με αριθμό μητρώου **151076** φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής **Μηχανικών** του Τμήματος **Μηχανικών Πληροφορικής και Υπολογιστών**, δηλώνω υπεύθυνα ότι:

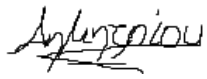
«Είμαι συγγραφέας αυτής της διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο Δηλών

Δημητρίου

.....



.....

Ηλίας Δημητρίου

Σύντομη Περιγραφή

Η σύνθεση εικόνας απο κείμενο αποτελεί ένα αρκετά σύνθετο πρόβλημα, κυρίως του κλάδου της Όρασης Υπολογιστών, με αρκετές πρακτικές εφαρμογές. Βασικός στόχος του αντικειμένου αυτού είναι η δημιουργία εικόνων από ένα μοντέλο, κατόπιν παροχής σε αυτό ορισμένων λεκτικών περιγραφών. Οι παραγόμενες εικόνες πρέπει να είναι υψηλής ποιότητας και συναφείς με τις λεκτικές περιγραφές.

Αρκετές προσεγγίσεις σύνθεσης εικόνας απο κείμενο, έχουν καταφέρει να κατασκευάσουν εικόνες που αντικατοπρίζουν εώς ένα σημείο την σημασία των δοθέντων λεκτικών περιγραφών, αλλά παρουσιάζουν αδυναμίες αναφορικά με την απεικόνιση λεπτομερειών των αντικειμένων που περιγράφονται.

Με την ανάπτυξη των Γεννητικών Ανταγωνιστικών Δικτύων(Generative Adversarial Networks-GAN's), έχει παρατηρηθεί σημαντική βελτίωση αναφορικά με την επίλυση αυτού του προβλήματος, καθώς έχουν αναπτυχθεί τεχνικές που είναι ικανές να παράξουν εικόνες τόσο αληθοφανείς και ταυτόχρονα σχετικές με τις περιγραφές τους που μπορούν να ξεγελάσουν μέχρι και τον άνθρωπο. Αυτές οι τεχνικές περιλαμβάνουν βαθιά συνελικτικούς και επαναλαμβανόμενους κωδικοποιητές κειμένου, που βοηθούν στη μάθηση ορισμένων συναρτήσεων που συσχετίζουν τις εικόνες με λεκτικές περιγραφές και όχι με ετικέτες κλάσεων, όπως είναι σύνηθες. Με αυτό τον τρόπο επιτυγχάνεται μία προσέγγιση θεώρησης της εικόνας και την αντίστοιχης περιγραφής αυτής ως μία οντότητα.

Στη παρούσα διπλωματική εργασία, γίνεται αξιοποίηση υλοποιημένου απο τρίτους κώδικα και αξιολόγηση των αποτελεσμάτων αυτού, μέσω κάποιων μετρικών, προκειμένου να γίνει μια σχετική σύγκριση μεταξύ ορισμένων μοντέλων σύνθεσης εικόνας απο κείμενο που υπάρχουν. Η υλοποίηση αυτή περιλαμβάνει την χρήση του αλγορίθμου CLS-GAN σε συνδυασμό με το StackGAN.

Λεξεις κλειδιά

Σύνθεση εικόνας απο κείμενο, Όραση Υπολογιστών, Γεννητικά Ανταγωνιστικά Δικτυα, CLS-GAN, StackGAN.

Abstract

Text-to-image synthesis is a challenging problem, mostly in the field of Computer Vision, with many practical applications. The basic goal of this research area is the creation of images from a model, after providing it with some text descriptions. The produced images must be of high-quality as well as relevant to the text descriptions.

Many text-to-image approaches have managed to produce images that reflect the meaning of the given text descriptions up to a point, but they still manifest weaknesses regarding the depiction of the described objects details

With the growth of Generative Adversarial Networks, a great improvement has been observed regarding the solution of this problem, since different techniques have been developed, which have proven capable of producing images so plausible and at the same time relevant to the text descriptions, that can fool even humans. These techniques use deep convolutional and recurrent text encoders to learn a correspondence function with images by conditioning the model conditions on text descriptions instead of class labels. In this way, a view that considers the image and the text description as one entity, is achieved.

The main goal of the present thesis, is to use a code developed by a third party, and afterwards evaluate the results of the model, through the use of some metrics, in order to compare them with other existing text-to-image models. This implementation includes the use of the CLS-GAN algorithm along with StackGAN.

Key Words

Text-to-image synthesis, Computer Vision, Generative Adversarial Networks, CLS-GAN, StackGAN.

Ευχαριστίες

Με την ολοκλήρωση της παρούσας διπλωματικής εργασίας θα ήθελα να ευχαριστήσω όλους όσους με υποστήριξαν σε όλη τη διάρκεια των σπουδών μου.

Αρχικά θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της εργασίας, κ. Αθανάσιο Βουλόδημο για την συνεχή στήριξη του σε όλα τα στάδια της εκπόνησης, αλλά και για το γεγονός οτι μου έδωσε την δυνατότητα να διευρύνω τις σπουδές μου εισάγοντας με σε ένα τόσο ενδιαφέρον θέμα. Ακόμα, θα ήθελα να ευχαριστήσω τους καθηγητές, κ. Γεώργιο Μπαρδή και κ. Παναγιώτα Τσελέντη, που με τίμησαν με τη παρουσία τους στην τριμελή επιτροπή εξέτασης.

Επιπρόσθετα, θα ήθελα να ευχαριστήσω τους γονείς μου και τον αδερφό μου που με υποστήριξαν με όποιο τρόπο μπορούσαν στις σπουδές μου. Τέλος, θέλω να ευχαριστήσω όλους τους φίλους μου που ήταν δίπλα μου και δε σταμάτησαν να με ενθαρρύνουν.

Περιεχόμενα

Κεφάλαιο 1

1. Εισαγωγή

Η μελέτη του ανθρώπινου εγκεφάλου αποτελεί αντικείμενο που προβληματίζει το είδος μας απο τα αρχαία χρόνια. Γι'αυτό το λόγο, αρκετοί ερευνητές, από διάφορους επιστημονικούς κλάδους, τις δεκαετίες 1940 και 1950 άρχισαν σταδιακά να συζητούν την πιθανότητα δημιουργίας ενός "τεχνητού εγκεφάλου". Προσωπικότητες όπως οι McCulloch και Pitts, με την περιγραφή της έννοιας του τεχνητού νευρώνα [MCP1943] , καθώς και ο Alan Turing με την αναπτυξη της υπολογιστικής θεωρίας και την δημιουργία του περιβόητου "Turing Test" [AT2009], έμμελε να θέσουν τα θεμέλια για την άνθιση της τεχνητής νοημοσύνης όπως την γνωρίζουμε σήμερα.

Η ανάπτυξη της τεχνητής νοημοσύνης έχει πλέον οδηγήσει σε δημιουργία αλγορίθμων και συστημάτων που είναι ικανά να εκτελέσουν διαδικασίες που εκτελούν οι άνθρωποι, παρέχοντας μάλιστα και καλύτερα αποτελέσματα από αυτούς. Αυτό φυσικά δεν παύει να ισχύει και για συστήματα του κλάδου της όρασης των υπολογιστών, τα οποία πλέον είναι σε θέση να πραγματοποιούν διαδικασίες όπως για παράδειγμα η αναγνώριση αντικειμένων, εντοπισμός ασθενειών από ακτινογραφίες ασθενών, βελτιστοποίηση οδηγικής εμπειρίας μέσω συστημάτων ανίχνευσης λωρίδων κ.ο.κ

1.1 Κίνητρο Ενασχόλησης

Η αναγνώριση αντικειμένων που απεικονίζονται σε εικόνες είναι κάτι που συστήματα όρασης υπολογιστών έχουν πραγματοποιήσει με μεγάλη επιτυχία. Παρόλα αυτά όμως, αυτό δεν οφείλεται στο γεγονός ότι τα συστήματα αυτά είναι σε θέση να κατανοησουν τον οπτικό κόσμο και τα χαρακτηριστικά αυτού. Αν μπορούσαν να κάνουν κάτι τέτοιο τότε τα συστήματα αυτά θα μπορούσαν να λύσουν το πρόβλημα της δημιουργίας συνθετικών εικόνων.

Αυτή ακριβώς την δυνατότητα, εως ένα βαθμό, δηλαδή την κατανόηση του οπτικού κόσμου και την δημιουργία συνθετικών εικόνων, έχουν τα γεννητικά μοντέλα. Πιο συγκεκριμένα, τα Γεννητικά Ανταγωνιστικά Δικτυα [GOO2014] είναι μοντέλα που έχουν την ικανότητα κατανόησης του οπτικού κόσμου και έχουν συμβάλει καθοριστικά στο κομμάτι της σύνθεσης εικόνας, παράγοντας μάλιστα σε ορισμένες περιπτώσεις εικόνες πανομοιότυπες με πραγματικές. Χαρακτηριστικό παράδειγμα αυτού αποτελούν οι εικόνες του Σχήματος 1.1.

Σχήμα 1.1:Συνθετικές εικόνες μοντέλου StackGAN-v2(Πηγή: https://github.com/hanzhanggit/StackGAN-v2)

Εφόσον λοιπόν τα Γεννητικά Ανταγωνιστικά Δικτύα μπορούν να δημιουργήσουν εικόνες τόσο υψηλού επιπέδου, διευρύνουμε το πρόβλημα της απλής σύνθεσης εικόνας σε σύνθεση εικόνας από κείμενο. Δίνοντας δηλαδή σε ένα μοντέλο μια λεκτική περιγραφή, προχωράει στη σύνθεση μιας όσο το δυνατόν πιο ρεαλιστικής εικόνας. Μια τέτοια επέκταση του προβλήματος έχει σαφέστατα πληθώρα πιθανών εφαρμογών, όπως είναι για παράδειγμα η σχεδίαση προσώπου δράστη απο περιγραφή μάρτυρα, η δημιουργία χαρακτήρων για ταινίες ή βιντεοπαιχνίδια, η σχεδίαση προϊόντων κατόπιν περιγραφής αυτών απο ενδιαφερόμενους πελάτες/καταναλωτές κ.α

Τα εξαιρετικά αποτελέσματα που παρέχουν τα συγκεκριμένα μοντέλα σε συνδυασμό με τη πληθώρα εφαρμογών που διαθέτουν, αποτελούν το βασικό κίνητρο για την εκπόνηση της παρούσας διπλωματικής εργασίας και ενδεχομένως τη μελλοντική ενασχόληση με το αντικείμενο.

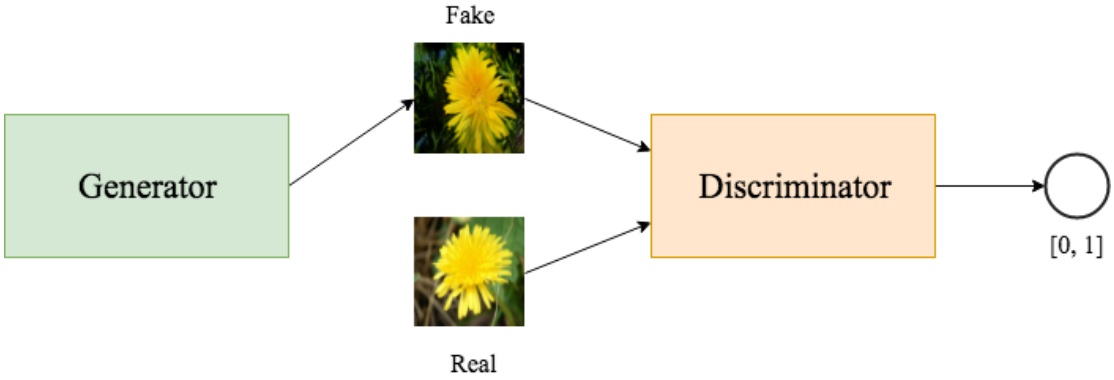
1.2 Δομή Διπλωματικής

Η παρούσα εργασία αποτελείται συνολικά από 7 κεφάλαια συμπεριλαμβανομένης και της εισαγωγής(1ο Κεφάλαιο). Έτσι έχουμε τα εξής:

- Στο 2ο Κεφάλαιο παρέχεται όλη η απαραίτητη πληροφορία αναφορικά με τα Γεννητικά Ανταγωνιστικά Δικτυα.
- Στο 3ο Κεφάλαιο γίνεται μια παρουσίαση του προβλήματος σύνθεσης εικόνας απο κείμενο μαζί με ορισμένα state-of-the-art μοντέλα που επιλύουν το πρόβλημα.
- Στο 4ο Κεφάλαιο γίνεται αναλυτική περιγραφή του μοντέλου που χρησιμοποιήθηκε στη παρούσα διπλωματική.
- Στο 5ο Κεφάλαιο γίνεται μια παρουσιάση των μετρικών αξιολόγησης που χρησιμοποιήθηκαν για την αξιολόγηση των συνθετικών εικόνων καθώς και το σύνολο δεδομένων που αξιοποιήθηκε.
- Στο 6ο Κεφάλαιο γίνεται παρουσίαση των αποτελεσμάτων του μοντέλου και σύγκριση αυτών με τα αποτελέσματα άλλων μοντέλων με βάση τις μετρικές.
- Τέλος στο 7ο Κεφάλαιο γίνεται αναφορά σε ορισμένα συμπεράσματα.

