Αυτοκωδικοποιητές και Παραγωγικά Δίκτυα Μάθησης με Αντιπαλότητα

Τεχνητά Νευρωνικά Δίκτυα και Μηχανική Μάθηση

Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ

Εθνικό Μετσόβιο Πολυτεχνείο

Γιώργος Αλεξανδρίδης

Αυτοκωδικοποιητές

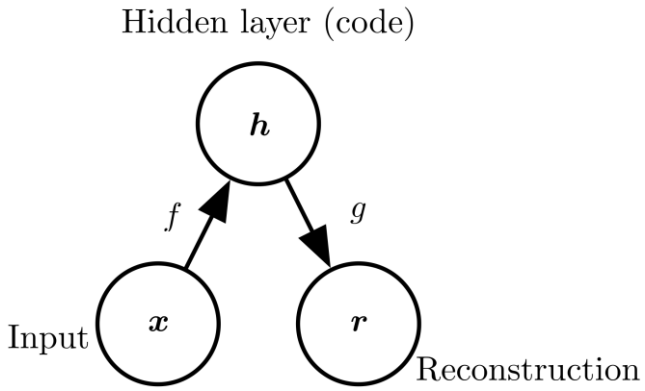
Autoencoders

Διακριτικά και Παραγωγικά Μοντέλα

- **Διακριτικά** (*discriminative*) μοντέλα
  - Λειτουργία ταξινόμησης
    - Δεδομένων των χαρακτηριστικών ενός δείγματος $x \in X$ του χώρου της εισόδου, προέβλεψε την κλάση (ετικέτα) του y
    - Εκμάθηση της κατανομής $p(Y|X = x) \Longrightarrow$ όρια μεταξύ των κλάσεων
  - Παραδείγματα: MLP, SVM, RBF, …
- **Παραγωγικά** (*generative*) μοντέλα
  - Λειτουργία ταξινόμησης
    - Δεδομένης μια κλάσης $y \in Y$, προέβλεψε τα χαρακτηριστικά των δειγμάτων X που ανήκουν σε αυτή
    - Εκμάθηση της κατανομής $p(X|Y = y)$ των επιμέρους κλάσεων των δεδομένων
  - Παραδείγματα: Απλός Μπεϋζιανός ταξινομητής, Γκαουσιανά μοντέλα μίξης, …

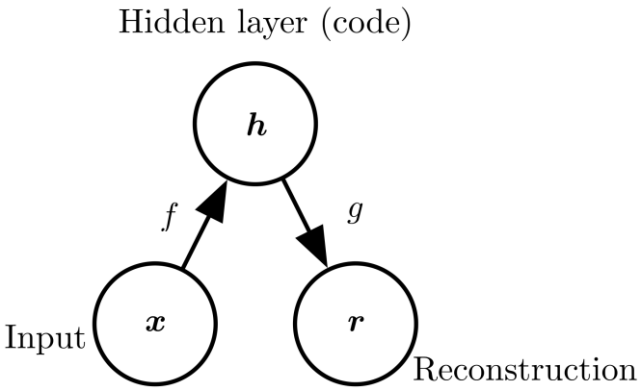
Αυτοκωδικοποιητές

- **Αυτοκωδικοποιητής** (*Autoencoder* ή ΑΚ)
  - Νευρωνικό δίκτυο που εκπαιδεύεται να αντιγράφει την είσοδό του στην έξοδό του
  - Εσωτερικά αποτελείται από κρυφό επίπεδο $\boldsymbol{h}$ στο οποίο αναπαρίσταται **κωδικοποιημένη** (*coded*) η είσοδος
  - Δύο μέρη
    1. Συνάρτηση **κωδικοποίησης** (*encoder function*) $\boldsymbol{h} = f(\boldsymbol{x})$
    2. Συνάρτηση **αποκωδικοποίησης** (*decoder function*) $\boldsymbol{r} = g(\boldsymbol{h})$