2. Γεννητικά Ανταγωνιστικά Δίκτυα

Τα ***Γεννητικά Ανταγωνιστικά Δίκτυα(GAN)*** έκαναν την εμφάνιση τους το 2014 με τη δημοσίευση του έργου [GOO2014] από τον Ian Goodfellow και τους συνεργάτες του. Η βασική ιδέα του μοντέλου στηρίζεται σε ένα παιχνίδι μεταξύ δύο οντοτήτων, του ***γεννήτορα*** (generator) και του ***διευκρινιστή*** (discriminator). Ο *γεννήτορας* απο τη μεριά του, παράγει ένα σύνολο εικόνων και επιδιώκει να ξεγελάσει τον *διευκρινιστή* κάνοντας τον να νομίζει πως οι εικόνες είναι πραγματικές και όχι συνθετικές. Απο την άλλη ο *διευκρινιστής*, δοθείσας μίας εικόνας, επιδιώκει να την αναγνωρίσει ως προς την αυθεντικότητα της. Διαισθητικά, το παιχνίδι αυτό μεταξύ των δύο οντοτήτων παίζεται επαναλαμβανόμενα, έχοντας σαν αποτέλεσμα την σύνθεση όλο και πιο ρεαλιστικών εικόνων απο τον *γεννήτορα*.(Σχήμα 2.1)



Σχήμα 2.1: Μακροσκοπική αναπαράσταση λειτουργίας των GANs(Πηγή:[CB2018])

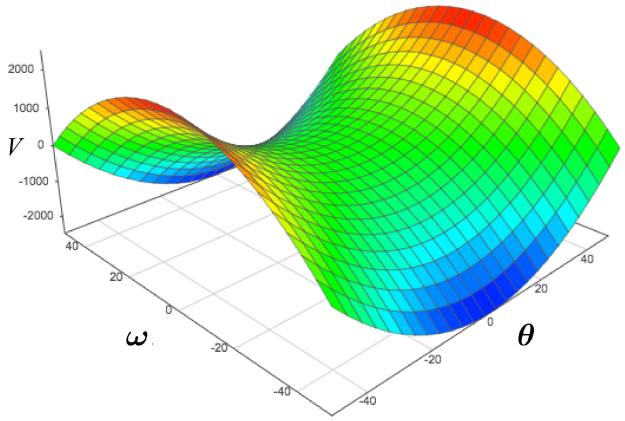
2.1 Μαθηματική προσέγγιση αρχιτεκτονικής

Έχοντας ως στόχο την εύρεση της κατανομής p_g του *γεννήτορα* πάνω σε ένα σύνολο δεδομένων x, γίνεται ο καθορισμός μιας κατανομής θορύβου $p_z(z)$. Έπειτα η συσχέτιση στο χώρο των δεδομένων εκφράζεται ως $G(z; \theta_g)$, όπου το G είναι μια συνάρτηση διαφοροποίησης που αντιπροσωπεύεται απο ένα πολυστρωματικό perceptron (MLP) με παραμέτρους θ_g. Επιπρόσθετα, καθορίζεται και ένα δεύτερο πολυστρωματικό perceptron $D(x; \theta_d)$ που στην έξοδο του δίνει ένα διάνυσμα μίας διάστασης. Το $D(x)$ αντιπροσωπεύει την πιθανότητα το x να προέρχεται απο πραγματικά δεδομένα και όχι απο την κατανομή p_g του *γεννήτορα*, δηλαδή ψεύτικα δεδομένα.

Ο *διευκρινιστής*, εκπαιδεύται προκειμένου να μεγιστοποιηθεί η πιθανότητα ανάθεσης της σωστής ετικέτας τόσο στα δεδομένα εκπαίδευσης όσο και στα δείγματα του *γεννήτορα*. Η εκπαίδευση του *γεννήτορα* γίνεται ταυτόχρονα, με στόχο την ελαχιστοποίηση του $log(1 - D(G(z)))$. Εν ολίγοις, ο *γεννήτορας* και ο *διευκρινιστής* παίζουν ένα παιχνίδι μεγιστοποιήσης-ελαχιστοποίησης της συνάρτησης τιμής $V(G,D)$ που περιγράφεται από τον τύπο (2.1)(όπως δίνεται στο [GOO2014]) :

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

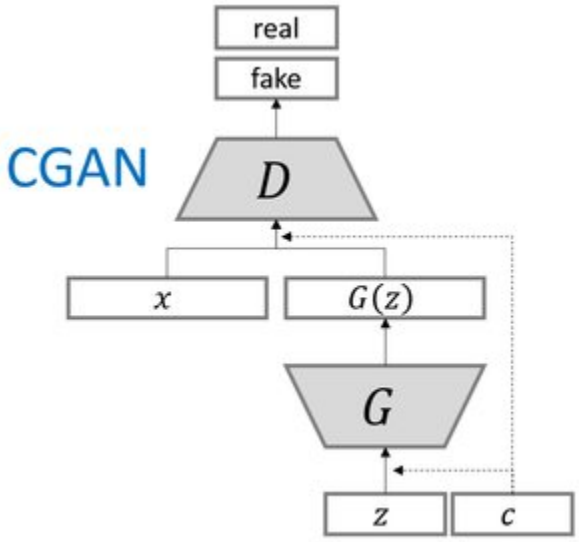
Σε μία ιδανική περίπτωση το παιχνίδι που περιγράφει η σχέση (2.1), θα έφτανε σε μία ισορροπία, ή σύμφωνα με τη θεωρία παιγνίων σε ένα ***Nash Equilibrium***, σε εκείνο το σημείο του χώρου όπου το V είναι ελάχιστο σε σχέση με την τιμή του G και μέγιστο σε σχέση με την τιμή του D. Σε εκείνο το σημείο ο *γεννήτορας* αλλά και ο *διευκρινιστής* δεν θα πραγματοποιούσαν καμμία αλλαγή στις παραμέτρους τους.



Σχήμα 2.2: Σημείο ισοροπίας Nash της συνάρτησης V. Το θ αντιστοιχεί στον *γεννήτορα* και το w στον *διευκρινιστή*(Πηγή:[CB2018])

2.2 Υπό συνθήκη Γεννητικό Ανταγωνιστικό Δίκτυο

Εάν ο *γεννήτορας* αλλά και ο *διευκρινιστής* ενός GAN τροφοδοτηθούν επιπρόσθετα με κάποιου είδους πληροφορίας y, τότε επιτυγχάνεται η επέκταση του μοντέλου που αναλύθηκε στην ενότητα 2.1 του παρόντος κεφαλαίο. Αυτή η επέκταση αποτελεί ένα **υπό συνθήκη GAN** (Σχήμα 2.3). Η επιπρόσθετη αυτή πληροφορία y ονομάζεται ***μεταβλητή συνθήκης*** και μπορεί να έχει την μορφή ετικετών, εικόνων ή οποιαδήποτε άλλη μορφή. Η ένταξη της μεταβλητής συνθήκης στο GAN, μπορεί να γίνει εύκολα με την δημιουργία ενός επιπλέον στρώματος στο δίκτυο τόσο στον *γεννήτορα* όσο και στον *διευκρινιστή*.



Σχήμα 2.3: Αρχιτεκτονική του GAN υπό συνθήκη. Η μεταβλητή συνθήκης απεικονίζεται με το $\boldsymbol{c}$(Πηγή: [GS2018])

Ο *γεννήτορας* πλεόν, σαν είσοδο δέχεται ενωποιημένη την κατανομή θορύβου $\boldsymbol{p_z(z)}$ μαζί με τη μεταβλητή συνθήκης $\boldsymbol{y}$. Την ίδια ακριβώς είσοδο λαμβάνει και ο *διευκρινιστής*, με την μόνη διαφορά να έγκειται στο γεγονός οτι η συνένωση της μεταβλητής συνθήκης γίνεται με το δείγμα εισόδου του *διευκρινιστή*.

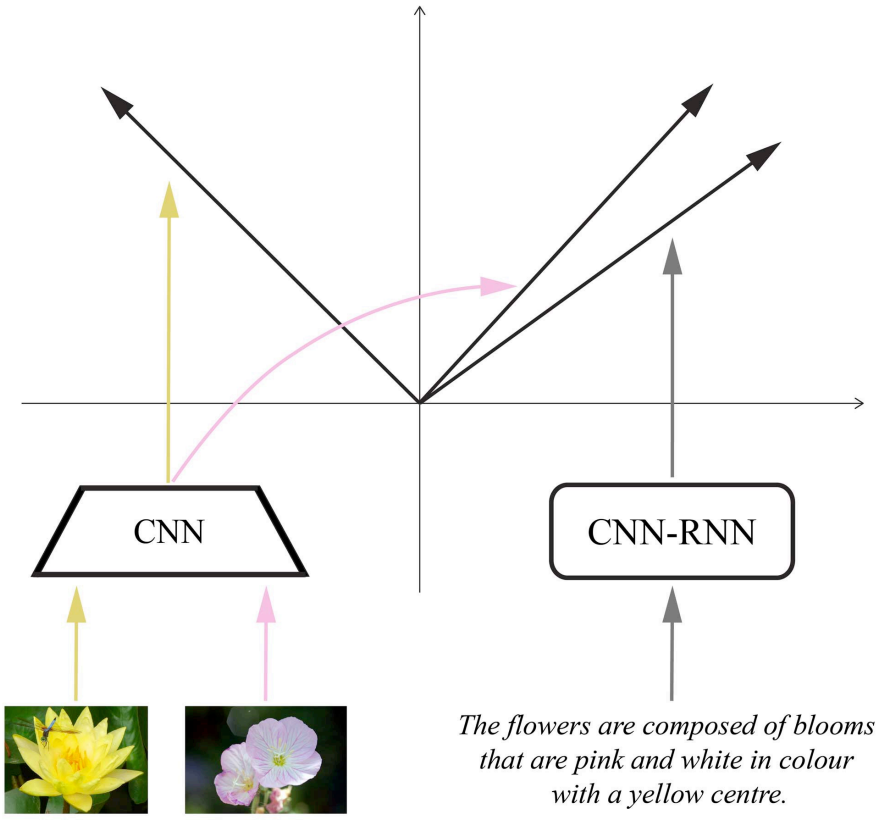
Έτσι, συμφωνα με τα παραπάνω, και όπως αναφέρεται και στο [MM2014], η σχέση (2.1) αλλάζει και παίρνει τη μορφή που δίνεται στη σχέση (2.2) :

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

2.3 Διανυσματική αναπαράσταση κειμένου

Οι **λεκτικές περιγραφές** για να μπορέσουν να ενσωματωθούν σε κάποιο μοντέλο πρέπει αρχικά να αναπαρασταθούν υπο **μορφή διανύσματος**. Η διαδικασία κατα την οποία γίνεται μετατροπή των περιγραφών αυτών σε διανύσματα ονομάζεται **ενσωμάτωση κειμένου** (text embedding).

Ο υπολογισμός των διανυσματικών αναπαραστάσεων πραγματοποιείται από ένα κωδικοποιητή char-CNN-RNN ο οποίος έχει προταθεί στο [SR2016]. Ουσιαστικά, ο κωδικοποιητής ταιριάζει τις λεζάντες με τις αντίστοιχες εικόνες σε ένα κοινό χώρο αναπαράστασης όπου τα διανύσματα των εικόνων και των λεζάντων που ταιριάζουν έχουν μεγαλύτερο εσωτερικό γινόμενο.(Σχήμα 2.4)



Σχήμα 2.4:Ενσωματώσεις κειμένου. Αν η εικόνα ταιριάζει με την λεκτική περιγραφή τότε τα διανύσματα τους είναι πιο κοντά(Πηγή: [GS2018])

Στη διαδικασία του ταιριάσματος, το Συνελικτικό Νευρωνικό Δίκτυο(CNN) επεξεργάζεται τις εικόνες και ένα υβριδικό Συνελικτικό-Αναδρομικό Νευρωνικό Δίκτυο (CNN-RNN) μετασχηματίζει τις λεκτικές περιγραφές.

2.4 Ορισμένα Συχνά προβλήματα

Τα GANs έχουν εμφανίσει ορισμένα προβλήματα, εκ των οποίων κάποια ακόμα ταλαιπωρούν ερευνητές με την επίλυση τους. Παρόλα αυτά όμως, έχουν γίνει διάφορες προτάσεις λύσεων οι οποίες έχουν βοηθήσει σε κάποιες περιπτώσεις.

2.4.1 Εκλειπόμενες κλίσεις

Σε ορισμένες περιπτώσεις Γεννητικά Ανταγωνιστικών Δικτύων, ο *διευκρινιστής* γίνεται πολύ 'ισχυρός' αρκετά γρήγορα. Αυτό οδηγεί σε μία σαφέστατη επικράτηση του *διευκρινιστή* έναντι του *γεννήτορα*, προκαλώντας το πρόβλημα των ***εκλειπόμενων***

κλίσεων (vanishing gradients). Επί της ουσίας, το πρόβλημα αυτό σημαίνει ότι ο *διευκρινιστής* δεν παρέχει αρκετές πληροφορίες (feedback) στον *γεννήτορα* με αποτέλεσμα ο δεύτερος να μη μαθαίνει και να βελτιώνει την απόδοση του. Μία λύση σε αυτό το πρόβλημα προτάθηκε στο [GOO2014] , όπου αναφέρεται πως είναι πιο αποτελεσματική η χρήση της συνάρτησης σφάλματος $max(log(D))$ έναντι της της $min(log(1-D))$ στον *γεννήτορα*.

2.4.2 Κατάρρευση Συστήματος

Συνήθως, είναι επιθυμητό το εκάστοτε GAN να είναι ικανό να παράξει με ευρεία ποικιλία εξόδων. Για παράδειγμα, σε ένα GAN που παράγει εικόνες προσώπων, θα ήταν επιθυμητό να λαμβάνεται στην έξοδο διαφορετικό πρόσωπο για κάθε τυχαία είσοδο που δίνεται.

Ωστόσο, εάν ο *γεννήτορας* παράξει στην έξοδο του κάτι αρκετά αληθοφανές, τότε ενδέχεται να μάθει να παράγει ***μόνο*** τέτοιου τύπου εξόδους. Αυτό οφείλεται στο γεγονός οτι ο *γεννήτορας* πάντα επιδιώκει την δημιουργία εξόδων που θα φαίνονται αληθοφανείς στο *διευκρινιστή*.

Επομένως, έαν ο *γεννήτορας* παράγει συνεχώς την ίδια έξοδο, ο *διευκρινιστής* θα υιοθετήσει μια στρατηγική απόρριψης αυτής της εξόδου. Σε περίπτωση όμως που στην επόμενη επανάληψη του δικτύου ο διευκρινιστής 'κολλήσει' σε κάποιο τοπικό ελάχιστο και δεν μπορεί να βρεί την καλύτερη δυνατή στρατηγική, τότε είναι πολύ εύκολο κάθε επόμενη επανάληψη να υπερπροσαρμοστεί στη τρέχουσα κατάσταση του διευκρινιστή, χωρίς να υπάρχει κάποια τρομερή διαφορά στα δείγματα. Αυτό το πρόβλημα 'παγίδευσης' του *διευκρινιστή* σε ίδια δείγματα ονομαζεται ***κατάρρευση συστήματος.***

Ένας τρόπος που έχει επιλυθεί αυτό το πρόβλημα, είναι η χρήση μίας συνάρτησης απώλειας στον *γεννήτορα* που ενσωματώνει τις τρέχουσες ταξινομήσεις του *διευκρινιστή* με μελλοντικές ταξινομήσεις του προκειμένου να μην οδηγείται ο *γεννήτορας* σε υπερ-βελτίωση του δείγματος που έχει ξεγελάσει μία έκδοση του *διευκρινιστή*.

3. Σύνθεση εικόνας από κείμενο με χρήση GANs

Ένα απο τα πιο απαιτητικά προβλήματα της Επεξεργασίας Φυσικής Γλώσσας και της Όρασης Υπολογιστών, αποτελεί η απόδοση λεζάντας/λεκτικής περιγραφής σε κάποια εικόνα (image captioning), δηλαδή δοθείσας μίας εικόνας, επιδιώκεται η παραγωγή μιάς σχετικής λεκτικής περιγραφής. Την αντίστροφη ακριβώς διαδικασία, πραγματοποιεί όπως αναφέρθηκε και στο 1ο Κεφάλαιο, η σύνθεση εικόνας από κείμενο(text to image synthesis), δηλαδή δοθείσας μιας λεζάντας/λεκτικής περιγραφής, επιδιώκεται η παραγωγή μιας σχετικής εικόνας.

Τα δύο αυτά προβλήματα, παρουσιάζουν μια αρκετά όμοια συμπεριφορά αναφορικά με την ποικιλία των πιθανών αποτελεσματων τους. Αυτό μπορεί να γίνει πολυ εύκολα κατανοητό με ένα παράδειγμα. Αν δοθεί σε ένα υπολογιστικό μοντέλο μία εικόνα και του ζητηθεί να την περιγράψει λεκτικά, τότε μπορεί να δώσει πολλές διαφορετικές λύσεις. Αντίστοιχα και στο πρόβλημα σύνθεσης εικόνας απο λεκτική περιγραφή, ένα αντίστοιχο υπολογιστικό μοντέλο θα καταλήξει σε πολλά διαφορετικά αποτελέσματα. Ωστόσο, μια πολύ σημαντική λεπτομέρεια ανάμεσα σε αυτά τα δύο προβλήματα αλλάζει ραγδαία το βαθμό δυσκολίας επίλυσης τους. Αυτή η λεπτομέρεια, είναι η δυνατότητα αντιμετώπισης του προβλήματος απόδοσης λεζάντας σε εικόνα με ένα πιο ακουλουθιακό τρόπο, έχοντας σε κάθε στάδιο σαν 'στήριγμα' τις προηγούμενες λέξεις που έχουν καθοριστεί στη λεκτική περιγραφή. Το γεγονός έλλειψης αυτής της δυνατότητας στο πρόβλημα σύνθεσης εικόνας από κείμενο, το καθιστά αρκετά δυσκολότερο σε σχέση με την απόδοση λεζάντας σε εικόνα.

Η σύνθεση εικόνας από τη φυσική γλώσσα έχει σαφέστατα πολλές πιθανές εφαρμογές σε μελλοντικές γενιές, όταν φυσικά η τεχνολογία θα είναι σε θέση να τις υποστηρίξει.

3.1 State-of-the-art μοντέλα

Από την εμφάνιση των Γεννητικών Ανταγωνιστικών Δικτύων και έπειτα, πολλοί ερευνητές πρότειναν διάφορα μοντέλα επίλυσης του προβλήματος σύνθεσης εικόνας από κείμενο, με ορισμένα από αυτά τα μοντέλα να παρέχουν εξαιρετικά αποτελέσματα. Στη παρούσα διπλωματική θα γίνει μία σύντομη παρουσίαση ορισμένων μοντέλων που θεωρούνται state-of-the-art, χωρίς όμως να παρέχονται περισσότερες λεπτομέρειες αναφορικά με την λειτουργία τους.

3.1.1 GAN-CLS [CLS2016]

Το μοντέλο GAN-CLS αποτέλεσε την πρώτη προσπάθεια επίλυσης του προβλήματος σύνθεσης εικόνας από κείμενο. Ο πιό άμεσος και απλός τρόπος εκπαίδευσης ενός υπο-συνθήκη GAN, είναι η προσέγγιση ζευγών εικόνας-κειμένου σαν ενιαίες παρατηρήσεις και η εκπαίδευση του *διευκρινιστή* να κρίνει τα ζεύγη ως αληθινά ή συνθετικά. Αυτό ακριβώς κάνει και το μοντέλο GAN-CLS. Αρχικά, ο *διευκρινιστής* δεν έχει σαφή αίσθηση εάν οι πραγματικές εικόνες εκπαίδευσης ταιριάζουν με το περιεχόμενο των διανυσματικών αναπαραστάσεων κειμένου.

Λαμβάνοντας αυτό κατά νού, στο GAN-CLS, επιπρόσθετα των αληθινών και ψεύτικων εισόδων στον *διευκρινιστή* κατά την διαδικασία της εκπαίδευσης, γίνεται είσοδος και ενός τρίτου τύπου εισόδου που αποτελείται από πραγματικές εικόνες με λεκτικές περιγραφές που δεν είναι συναφείς με τις εικόνες αυτές, και τις οποιές ο *διευκρινιστής* πρέπει να μάθει να αναγνωρίζει ως ψεύτικες. Έτσι, με την βελτιστοποίηση του ταιριάσματος εικόνας-κειμένου σε συνδυασμό με την αληθοφάνεια της εικόνας, ο *διευκρινιστής* μπορεί να δώσει στον *γεννήτορα* επιπλέον πληροφορίες.

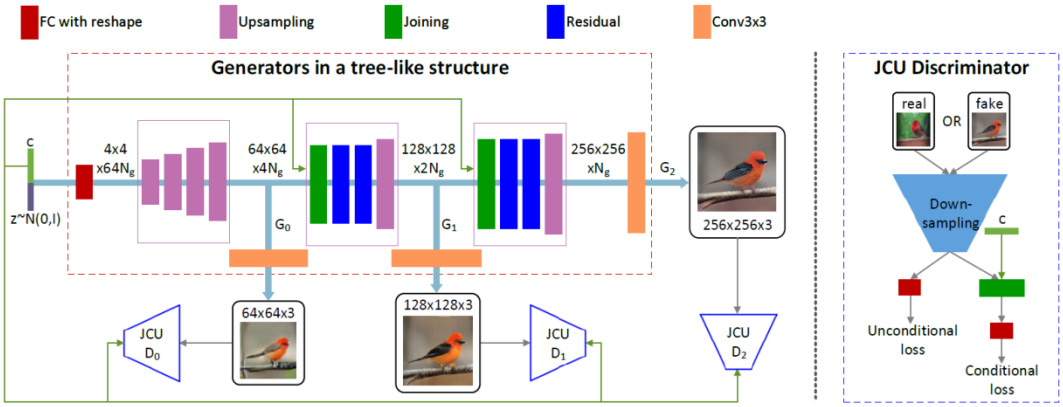
3.1.2 StackGAN [HZ2016]

Ακολουθώντας μία διαφορετική προσέγγιση η οποία στόχευε στη παραγωγή εικόνων υψηλής ανάλυσης με ρεαλιστικές λεπτομέρειες, οι συγγραφείς του StackGAN πρότειναν μία αρχιτεκτονική, όπου η διαδικασία σύνθεσης εικόνων από κείμενο χωρίζεται σε δύο στάδια. Το πρώτο στάδιο, σχεδιάζει κάποια βασικά σχήματα και χρώματα ενός αντικειμένου καθώς και του φόντου από ένα τυχαίο δίανυσμα θορύβου, δίνοντας έτσι μία πρώτη εικόνα χαμηλής ανάλυσης. Το δεύτερο στάδιο απο τη μεριά του, λαμβάνει ως είσοδο την εικόνα που δημιούργησε το πρώτο στάδιο, και διορθώνει τα σφάλματα που υπάρχουν και σχεδιάζει πιο αποτελεσματικά τις λεπτομέρειες του εκάστοτε αντικειμένου, διαβάζοντας και πάλι την λεκτική περιγραφή που αντιστοιχεί στην εικόνα, δημιουργώντας έτσι μία πιο αληθοφανή και υψηλότερης ανάλυσης εικόνα.