Αυτοκωδικοποιητές

- Εκπαίδευση
  - Μέσω οπίσθιας διάδοσης του σφάλματος
    - Όπως στα ΤΝΔ πρόσθιας τροφοδότησης
  - Μέσω επανακυκλοφορίας (recirculation)
    - Συγκρίνεται η ενεργοποίηση των νευρώνων στην αρχική είσοδο και στην αναπαράσταση
- Τετριμμένη λύση: Μάθε την $g(f(\boldsymbol{x})) = \boldsymbol{x}, \forall \boldsymbol{x}$
  - Δεν έχει νόημα γι' αυτό οι ΑΚ σχεδιάζονται έτσι ώστε **να μην μπορούν** να αντιγράψουν τέλεια
- Παρουσιάστηκαν στα μέσα του 1980
- Χρήση σε προβλήματα **μείωσης διαστατικότητας** (*dimensionality reduction*) και **εξαγωγής χαρακτηριστικών** (*feature extraction*)



Διαδικασία Μάθησης

- Δεν μας ενδιαφέρει τόσο η διαδικασία αντιγραφής, όσο να αποτυπωθούν στο $\boldsymbol{h}$ χρήσιμες ιδιότητες του $\boldsymbol{x}$
- **Υποπλήρης** (*undercomplete*) ΑΚ: Περιορίζουμε το $\boldsymbol{h}$ ώστε να έχει μικρότερες διαστάσεις από το $\boldsymbol{x}$
  - Μαθαίνει τα **προεξέχοντα** (*salient*) χαρακτηριστικά της εισόδου
- **Διαδικασία μάθησης**: Ελαχιστοποίηση συνάρτησης απώλειας $L(x, g(f(x))$
  - Αν g γραμμική και L Μέσο Τετραγωνικό Σφάλμα (ΜΤΣ), τότε ο υποπλήρης ΑΚ καλύπτει τον ίδιο **υποχώρο** με την PCA
  - Αν f, g μη γραμμικές, ο υποπλήρης ΑΚ μαθαίνει μια πιο **ισχυρή μη-γραμμική γενίκευση** της PCA
- **Υπερπλήρης** (*overcomplete*) ΑΚ: Θέτουμε $\dim(\boldsymbol{h}) \geq \dim(\boldsymbol{x})$
- Η επιλογή της διάστασης του ΑΚ καθώς και της χωρητικότητας των f,g εξαρτάται από την πολυπλοκότητα της **υποκείμενης** (*underlying*) κατανομής των δεδομένων που μοντελοποιούνται

Εκμάθηση Πολλαπλότητας

- **Γενική υπόθεση ΑΚ**
  - Τα δεδομένα είναι συγκεντρωμένα γύρω από **πολλαπλότητες** (*manifolds*) χαμηλότερων διαστάσεων
- Πολλαπλότητες χαρακτηρίζονται από τις **εφαπτόμενες πλευρές** τους (*tangent planes*)
  - Πολλαπλότητα **διάστασης** d ορίζεται από d **διανύσματα βάσης**
  - Διανύσματα βάσης ορίζουν τις επιτρεπόμενες αποκλίσεις εντός της πολλαπλότητας
    - Πόσο μπορεί να «αλλάξει» το x παραμένοντας ενός της πολλαπλότητας

Εκμάθηση Πολλαπλότητας

- Συμβιβασμός μεταξύ **δύο** αντίρροπων δυνάμεων
  1. Εκμάθησης αναπαράστασης $\boldsymbol{h}$ από τα δεδομένα $\boldsymbol{x}$
     - έτσι ώστε το $\boldsymbol{x}$ να μπορεί να ανακτηθεί προσεγγιστικά από το $\boldsymbol{h}$ μέσω του ΑΚ
  2. Ικανοποίησης περιορισμών που περιορίζουν την χωρητικότητα του ΑΚ
     - λχ ποινές ομαλοποίησης
     - ΑΚ μαθαίνει αναπαραστάσεις που είναι **λιγότερο «ευαίσθητες»** στη μεταβολή της εισόδου
- Τελικά, ο ΑΚ μαθαίνει μόνο τις **διακυμάνσεις** που είναι απαραίτητες για την ανακατασκευή των δειγμάτων εκπαίδευσης