Στο σημείο αυτό είναι σημαντικό να αναφερθεί πως τα δύο προαναφερθέντα μοντέλα, παρουσιάζονται αναλυτικά στο 4ο Κεφάλαιο, καθώς ο συνδυασμός τους αποτελεί το εξεταζόμενο μοντέλο της παρούσας διπλωματικής εργασίας.

3.1.3 StackGAN++ [HZ2018]

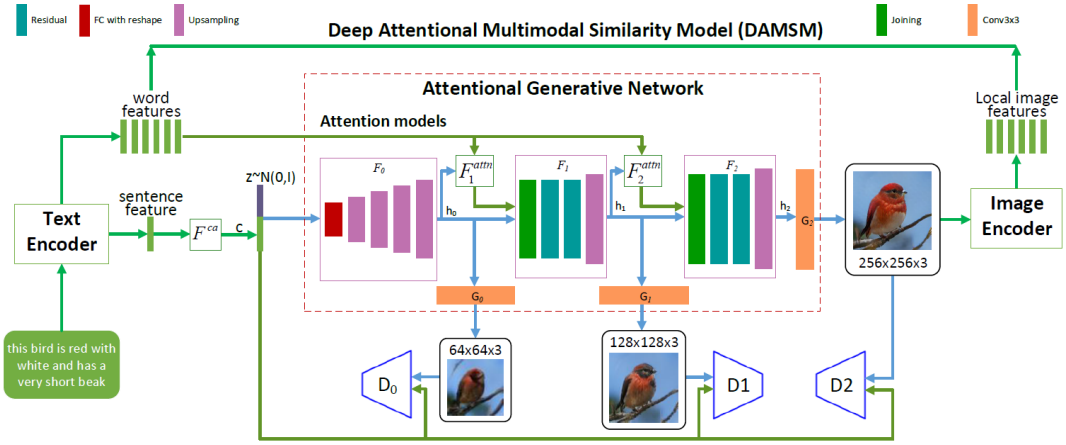
Επεκτείνοντας την ιδέα του StackGAN, οι ίδιοι συγγραφείς δημιούργησαν ένα προχωρημένο πολυεπίπεδο Γεννητικό Ανταγωνιστικό Δίκτυο που αποτελείται από πολλαπλούς *γεννήτορες* και *διευκρινιστές* σε μία δενδρική δομή. Η αρχιτεκτονική, συνθέτει εικόνες σε διάφορες κλίμακες για την ίδια σκηνή. Οι δοκιμές έχουν δείξει πως αυτή η νέα αρχιτεκτονική που έχει προταθεί υπερέχει κατά πολύ των άλλων state-of-the-art μονέλων στην σύνθεση ρεαλιστικών εικόνων.



Σχήμα 3.1:Μακροσκοπική αναπαράσταση αρχιτεκτονικής StackGAN++ (Πηγή: [HZ2018])

3.1.4 AttnGAN [TX2017]

Το AttnGAN αποτελείται από μία αρχιτεκτονική παρόμοια με αυτή του StackGAN++, με τη μόνη διαφορά να παρατηρείται στην προσθήκη ενός *μοντέλου εστίασης*(attention model) πάνω στην υπάρχουσα αρχιτεκτονική. Το μοντέλο εστίασης αναπαράγει τον μηχανισμό εστίασης του ανθρώπου και επιτρέπει στο δίκτυο να επικεντρωθεί σε μία λέξη από μια πρόταση ή ένα τμήμα κειμένου, που αφορά μία εικόνα, τη φορά. Με αυτό το τρόπο εξασφαλίζεται ένα πιο ακριβές ταίριασμα σε επίπεδο εικόνας-λέξης και όχι επίπεδο πρότασης-εικόνας όπως συνηθίζεται σε άλλα μοντέλα.

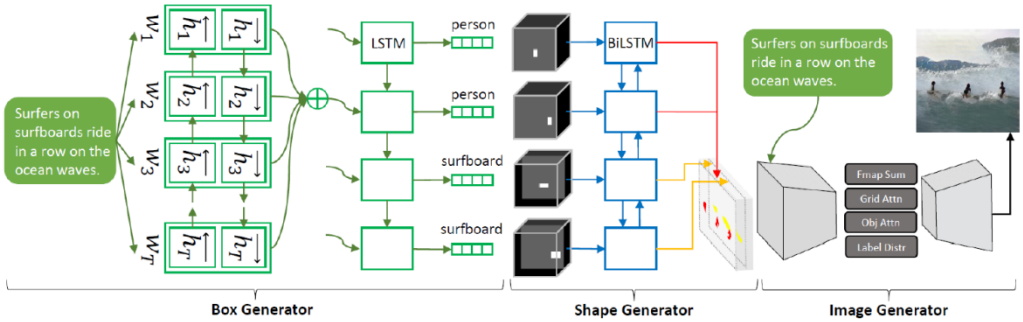


Σχήμα 3.2:Μακροσκοπική αναπαράσταση αρχιτεκτονικής AttnGAN (Πηγή: [TX2017])

3.1.5 Obj-GAN [WL2019]

Το μοντέλο αυτό επιτρέπει την σύνθεση εικόνας που αφορά περίπλοκες σκηνές. Η λειτουργία του επικεντρώνεται κυρίως στο εκάστοτε αντικείμενο που περιγράφεται (object-centered text-to-image synthesis) ακολουθώντας μια εξελικτική πορεία

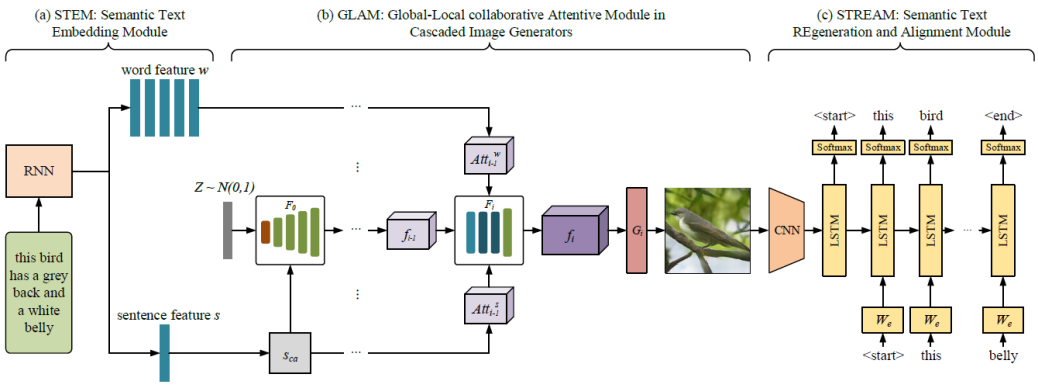
σύνθεσης εικόνας. Στην αρχή, γίνεται σχεδιασμός κάποιον κιβωτίων (bounding boxes) που θα φιλοξενήσουν στη συνέχεια τα αντικείμενα που πρόκειται να σχεδιαστούν. Πριν το σχεδιασμό των αντικειμένων, δημιουργούνται οι φιγούρες αυτών χωρίς λεπτομέρειες. Τελος, αφού ολοκληρωθεί και ο σχεδιασμός των αντικειμένων, το μοντέλο προχωράει σε προσθήκη κάποιων στοιχείων του φόντου και ορισμένων επιπρόσθετων αισθητικών λεπτομερειών.



Σχήμα 3.3:Μακροσκοπική αναπαράσταση αρχιτεκτονικής Obj-GAN (Πηγή: [WL2019])

3.1.6 MirrorGAN [TQ2019]

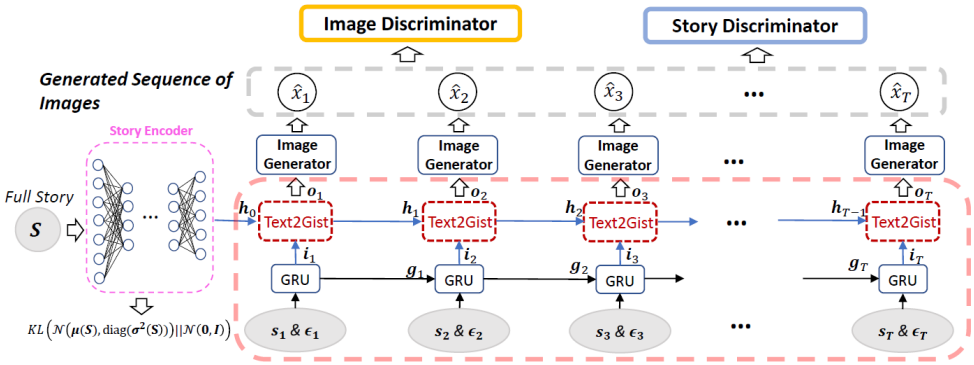
Το συγκεκριμένο μοντέλο αποτελείται απο τρία τμήματα. Το πρώτο τμήμα, λαμβάνει τις λεκτικές περιγραφές και παράγει διανυσματικές αναπαραστάσεις αυτών. Το δεύτερο τμήμα, που πραγματοποεί ουσιαστικά τη σύνθεση της εκάστοτε εικόνας, διαθέτει ένα σύνολο απο στοιβαγμένα Γεννητικά Ανταγωνιστικά Δικτυα και ένα μοντέλο εστίασης. Τέλος, το τρίτο τμήμα παίρνει τις παραγόμενες εικόνες και δημιουργεί μία νέα λεκτική περιγραφή(image captioning.). Ετσι, το μοντέλο αυτό συνδυάζει το πρόβλημα σύνθεσης εικόνας απο κείμενο με το πρόβλημα της απόδοσης λεζάντας σε εικόνα. Με αυτό το τρόπο το μοντέλο αυτό προσθέτει έναν επιπλέον παράγοντα στη συνάρτηση σφάλματος του, που αφορά την ομοιότητα της πραγματικής λεκτικής περιγραφής με την συνθετική.



Σχήμα 3.4:Μακροσκοπική αναπαράσταση αρχιτεκτονικής MirrorGAN (Πηγή: [TQ2019])

3.1.7 StoryGAN [YL2018]

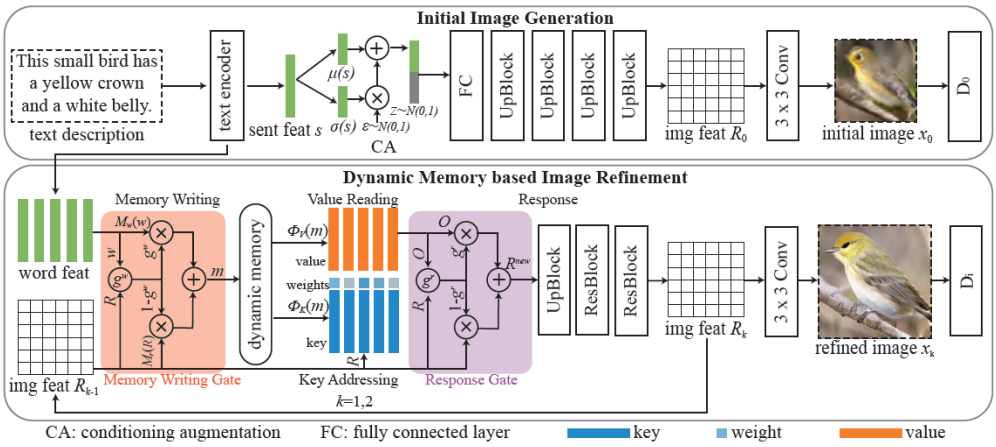
Το StoryGAN όπως δηλώνει και το όνομα του, αφορά την οπτικοποίηση μίας ιστορίας. Είναι δηλαδή ένα μοντέλο το οποίο συνθέτει εικόνες απο μία σύντομη ιστορία. Κάθε πρόταση της εκάστοτε ιστορίας οπτικοποιείται σε μία εικόνα. Στο τέλος οι εικόνες δεν φτάνει να είναι απλά υψηλής ποιότητας και σχετικές με τις προτάσεις που τις περιγράφουν, πρέπει επιπρόσθετα οι εικόνες μεταξύ τους να έχουν μία συλλογική συνοχή αναφορικά με τις σκηνές αλλά και με τους απεικονιζόμενους χαρακτήρες προκειμένου να γίνεται κατανοητή η ιστορία.



Σχήμα 3.5:Μακροσκοπική αναπαράσταση αρχιτεκτονικής StoryGAN (Πηγή: [YL2018])

3.1.8 DM-GAN [MZ2019]

Συχνά τα μοντέλα σύνθεσης εικόνας απο κείμενο, βασίζουν την ποιότητα των τελικών παραγόμενων εικόνων τους στις εικόνες που παράγονται στα πρώτα στάδια. Αυτό σημαίνει φυσικά, πως αν οι εικόνες των πρώτων σταδίων δεν είναι καλής ποιότητας, τότε και η τελική εικόνα θα παρουσιάζει προβλήματα. Το DM-GAN(Dynamic Memory) χρησιμοποιώντας ένα σύνολο *θυρών μνήμης*(memory gates), που λαμβάνουν ως είσοδο τη συνθετική εικόνα του κάθε σταδίου μαζί με τη λεκτική περιγραφή της, μπορεί να εντοπίσει και στη συνέχεια να διορθώσει όλα τα τμήματα των εικόνων που δεν έχουν αποτυπωθεί σωστά.



Σχήμα 3.6:Μακροσκοπική αναπαράσταση αρχιτεκτονικής CSL-GAN (Πηγή:[MZ2019])

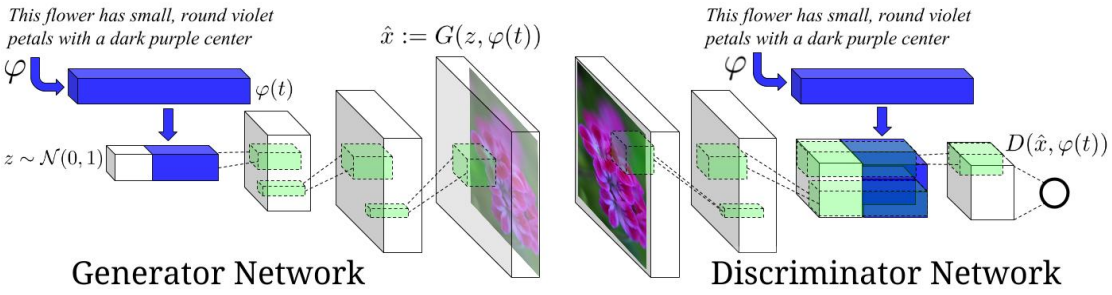
Κεφάλαιο 4

4. Συνδυασμός GAN-CLS & StackGAN

Στο κεφάλαιο αυτό θα γίνει μία πιο ενδελεχής παρουσίαση των μοντέλων GAN-CLS και StackGAN που παρουσιάστηκαν στο 3ο Κεφάλαιο. Ο συνδυασμός αυτών των δύο μοντέλων αποτελεί και το πειραματικό μέρος της παρούσας διπλωματικής εργασίας, καθώς έγινε χρήση έτοιμου κώδικα υλοποιημένου από τρίτους. Επιπρόσθετα έχουν προστεθεί κάποια python scripts για την επεξεργασία των συνθετικών εικόνων του μοντέλου καθώς και scripts για τον υπολογισμό μετρικών που θα παρουσιαστούν σε επόμενο κεφάλαιο. Ολοι οι κώδικες αναφορικά με τα προαναφερθέντα, μπορούν να βρεθούν στο παρακάτω σύνδεσμο: https://github.com/EliasDimitriou14/Text2Image_Thesis.

4.1 GAN-CLS(Conditional Latent Space)

Η πρώτη προσπάθεια επίλυσης του προβλήματος σύνθεσης εικόνας από κείμενο έγινε απο τον Scott Reed και του συνεργάτες του στο έργο [CLS2016]. Στο συγκεκριμένο έργο οι συγγραφείς χωρίζουν το πρόβλημα σε δύο υποπροβλήματα: την εύρεση μίας διευκρινιστικής οπτικής αναπαράστασης των λεκτικών περιγραφών και την χρήση αυτής της αναπαράστασης για την παραγωγή ρεαλιστικών εικόνων. Πρέπει να σημειωθεί, οτι τα αποτελέσματα αυτής της προσέγγισης ήταν αρκετά ενθαρρυντικά.



Σχήμα 4.1:Μακροσκοπική αναπαράσταση αρχιτεκτονικής CSL-GAN
(Πηγή:https://github.com/hellochick/text-to-image)

4.1.1 Αρχιτεκτονική GAN-CLS

Η αρχιτεκτονική του GAN- CLS και στη μεριά του *γεννήτορα* αλλά και στου *διευκρινιστή* είναι παρόμοια με αυτή των βαθιά συνελικτικών-Γεννητικά Ανταγωνιστικών Δικτύων(DC-GAN) όπως αυτά παρουσιάζονται στο έργο [AR2015]. Μια μακροσκοπική οπτική της αρχιτεκτονικής φαίνεται στο Σχήμα 4.1.

Στον *γεννήτορα*, αρχικά γίνεται δειγματοληψία ενός διανύσματος θορύβου $z \in \mathbb{R} \sim \mathcal{N}(0,1)$ 128 διαστάσεων. Η λεκτική περιγραφή t περνάει μέσα απο τον κωδικοποιητή φ και η έξοδος $\varphi(t)$ συμπιέζεται σε 128 διαστάσεις με την χρήση ενός πλήρως συνδεδεμένου στρώματος, με συνάρτηση ενεργοποιήσης μια *διαρρέουσα διορθωμένη γραμμική μονάδα* (leaky ReLU). Στη συνέχεια, γίνεται συνένωση της διανυσματικής αναπαράστασης με το διάνυσμα του θορύβου z. Το νέο αυτό διάνυσμα με τη σειρά του μετασχηματίζεται γραμμικά και περνάει απο μία σειρά αποσυνελίξεων με ενεργοποιήσεις της leaky ReLU μέχρι να ληφθεί ένας τανυστής διαστάσεων 64 *64* 3. Τέλος, προκειμένου τα εικονοστοιχεία να παίρνουν τιμές στο εύρος [-1, 1], οι τιμές του τανυστή περνούν απο μία συνάρτηση ενεργοποίησης υπερβολικής εφαπτομένης($tanh$).

Στον *διευκρινιστή*, η εικόνα εισόδου περνάει απο ένα σύνολο συνελικτικών στρωμάτων με στόχο η χωρική ανάλυση να φτάσει σε μέγεθος 4*4. Τότε, οι διανυσματικές αναπαραστάσεις κειμένου συμπιέζονται σε ένα διάνυσμα μεγέθους 128 διαστάσεων με τη χρήση ενός πλήρως συνδεδεμένου στρώματος ενεργοποιήσεων της leaky ReLU, όπως ακριβώς και στη περίπτωση του *γεννήτορα*. Εν συνεχεία, οι διανυσματικές αναπαραστάσεις κειμένου επαναλαμβάνονται και συνενώνονται στα συνελικτικά χαρακτηριστικά του δικτύου. Ο συνενωμένος τανυστής στη συνέχεια, περνάει απο κάποια επιπρόσθετα στάδια συνελίξεων μεχρι να ληφθεί ένα μονόμετρο μέγεθος. Τέλος, σε αυτό το μονόμετρο μέγεθος γίνεται εφαρμογή μιάς σιγμοειδούς συνάρτησης ενεργοποίησης, ώστε το μέγεθος αυτό να λάβει τιμές στο εύρος [0,1] ώστε να αντιπρωσωπεύει μία έγκυρη πιθανότητα.

Σε αυτό το σημείο σημειώνεται οτι οι διανυσματικές αναπαραστάσεις κειμένου που αναφέρονται παραπάνω ακολουθούν την λογική που αναλύθηκε στην ενότητα 2.3.

4.1.2 Συνάρτηση Σφάλματος GAN-CLS

Ο *διευκρινιστής* του GAN-CLS παρουσιάζει μια διαφοροποίηση στη συνάρτηση σφάλματος του σε σχέση με αυτή που παρουσιάστηκε στην σχέση (2.2). Η διαφοροποίηση αυτή έχει ως στόχο την ενίσχυση της δυνατότητας του *διευκρινιστή* να μπορεί να ταιριάζει εικόνα με αντίστοιχο κείμενο. Η τροποποιημένη σχέση (4.1) σύμφωνα με το [CLS2016] φαίνεται παρακάτω:

$$L_D = \log(\mathrm{D}(\mathrm{x}, \mathrm{h})) + \frac{1}{2}\log(1 - \mathrm{D}(\mathrm{x}, \hat{\mathrm{h}})) + \frac{1}{2}\log(1 - \mathrm{D}(\mathrm{G}(\mathrm{z}, \mathrm{h}), \mathrm{h})) \quad (4.1)$$

Στη παραπάνω σχέση το $x \sim p_{data}$ προέρχεται απο την κατανομή των πραγματικών δεδομένων, το $h \sim p_{match\_embedding}$ από την κατανομή των κωδικοποιημένων λεκτικών περιγραφών υπό μορφή διανύσματων που ταιριάζουν με την εικόνα, το $\hat{\mathrm{h}} \sim p_{mis-match\_embedding}$ από την κατανομή των κωδικοποιημένων λεκτικών περιγραφών υπό μορφή διανύσματων που δεν ταιριάζουν με την εικόνα και το $z \sim p_z$ προέρχεται από την κατανομή του θορύβου.

Με αυτό το τρόπο ο *διευκρινιστής* πλέον δύναται να αναγνωρίσει δύο ειδών λάθη. Το πρώτο αφορά ρεαλιστικές εικόνες οι οποίες δεν ταιριάζουν με τις λεκτικές περιγραφές και το δεύτερο αφορά μή ρεαλιστικές εικόνες.

4.1.3 GAN με πολλαπλή παρεμβολή(GAN-INT)

Τα βαθιά δίκτυα, όπως έχει αποδειχθεί, έχουν παρουσιάσει την τάση να μαθαίνουν αναπαραστάσεις για ζευγάρια εισόδου των οποίων οι διανυσματικές αναπαραστάσεις βρίσκονται κοντά σχετικά με την πολλαπλότητα των δεδομένων. Αυτό το γεγονός, οδήγησε τους συγγραφείς του [CLS2016] στη δημιουργία ενός συνόλου επιπρόσθετων διανυσματικών αναπαραστάσεων κειμένου από την παρεμβολή μεταξύ αναπαραστάσεων του συνόλου εκπαίδευσης. Αυτές οι παρεμβαλόμενες αναπαραστάσεις κειμένου, δεν είναι αναγκαίο να αναταποκρίνονται σε κάποιο πραγματικό κείμενο γραμμένο απο άνθρωπο και επομένως δεν εμφανίζεται κάποιο επιπλέον υπολογιστικό κόστος για την διαδικασία απόδοσης ετικετών. Αυτό μπορεί να επιτευχθεί με την προσθήκη ενός παραπάνω όρου προς ελαχιστοποιήση στον *γεννήτορα*:

$$E_{t1,t2\sim pdata}\left[\log\left(1 - D(G(z, \beta t_1 + (1-\beta)\, t_2))\right)\right] \qquad (4.2)$$

όπου $\boldsymbol{z}$ προέρχεται απο την κατανομή του θορύβου και το $\boldsymbol{\beta}$ παρεμβάλεται ανάμεσα στις ενσωματώσεις $\boldsymbol{t_1}$ και $\boldsymbol{t_2}$. Επίσης , κατόπιν δοκιμών, οι συγγραφείς παρατήρησαν οτι το $\boldsymbol{\beta=0.5}$ λόγω καλών αποτελεσμάτων. Επιπρόσθετα, είναι σημαντικό να αναφερθεί οτι τα στοιχεία $\boldsymbol{t_1}$ και $\boldsymbol{t_2}$ μπορεί να προέρχονται απο διαφορετικές εικόνες και κατηγορίες.

Εξαιτίας του γεγονότος οτι οι παρεμβαλόμενες αναπαραστάσεις είναι συνθετικές, ο *διευκρινιστής* δεν διαθέτει πραγματικά ζεύγη εικόνας-κειμένου για να εκπαιδευτεί. Παρόλλα αυτά όμως, ο *διευκρινιστής* μαθαίνει να κάνει προβλέψεις αναφορικά με την συνάφεια εικόνας-κειμένου. Αν λοιπόν ο *διευκρινιστής* επιτύχει σε αυτή του την ενέργεια, τότε ικανοποιώντας τον *διευκρινιστή* στις παρεμβαλόμενες αναπαραστάσεις κειμένου ο *γεννήτορας* μπορεί να μάθει να συμπληρώνει τα κενά που υπάρχουν στη πολλαπλότητα των δεδομένων στα διάφορα σημεία εκπαίδευσης.