Ομαλοποιημένοι Αυτοκωδικοποιητές

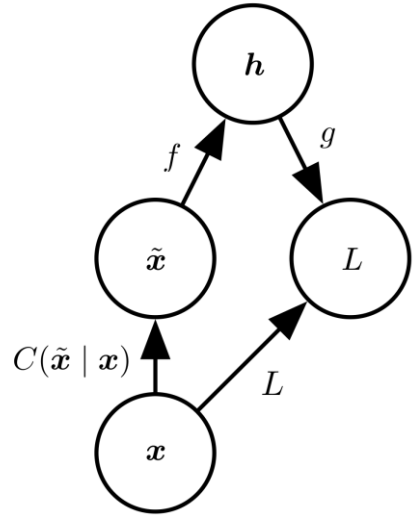
- **Ομαλοποιημένοι** (*regularized*) ΑΚ
  - Περιορισμός της **χωρητικότητάς** τους μέσω συνάρτησης απώλειας
- Συνάρτηση απώλειας επιβάλλει πρόσθετες ιδιότητες πέραν της αντιγραφής
  - **Αραιότητα** (*sparsity*) της αναπαράστασης
  - **Ανοχή** σε **θόρυβο** ή σε απουσιάζουσες τιμές
- Ένας *ομαλοποιημένος* ΑΚ μπορεί να είναι *μη-γραμμικός* και *υπερπλήρης* αλλά παρόλα αυτά να **μαθαίνει χαρακτηριστικά** της κατανομής των δεδομένων

Αραιοί Αυτοκωδικοποιητές

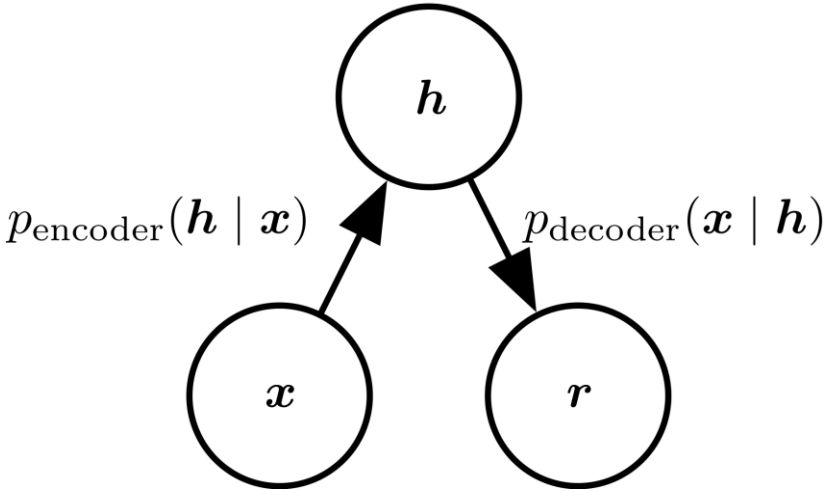
- **Αραιός** (*sparse*) ΑΚ
  - Προσθήκη όρου ποινής αραιότητας $\Omega(\boldsymbol{h})$ του επιπέδου κωδικοποίησης $\boldsymbol{h}$ στη διαδικασία μάθησης
  - $L(\boldsymbol{x},\ g(f(\boldsymbol{x})) + \Omega(\boldsymbol{h})$
- Προσθήκη όρων ομαλοποίησης στη διαδικασία μάθησης
  - **Προσοχή**: Η ποινή αραιότητας δεν εφαρμόζεται στις παραμέτρους του μοντέλου αλλά στις λανθάνουσες μεταβλητές
  - Αποτελεί **συνέπεια** της θεώρησης για την **κατανομή** του μοντέλου όσον αφορά τις λανθάνουσες μεταβλητές
- Η **εκπαίδευση** ενός **ΑΚ** είναι αντίστοιχη της εκπαίδευσης ενός **παραγωγικού μοντέλου**
  - Τα χαρακτηριστικά που μαθαίνει ο ΑΚ είναι επί της ουσίας οι λανθάνουσες μεταβλητές που χαρακτηρίζουν την είσοδο

Αυτοκωδικοποιητές απαλοιφής θορύβου

- **ΑΚ απαλοιφής θορύβου** (*Denoising Autoencoders* ή DAE)
  - Ελαχιστοποίηση της $L(\boldsymbol{x}, g(f(\tilde{x}))$
  - $\tilde{x}$ αλλοίωση του $\boldsymbol{x}$ μέσω της προσθήκης θορύβου
  - $C(\boldsymbol{x}|\tilde{x})$: Υπό **συνθήκη κατανομή** παραγωγής αλλοιωμένων $\tilde{x}$ δεδομένου $\boldsymbol{x}$
- Ο ΑΚ μαθαίνει την **κατανομή αποκατάστασης** (*reconstruction distribution*)
  - Λήψη δείγματος $\boldsymbol{x}$ από τα δεδομένα εκπαίδευσης και αλλοιωμένης του μορφής $\tilde{x}$ από την $C(\tilde{x}|\boldsymbol{x} = x)$
  - Χρήση $(\boldsymbol{x}, \tilde{x})$ ως **δείγμα εκπαίδευσης** για την εκτίμηση $p_r(\widetilde{x}|\boldsymbol{x}) = p_d(\boldsymbol{x}|\boldsymbol{h})$
    - p_d ορίζεται από $g(\boldsymbol{h})$
- Αν θέσουμε f ντετερμινιστική, ο DAE συμπεριφέρεται ως ένα ΤΝΔ πρόσθιας τροφοδότησης
  - Συνεπώς μπορούμε να χρησιμοποιήσουμε αντίστοιχές τεχνικές μάθησης (λχ κατάβαση κλίσης)



Στοχαστικοί αυτοκωδικοποιητές



- Διαδικασία κωδικοποίησης και αποκωδικοποίησης στοχαστική
  - Συνάρτηση Κωδικοποίησης $f \Longrightarrow$ κατανομή κωδικοποίησης $p_{encoder}(\boldsymbol{h}|\boldsymbol{x})$
  - Αντίστοιχα $g \Longrightarrow p_{decoder}(\boldsymbol{x}|\boldsymbol{h})$
- Κάθε μοντέλο λανθανουσών μεταβλητών $p_m(\boldsymbol{x}, \boldsymbol{h})$ μπορεί να ορίσει στοχαστικό κωδικοποιητή και αποκωδικοποιητή
  - $p_{encoder}(\boldsymbol{h}|\boldsymbol{x}) = p_m(\boldsymbol{h}|\boldsymbol{x})$
  - $p_{decoder}(\boldsymbol{x}|\boldsymbol{h}) = p_m(\boldsymbol{x}|\boldsymbol{h})$
- Στη γενική περίπτωση, ωστόσο, οι $p_{encoder}$ και $p_{decoder}$ δεν αποτελούν υπό συνθήκη κατανομές μιας ενιαίας κοινής κατανομής

Παραγωγικά Δίκτυα Μάθησης με Αντιπαλότητα

Generative Adversarial Networks (GANs)