4.2 Στοιβαγμένα GAN (Stacked GANs)

Παρά το γεγονός ότι το GAN-CLS κατάφερε να δώσει πολύ καλά πρώτα αποτελέσματα στην επίλυση του προβήματος της σύνθεσης εικόνας από κείμενο, η προσέγγιση που ακολουθεί με την παραγωγή μιάς πλήρως λεπτομερούς εικόνας απευθείας δεν είναι και η καλύτερη. Αυτό ακριβώς έδειξαν ο Han Zhang και οι συνεργάτες του στο έργο [HZ2016], με την πρόταση των στοιβαγμένων Γεννητικών Ανταγωνιστικών Δικτύων.

Τα StackGAN, όπως γίνεται εύκολα κατανοητό και από το όνομα τους, διαθέτουν μία αρχιτεκτονική που στηρίζεται σε χρήση περισσότερων του ενός GANs. Το πρώτο GAN, συχνά αναφερόμενο και ως Stage I, δημιουργεί εικόνες χαμηλής ανάλυσης, 64 *64 εικονοστοιχείων, από λεκτικές περιγραφές με παρόμοιο τρόπο με το GAN-CLS. Το δεύτερο GAN, συχνά αναφερόμενο και ως Stage II, διαθέτει ένα *γεννήτορα* που στην είσοδο του λαμβάνει την παραγόμενη εικόνα του Stage I *γεννήτορα* και συνθέτει μία εικόνα μεγάλυτερης ανάλυσης, 256*256 εικονοστοιχείων, με περισσότερες λεπτομέρειες και μεγαλύτερη συνάφεια με την εκάστοτε λεκτική περιγραφή.

4.2.1 Επαύξηση Ενσωματώσεων Κειμένου

Στο [HZ2016] οι συγγραφείς προτείνουν επιπρόσθετα μία τεχνική *επάυξησης συνθήκης* (conditioning augmentation-CA), με στόχο την αύξηση του διανυσματικού χώρου αναπαράστασης του κειμένου.

Ουσιαστικά, με την χρήση της προαναφερθείσας τεχνικής, επιδιώκεται η παραγωγή περισσότερων δειγμάτων εκπαίδευσης από ένα μικρό σύνολο ζευγών εικόνας-λεκτικής περιγραφής, προκειμένου να μπορεί το μοντέλο να παρουσιάζει μεγαλύτερη ανεκτικότητα στις τυχόν διαταραχές που προκύπτουν απο τον διανυσματικό χώρο των συνθηκών. Επιπλέον, για να ενισχυθεί η ανεκτικότητα αυτή ακόμα περισσότερο και να αποφευχθεί η υπερπροσαρμογή του μοντέλου, οι συγγραφείς προσέθεσαν την απόκλιση Kullback-Leibler, ανάμεσα στην κανονική Γκαουσιανή κατανομή και την Γκαουσιανή κατανομή των διανυσματικών αναπαραστάσεων, ως όρο κανονικοποίησης στον *γεννήτορα,* κατά την διαδικασία της εκπαίδευσης. Στην σχέση (4.3) φαίνεται πώς υπολογίζεται η απόκλιση Kullback-Leibler.

$$\mathbf{D_{KL}}\left(\mathrm{N}(\mu(\boldsymbol{\phi_t}), \Sigma(\boldsymbol{\phi_t}))\,||\mathrm{N}(0,1)\right) \qquad (4.3)$$

Σύμφωνα με τη παραπάνω σχέση λοιπόν, για κάθε μία διανυσματική αναπαράσταση κειμένου $\boldsymbol{\phi_t}$, δειγματοληπτώντας τυχαία την Γκαουσιανή κατανομή $\boldsymbol{N(\mu(\phi_t), \Sigma(\phi_t))}$, όπου $\boldsymbol{\mu(\phi_t)}$ είναι η μέση τιμή και $\boldsymbol{\Sigma(\phi_t)}$ ο διαγώνιος πίνακας συνδιακύμανσης, δύναται να παραχθούν επαυξημένες διανυσματικές αναπαραστάσεις κειμένου.

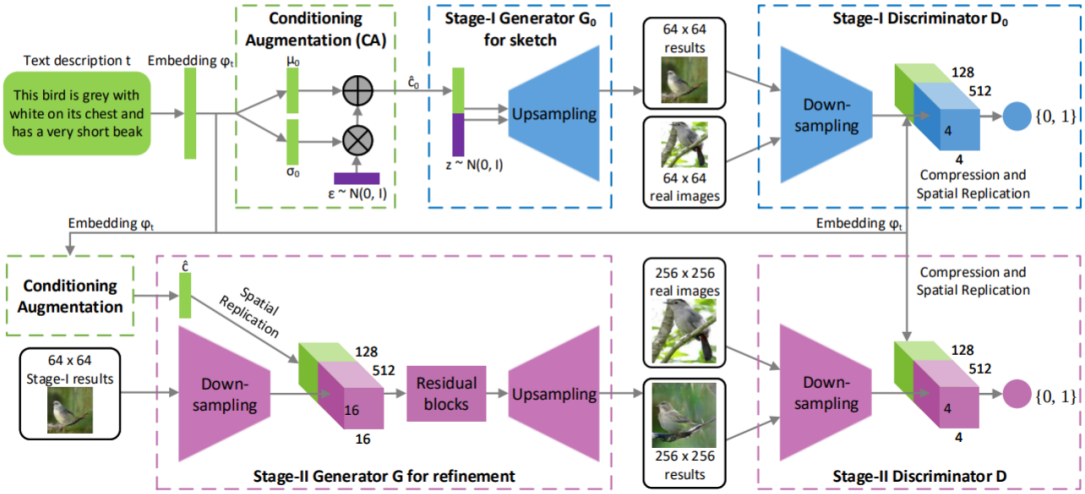
4.2.2 Αρχιτεκτονική StackGAN

Στο Σχήμα 4.2 παρουσιάζεται μακροσκοπικά η αρχιτεκτονική του StackGAN. Πέρα απο την προσθήκη της τεχνικής της επαύξησης συνθήκης, η αρχιτεκτονική του Stage I είναι πανομοιότυπη με αυτή του GAN-CLS όπως αυτή παρουσιάστηκε στην ενότητα (4.1.1).

Στο Stage II του μοντέλου,ο *γεννήτορας* αρχικά λαμβάνει εικόνες μεγέθους 64*64 και τις υποδειγματοληπτεί εώς ότου να φτάσουν σε μια χωρική ανάλυση 4*4 εικονοστοιχείων. Την συνέχεια, γίνεται συννένωση της 4*4 εικόνας με την αντίστοιχη επαυξημένη διανυσματική αναπαράσταση κειμένου, προκειμένου να βελτιωθεί ο συσχετισμός της εικόνας με το κείμενο του Stage I . Έπειτα, η συννενωμένη εικόνα με την αντίστοιχη επαυξημένη διανυσματική αναπαράσταση κειμένου περνάει απο 3 επιπρόσθετα στρώματα και υπερδειγματοληπτείται μέχρι να ληφθεί ένας τανυστής 256*256*3. Τέλος, με την εφαρμογή της συνάρτησης ενεργοποίησης υπερβολικής εφαπτομένης(*tanh*) η έξοδος λαμβάνει τιμές στο εύρος [-1, 1].

Ο *διευκρινιστής* του Stage I είναι ίδιος με αυτόν που περιγράφηκε στο GAN-CLS. Ο *διευκρινιστής* του Stage II είναι και αυτός όμοιος, με την μόνη διαφορά οτι πλέον υπάρχουν περισσότερα συνελικτικά στρώματα υποδειγματοληψίας για να μπορέσει να γίνει η προσαρμογή των εικόνων μεγαλύτερης ανάλυσης της εισόδου.

Τέλος και στον *γεννήτορα* και στον *διευκρινιστή* γίνεται κανονικοποίηση ομάδας(batch normalization), ενώ αναφορικά με την χρήση συναρτήσεων ενεργοποιήσης στον *γεννήτορα* γίνεται χρήση ReLU και στον *διευκρινιστή* leaky ReLU αντίστοιχα.



Σχήμα 4.2:Μακροσκοπική αναπαράσταση αρχιτεκτονικής StackGAN (Πηγή: [HZ2016])

Κεφάλαιο 5

5. Μετρικές αξιολόγησης & σύνολο δεδομένων

Σε αυτό το κεφάλαιο θα γίνει παρουσίαση των μετρικών αξιολόγησης που χρησιμοποιήθηκαν στη παρούσα διπλωματική, καθώς και του συνόλου δεδομένων που αξιοποιήθηκε για την εκπαίδευση του μοντέλου.

5.1 Σύνολο δεδομένων

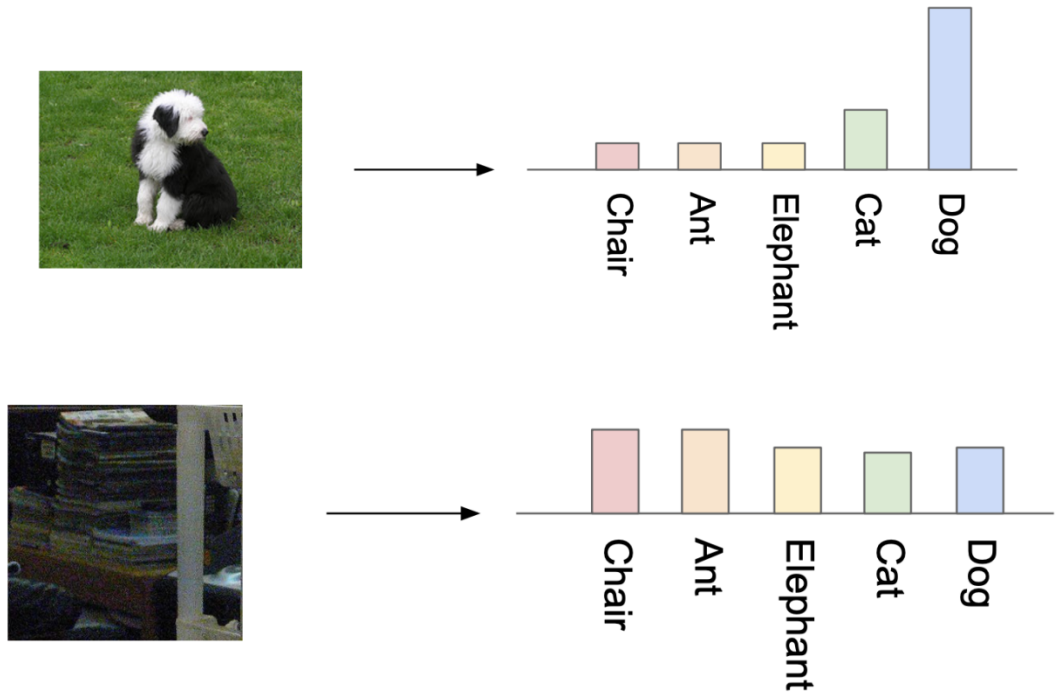
Το σύνολο δεδομένων που χρησιμοποιήθηκε στη παρούσα διπλωματική εργασία είναι το Oxford-102[MEN2008], το οποίο περιλαμβάνει εικόνες διαφορετικής κλίμακας, πόζας και φωτισμού, από 102 διαφορετικές κατηγορίες λουλουδιών που εμφανίζονται στο Ηνωμένο Βασίλειο. Κάθε μία από τις κατηγορίες λουλουδιών αποτελείται από 40 έως 258 εικόνες. Συνολικά το σύνολο δεομένων αποτελείται από 8189 εικόνες.