Παραγωγικά Μοντέλα Μάθησης με Αντιπαλότητα

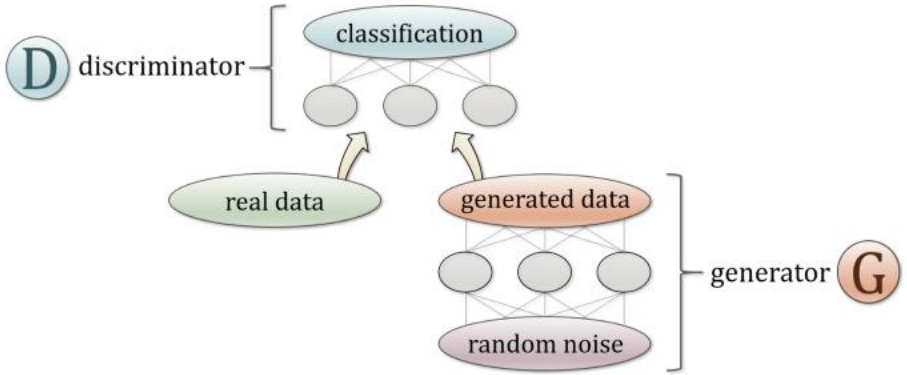


Figure 1. A basic GAN architecture.

- **Generative Adversarial Networks** (GANs)
- Σύστημα Μηχανικής Μάθησης που προτάθηκε το **2014** από τον **Goodfellow**, το οποίο αποτελείται από 2 διακριτά νευρωνικά δίκτυα
  - **Διευκρινιστής** (*discriminator*)
    - Αποφασίζει αν τα δεδομένα εισόδου του είναι αληθινά
  - **Γεννήτορας** (*generator*)
    - Προσπαθεί να δημιουργήσει δεδομένα τα οποία ο διευκρινιστής θα θεωρήσει ως αληθινά

Εκπαίδευση GAN

- Στην απλούστερη περίπτωση, ο γεννήτορας δεν «βλέπει» ποτέ τα πραγματικά δεδομένα
  - Πρέπει να «μάθει» να τα δημιουργεί
- Τεχνική Μάθησης: **Απώλεια Μάθησης με Αντιπαλότητα** (*Adversarial loss*)
  - Αρχικά ο γεννήτορας παράγει τυχαίο θόρυβο
  - Με βάση την απόκριση που λαμβάνει από τον διευκρινιστή, μαθαίνει τελικά την κατανομή των χαρακτηριστικών της εισόδου
- Για την αποφυγή υπερπροσαρμογής, μπορούν να χρησιμοποιηθούν **τεχνικές ομαλοποίησης** κατά τη διαδικασία της μάθησης
  - λ.χ. DropOut
- Συνεχές ανταγωνιστικό «παιχνίδι»
  - Διευκρινιστής μαθαίνει να εντοπίζει *έκτοπες τιμές*, *σφάλματα* και γενικότερα κάθε τι εκτός της «*κανονικότητας*» των δεδομένων
  - Ο γεννήτορας μαθαίνει την κατανομή των δεδομένων εισόδου

Πρόβλημα της Κυριαρχίας

- Αν το ένα δίκτυο **κυριαρχήσει** επί του άλλου, η εκπαίδευση σταματά και για τα δύο
- Περίπτωση δυαδικής ταξινόμησης
  - Ο διευκρινιστής έχει σιγμοειδή συνάρτηση ενεργοποίησης
  - Αν για τον *οποιονδήποτε λόγο*, το ανταγωνιστικό παιχνίδι «κολλήσει» σε μια από τις δύο **ακριανές περιοχές**, η εκπαίδευση σταματά