5.2 Inception Score(IS)

Το Inception Score [TS2016] αποτελεί μιά πολύ δημοφιλή μετρική αξιολόγησης των Γεννητικών Ανταγωνιστικών Δικτύων. Ο υπολογισμός της μετρικής ξεκινάει ουσιαστικά από το μοντέλο ταξινόμισης Inception V3[CS2015], καθώς η εκτίμηση της ποιότητας των συνθετικών εικόνων βασίζεται στο πόσο καλά μπορεί ο ταξινομητής να τις τοποθετήσει σε προκαθορισμένες κλάσεις. Με την χρήση της συγκεκριμένης μετρικής, εξετάζεται το εκάστοτε σύνολο συνθετικών εικόνων ως προς την ποιότητα τους αλλά και ως προς την ποικιλία τους. Αρχικά, γίνεται η εκτίμιση της ποιότητας των εικόνων με τον προσδιορισμό μίας *κατανομής πιθανότητας*(probability distribution), όπως φαίνεται στο (Σχήμα 5.1), του εκάστοτε δείγματος για κάθε κλάση από το Inception V3. Αν ο ταξινομητής δύναται να καθορίσει εύκολα τη κλάση που ανήκει το κάθε δείγμα τότε ιδανικά δημιουργείται μια κατανομή που ονομάζεται *στενή*(narrow distribution). Στη συνέχεια, γίνεται άθροιση των *κατανομών των ετικετών* που προέκυψαν από την εκτίμιση ποιότητας των εικόνων υπολογίζοντας έτσι την *οριακή κατανομή*. Αν τώρα η οριακή αυτή κατανομή, έχει μια *ομοιόμορφη μορφή* τότε αυτό σημαίνει πως το εκάστοτε σύστημα μπορεί να παράξει ένα ευρύ φάσμα εικόνων. Έχοντας πλέον την κατανομή ετικετών της κάθε εικόνας αλλά και την οριακή κατανομή, είναι εφικτός ο υπολογισμός του Inception Score του συνόλου δεδομένων, συγκρίνοντας απλώς τις δύο αυτές κατανομές. Η σύγκριση αυτή γίνεται με την βοήθεια της απόκλισης Kullback-Leibler που παρουσιάστηκε στην ενότητα 4.2.1. Αν η τιμή της απόκλισης είναι μεγάλη, τότε θα είναι υψηλό και το Inception Score. Η σχέση 5.1 είναι αυτή που υπολογίζει την μετρική.

$$IS = \exp\left(\mathbf{E_x}\, \mathbf{D_{KL}}(p(y|x) \,||\, p(y))\right) \qquad (5.1)$$

Στη παραπάνω σχέση το $\boldsymbol{x}$ εκφράζει τη συνθετική εικόνα, το $\boldsymbol{y}$ είναι η ετικέτα που αποδόθηκε από τον ταξινομητή, $\boldsymbol{p(y)}$ αποτελεί την οριακή κατανομή και $\boldsymbol{p(y|x)}$ είναι η κατανομή ετικέτας του δείγματος.



Σχήμα 5.1: Υπολογισμός κατανομής πιθανότητας δείγματος για κάθε κλάση. Η πρώτη εικόνα αποτελεί παράδειγμα στενής κατανομής. (Πηγή: https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score372dff6a8c7a)

Τέλος είναι σημαντικό να αναφερθεί ότι όσο υψηλότερο είναι το IS τόσο το καλύτερο, με υψηλότερη θεωρητική τιμή το άπειρο και χαμηλότερη το 0.

5.3 Fréchet Inception Distance(FID)

Η Fréchet Inception Distance, FID για συντομία, αποτελεί μία μετρική κατάλληλη για την αξιολόγηση της ποιότητας συνθετικών εικόνων καθώς δημιουργήθηκε με σκοπό την αξιολόγηση της απόδοσης των Γεννητικών Ανταγωνιστικών Δικτίων. Η πρώτη παρουσίαση της μετρικής, έγινε στο έργο [MH2018] του Martin Heusel. Η μετρική αποσκοπεί στην αξιολόγηση συνθετικών εικόνων πραγματοποιώντας συγκρίσεις στατιστικών στοιχείων που προέρχονται απο συλλογές συνθετικών και πραγματικών εικόνων. Η FID, ουσιαστικά υπολογίζει την απόσταση μεταξύ μιας κατανομής χαρακτηριστικών $\boldsymbol{p(.)}$ που προέρχεται απο τις συνθετικές εικόνες και μίας

κατανομής χαρακτηριστικών $\boldsymbol{p_r(.)}$ που προέρχεται από πραγματικές εικόνες. Αρχικά, γίνεται λήψη γκαουσιανής κατανομής από τις δύο κατανομές αυτές με μέση τιμή και συνδιακύμανση $\boldsymbol{(m, C)}$ και $\boldsymbol{(m_r, C_r)}$ αντίστοιχα. Στη συνέχεια με χρήση της απόστασης Fréchet, ή αλλίως Wasserstein-2, δύναται να γίνει υπολογισμός της απόστασης των κατανομών αυτών. Η σχέση 5.2 είναι αυτή που υπολογίζει την μετρική FID.

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

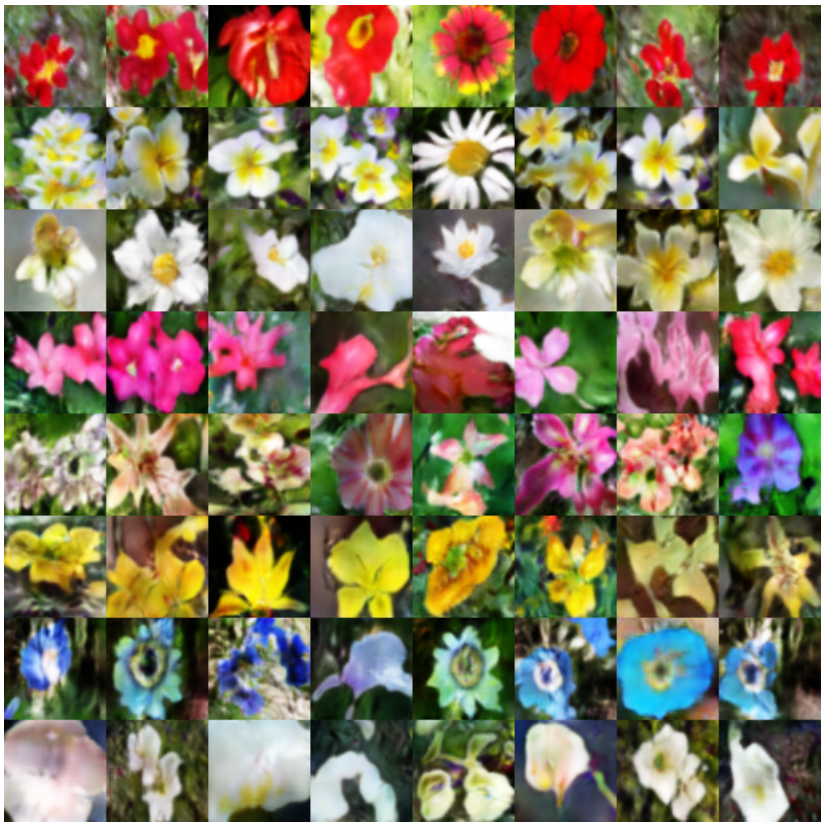
Στη παραπάνω σχέση το $\boldsymbol{T_r}$ εκφράζει το άθροισμα των στοιχείων της διαγωνίου του πίνακα.

Κεφάλαιο 6

6. Πειραματική διαδικασία & αποτελέσματα

Το παρόν κεφάλαιο αποτελεί την παρουσίαση ορισμένων στοιχείων που αφορούν το πειραματικό μέρος της εργασίας, με έμφαση στα αριθμητικά αποτελέσματα των μετρικών που παρουσιάστηκαν στο 5ο Κεφάλαιο και στα προβλήματα που αντιμετωπίστηκαν.

Το μοντέλο που χρησιμοποιήθηκε στα πλαίσια της εργασίας, κατά την διαδικασία της εκπαίδευσης του, ακολουθεί την λογική των εποχών δηλαδή μιας σειράς επαναλήψεων κατά την οποία το δίκτυο τροφοδοτείται με τις πραγματικές εικόνες που συνοδεύονται απο λεκτικές περιγραφές. Αυτό γίνεται προκειμένου να μπορέσει να "μάθει" τι πρέπει στη συνέχεια να συνθέσει, εξάγοντας ορισμένα χαρακτηριστικά από τις πραγματικές εικόνες. Οι εποχές εκπαίδευσης συνήθως πρέπει να είναι πολλές π.χ 600, καθώς με το πέρας της κάθε εποχής βελτιώνονται οι παράμετροι του μοντέλου και σταδιακά δημιουργεί όλο και καλύτερες εικόνες. Στο τέλος κάθε εποχής, το μοντέλο παρέχει και μία *εικόνα δείγματος* σαν αυτή του Σχήματος 6.1.



Σχήμα 6.1: Συνθετική εικόνα που παράγει το μοντέλο στο τέλος κάθε εποχής. Η εικόνα έχει τη μορφή πλέγματος 8*8 και σε κάθε σειρά απεικονίζονται εκδόσεις συνθετικών εικόνων του ίδιου λουλουδιού.(Πηγή: https://github.com/hellochick/text-to-image)

Η εικόνα του Σχήματος 6.1 αποτελεί εικόνα δείγματος μετά από εκτέλεση του μοντέλου για 800 εποχές. Στο τέλος της κάθε εποχής, γίνεται δημιουργία μιας αντίστοιχης εικόνας, και όπως είναι προφανές οι μεταγενέστερες εποχές εκπαίδευσης παρέχουν σαφέστατα καλύτερες εικόνες λόγω της βελτίωσης των παραμέτρων εκπαίδευσης που προσαρμόζονται στο τέλος κάθε εποχής. Οι εικόνες αυτές,

δημιουργούνται με βάση 8 τυχαίες λεκτικές περιγραφές, αναφερόμενες απο τον δημιουργό και ως *πρόταση δείγματος*. Επομένως, αν το μοντέλο τρέξει για 500 εποχές τότε θα δώσει 500 εικόνες δείγματος, με σταδιακή βελτίωση του απεικονιζόμενου αντικειμένου, για τις ίδιες 8 λεκτικές περιγραφές που έχουν τεθεί πριν από την εκτέλεση. Προφανώς, αν γίνει αλλαγή των λεκτικών περιγραφών και επανεκτέλεση του μοντέλου, στην έξοδο θα ληφθούν εικόνες για τα καινούρια λουλούδια που περιγράφονται. Στη παρούσα διπλωματική εργασία έγινε χρήση 64 διαφορετικών και τυχαία επιλεγμένων λεκτικών περιγραφών λουλουδιών για να επιτευχθεί κατά κάποιο τρόπο ποικιλία στις συνθετικές εικόνες.

6.1 Προβλήματα εκτέλεσης

Ένα βασικό πρόβλημα που παρουσιάστηκε κατά την εκτέλεση του μοντέλου αφορούσε τον πραγματικό χρόνο εκτέλεσης. Η ολοκλήρωση μιας και μόνο εποχής εκτέλεσης απαιτούσε 45 λεπτά περίπου. Εξαιτίας αυτού, ήταν πρακτικά αδύνατη η εκπαίδευση του μοντέλου σε μεγάλο πλήθος εποχών, γεγονός που περιόρισε σε τεράστιο βαθμό το σύνολο των εικόνων, ποσοτικά και ποιοτικά, που δημιουργήθηκαν τελικώς προκειμένου να γίνει η αξιολόγηση. Συνολικά το μοντέλο εκπαιδεύτηκε σε 126 εποχές, οι οποίες ήταν σπασπένες σε 6 διαδοχικές εκτελέσεις των 21 εποχών. Είναι σημαντικό να αναφερθεί, πως ενδέχεται σε έναν πιο σύγχρονο υπολογιστή το μοντέλο να τρέχει πολύ πιο γρήγορα. Επιπρόσθετα, ο αναγνώστης ενθαρρύνεται να δοκιμάσει να εκτελέσει το μοντέλο με παράλληλο τρόπο, ώστε ενδεχομένως να επιταχύνει την διαδικασία, έαν φυσικά διαθέτει Nvidia GPU και CUDA που είναι απαραίτητα εργαλεία σε μία τέτοια περίπτωση.

Επειδή λοιπόν το πλήθος των εικόνων μετά από ένα τόσο μικρό αριθμό εποχών είναι μικρό, αναπτύχθηκε κώδικας ο οποίος "κόβει" την κάθε εικόνα δείγματος σε 64 κομμάτια, δηλαδή σε 64 εικόνες των λουλουδιών που απεικονίζονται στην εκάστοτε εικόνα. Με αυτό τον τρόπο στο τέλος της εκπαίδευσης συγκεντρώθηκε ένα πλήθος 8000+ εικόνων εκ των οποίων χρησιμοποιήθηκαν για την αξιολόγηση 7370 για λόγο που θα εξηγηθεί στην ενότητα 6.2 που ακολουθεί.

6.2 Αποτελέσματα μετρικών & συγκρίσεις μοντέλων

Σε αυτή την ενότητα γίνεται παρουσίαση των αριθμητικών αποτελεσμάτων των μετρικών IS και FID για το σύνολο των συνθετικών εικόνων που δημιούργησε το μοντέλο. Στη περίπτωση του IS, αξιολογήθηκαν 7370 συνθετικές εικόνες μεγέθους 64*64. Το ίδιο ακριβώς σύνολο συνθετικών εικόνων χρησιμοποιήθηκε και στην περίπτωση της FID, σε συνδυασμό με το σύνολο εικόνων εκπαίδευσης καθώς η μετρική αυτή προβλέπει την σύγκριση δύο διαφορετικών κατανομών, μία που προέρχεται από τις συνθετικές εικόνες και μία που προέρχεται από τις πραγματικές. Αναφορικά με το πλήθος των συνθετικών εικόνων, χρησιμοποιήθηκαν ακριβώς 7370 μόνο και μόνο διότι ο δημιουργός του μοντέλου παρείχε σαν εικόνες εκπαίδευσης 7370 εικόνες σε μορφή numpy arrays όπως τις αναγνωρίζει η python. Επομένως για να μπορέσει να γίνει χρήση της FID, που θέλει τα συγκρινόμενα σύνολα εικόνων να

πληρούν τις ίδιες προδιαγραφές, αξιοποίθηκαν 7370 συνθετικές εικόνες. Τα αριθμήτικά αποτελέσματα φαίνονται στον Πίνακα 6.1.

Model	Oxford-102	
	IS ↑	FID ↓
CLS-GAN - StackGAN	3.81 ± 0.17	87.48

Πίνακας 6.1: Inception Score (IS) και Fréchet Inception Distance (FID) για το δίκτυο CLS-GAN - StackGAN της εργασίας

Ορισμένα από τα μοντέλα που αναφέρθηκαν στην ενότητα 3.1, έχουν χρησιμοποιηθεί για σύνθεση εικόνας από κείμενο για το σύνολο δεδομένων Oxford-102, και έχουν επίσης αξιολογηθεί από τις ίδιες μετρικές αξιολόγησεις που χρησιμοπουούνται στη παρούσα εργασία. Ακολουθεί ο σχετικός Πίνακας 6.2 αριθμητικών αποτελεσμάτων για κάποια από αυτά τα μοντέλα.

Model	Oxford-102	
	IS ↑	FID ↓
GAN-INT-CLS	2.66 ± 0.03	79.55
StackGAN	3.20 ± 0.01	55.28
StackGAN++	3.26 ± 0.01	48.68

Πίνακας 6.2: Inception Score (IS) και Fréchet Inception Distance (FID) για το ορισμένα state-of-the-art μοντέλα σύνθεσης εικόνας από κείμενο

Όπως αναφέρεται και στο έργο [TS2016], μία καλή τεχνική αξιολόγησης προβλεπει ένα μεγάλο σύνολο εικόνων για αξιολόγηση. Έχοντας αυτό κατά νού, οι συγγραφείς των αντίστοιχων έργων που κάνουν χρήση των μοντέλων του Πίνακα 6.2 , έβγαλαν αποτελέσματα για τις μετρικές IS και FID για σύνολα εικόνων που μπορεί να ξεπερνούσαν ακόμα και τις 30.000 εικόνες. Εξαιτίας αυτού λοιπόν, είναι πολύ λογικό να αναλογιστεί κανείς γιατί τα αποτελέσματα του Πίνακα 6.1 δεν είναι και πολύ αντιπροσωπευτικά των πραγματικών δυνατοτήτων του μοντέλου CLS-GAN - StackGAN καθώς είναι αισθητά δαφορετικά με αυτά του Πίνακα 6.2. Επιπρόσθετα του διαφορετικού πλήθους συνθετικών εικόνων προς αξιολόγηση, είναι σημαντικό να αναφερθεί ότι οι εικόνες του μοντέλου της εργασίας καθώς και του GAN-INT-CLS είνα διαστάσεων 64*64, ενώ οι εικόνες των μοντέλων StackGAN και StackGAN++ είναι διατάσεων 256*256.

6.3 Αποτελέσματα συνθετικών εικόνων

Παρακάτω ακολουθούν μερικές ενδεικτικές συνθετικές εικόνες που δημιούργησε το εξεταζόμενο μοντέλο της εργασίας για διάφορα σύνολα λεκτικών περιγραφών σε διαφορετικά στάδια εκπαίδευσης του.

```
["the flower shown has yellow anther red pistil and bright red petals."] * int(sample_size/ni) + \
["this flower has petals that are yellow, white and purple and has dark lines"] * int(sample_size/ni)
["the petals on this flower are white with a yellow center"] * int(sample_size/ni) + \
["this flower has a lot of small round pink petals."] * int(sample_size/ni) + \
["this flower is orange in color, and has petals that are ruffled and rounded."] * int(sample_size/ni)
["the flower has yellow petals and the center of it is brown."] * int(sample_size/ni) + \
["this flower has petals that are blue and white."] * int(sample_size/ni) +\
["these white flowers have petals that start off white in color and end in a white towards the tips."]
```

Σχήμα 6.2: Συνθετική εικόνα που παράγει το μοντέλο για τις αντίστοιχες λεκτικές περιγραφές, όπως δόθηκαν στην υλοποιήση του κώδικα. Η παραπάνω εικόνα προέρχεται από τις πρώτες εποχές εκπαίδευσης του μοντέλου.

```
["this flower has a vertical growth of pink petals in a waxy chevron pattern."] * int(sa
["this flower has petals that are white and has flowery stigma"] * int(sample_size/ni) +
["this flower has purple petals with a yellow stamen in the center on a green pedicel."]
["the flower has petals that are burgundy with yellow anther and burgundy filaments."] *
["this flower has bright pink leaves with pointed tips surrounding tiny white blooms."]
["this flower is white and yellow in color, and has petals that are curved upward."] * i
["this flower has spiraling layers of pale purple petals around a yellow stamen."] * int
["this flower has large pink petals that turn dark pink towards the center."] * int(samp
```

Σχήμα 6.3: Συνθετική εικόνα που παράγει το μοντέλο για τις αντίστοιχες λεκτικές περιγραφές, όπως δόθηκαν στην υλοποιήση του κώδικα. Η παραπάνω εικόνα προέρχεται από μεταγενέστερες εποχές εκπαίδευσης του μοντέλου.

```
["this flower has petals that are yellow and very stringy"] * int(sample_size/ni) + \
["this flower is white in color, and has petals that are that are ruffled."] * int(sample_s
["this flower has small white stamen, a large white petal, and a yellow ovary."] * int(samp
["this white flower has many petals with prominent yellow color anthers"] * int(sample_size
["this flower has a ring of slightly overlapping rounded peach petals with darker veins."]
["this flower is pink in color, with petals that are leaf like."] * int(sample_size/ni) + \
["this flower has petals with pointed yellow tips and a red color closer to the base."] * i
["this flower is pink and yellow in color, with petals that are curled around the center."]
```

Σχήμα 6.4: Συνθετική εικόνα που παράγει το μοντέλο για τις αντίστοιχες λεκτικές περιγραφές, όπως δόθηκαν στην υλοποιήση του κώδικα. Η παραπάνω εικόνα προέρχεται από τις τελευταίες εποχές εκπαίδευσης του μοντέλου.

Απο τις παραπάνω συνθετικές εικόνες μπορεί κάποιος εύκολα να διακρίνει τη σαφέστατη βελτίωση της ποιότητας των εικόνων όσο προχωρούν οι εποχές εκπαίδευσης του μοντέλου. Παρόλα αυτά όμως, υπάρχουν και συνθετικές εικόνες που είναι αρκετά δύσκολο να διακρίνει κάποιος τι απεικονίζεται, όπως για παράδειγμα στο Σχήμα 6.4 στην υποεικόνα που βρίσκεται στη 4η γραμμή και 3η στήλη. Το γεγονός όμως οτι παρά τις δυσκολίες στην εκπαίδευση που αφορούσαν το μικρό σύνολο εποχών, το μοντέλο παρέχει αρκετά καλά αποτελέσματα είναι άκρως ενθαρρυντικό για την απόδοση του σε περίπτωση πιο αποτελεσματικής εκπαίδευσης.

Κεφάλαιο 7

7. Συμπεράσματα

Είναι σαφές πως τα GAN's, παρά το γεγονός οτι αποτελούν μία σχετικά πρόσφατη τεχνολογία έχουν συνεισφέρει αρκετά στην επιστήμη των υπολογιστών και όπως φαίνεται θα παίξουν πολύ σημαντικό ρόλο σε μελλοντικές κατευθύνσεις, όταν και θα έχουν μελετηθεί ακόμα περισσότερο.

Στα πλαίσια της παρούσας διπλωματικής εργασίας, πραγματοποιήθηκε μελέτη των GAN's και των δυνατοτήτων αυτών στην επίλυση ενός αρκετά σύνθετου προβλήματος, όπως είναι η σύνθεση εικόνας από κείμενο. Μελετήθηκε ο τρόπος λειτουργίας των δικτύων αυτών, καθώς και η σύνδεση που έχουν με τα πεδία της 'Ορασης Υπολογιστών και Επεξεργασίας Φυσικής Γλώσσας, και έγινε παρουσίαση των δυνατοτήτων τους αναπαράγοντας συνθετικές εικόνες με χρήση ενός συνδυασμού μοντέλων. Σίγουρα η εργασία αυτή ακουμπάει μονάχα την επιφάνεια των δυνατοτήτων των GAN's, και παρουσιάζει ελλείψεις αναφορικά με το αρχικό πλάνο υλοποίησης, όπως για παράδειγμα τη χρήση επιπρόσθετων συνόλων δεδομένων για σύνθεση εικόνας από κείμενο, αλλά αποτελεί μία αρκετά καλή πρώτη επαφή με το αντικείμενο το οποίο ήταν άκρως απαιτητικό και σύνθετο.

Βιβλιογραφία

[AM2020]
Αθανάσιος Μασούρης, "Δημιουργία Εικόνας από Κείμενο με χρήση Γεννητικών Ανταγωνιστικών Δικτύων", 2020

[AR2015]
Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks",2015

[AT2009]
AM Turing, "Computing machinery and intelligence", Parsing the Turing test, 2009

[CB2018]
Cristian Bodnar, "Text to Image Synthesis Using Generative Adversarial Networks",2018

[CS2015]
Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna, "Rethinking the Inception Architecture for Computer Vision", 2015

[CLS2016]

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, Honglak Lee. " Generative Adversarial Text to Image Synthesis". 2016.

[GOO2014]
Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, "Generative Adversarial Networks", 2014

[GS2018]
Gerasimos Spanakis, Ajkel Mino, " LoGAN: Generating Logos with a Generative Adversarial Neural Network Conditioned on color",2018

[HZ2016]
Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, Dimitris Metaxas, " StackGAN: "Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks " , 2016

[HZ2018]
Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, , Dimitris N. Metaxas, "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks" , 2018

[MCP1943]
WS McCulloch, W Pitts, "A logical calculus of the ideas immanent in nervous activity", The bulletin of mathematical biophysics, 1943

[MEN2008]
Maria-Elena Nilsback, Andrew Zisserman, "Automated flower classification over a large number of classes", 2008

[MH2018]
Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Sepp Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium", 2018

[MM2014]
Mehdi Mirza, Simon Osindero, " Conditional Generative Adversarial Nets ",2014

[MZ2019]
Minfeng Zhu, Pingbo Pan, Wei Chen and Yi Yang, "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthesis", 2019

[SR2016]
Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee. "Learning deep representations of _ne-grained visual descriptions", 2016.

[TQ2019]
Tingting Qiao, Jing Zhang, Duanqing Xu and Dacheng Tao, "MirrorGAN: Learning Text-To-Image Generation by Redescription", 2019.

[TS2016]
Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, "Improved Techniques for Training GANs", 2016

[TX2017]
Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He, "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks", 2017

[WL2019]
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

[YL2018]
Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, Jianfeng Gao, "StoryGAN: A Sequential Conditional GAN for Story Visualization", 2018