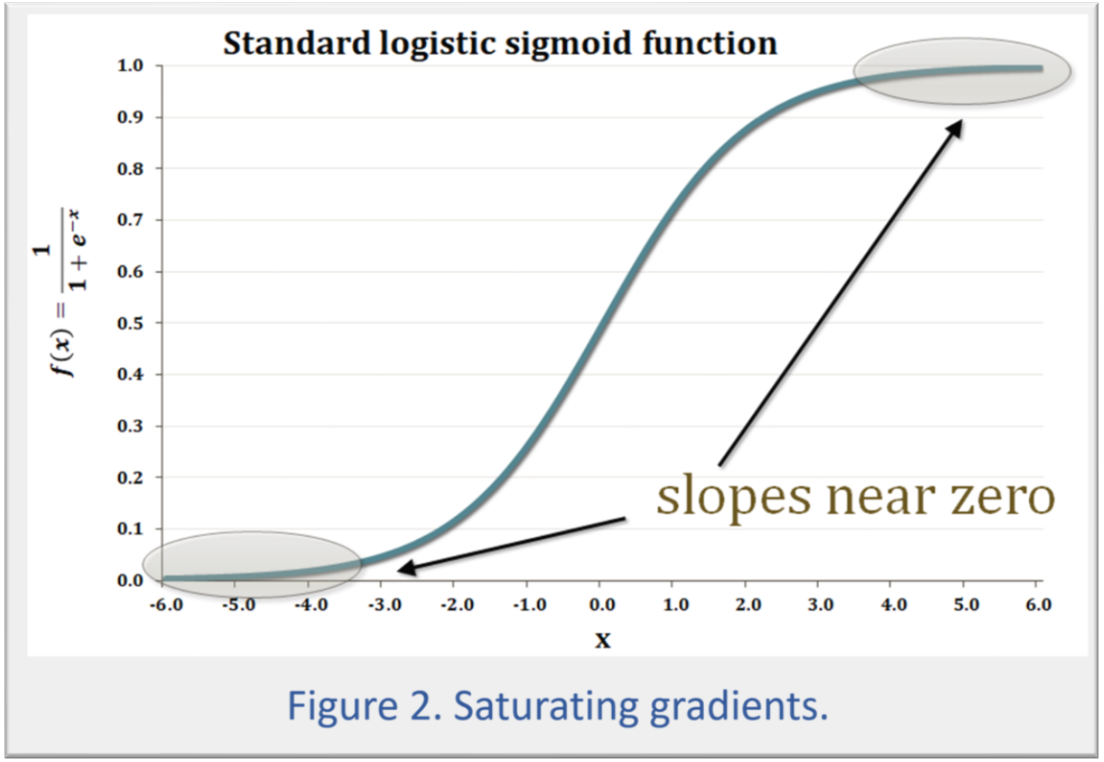


Figure 2. Saturating gradients.

Wasserstein GAN (WGAN)

- Προτάθηκε το **2017** από τον **Wasserstein**
- Η **συνάρτηση αποτίμησης** του διευκρινιστή δεν είναι πλέον η σιγμοειδής αλλά η γραμμική
  - Σε περίπτωση κυριαρχίας του ενός δικτύου επί του άλλου, ο γεννήτορας δε θα λαμβάνει συνεχώς τις ίδιες τιμές (0 ή 1)
- Επειδή ο διευκρινιστής δεν επιστρέφει τιμές στο [0, 1] αλλά σε ένα ευρύτερο σύνολο, ονομάζεται πλέον **κριτής** (*critic*)

Wasserstein GAN (WGAN)

- **Συνάρτηση απώλειας κριτή**
  - D_{REAL}: Εκτίμηση κριτή κατά πόσο το δείγμα είναι πραγματικό
  - D_{FAKE}: Εκτίμηση κριτή κατά πόσο το δείγμα είναι ψεύτικο
- Μεταβολή προς την αντίθετη κατεύθυνση της **κλίσης**
  - $L_C = -(D_{REAL} - D_{FAKE})$
  - $L_{\mathrm{G}} = -D_{FAKE}$
- Ένας καλός κριτής επιστρέφει ψηλές τιμές για τα πραγματικά δείγματα και χαμηλές για τα ψεύτικα
- Στόχος γεννήτορα είναι ο **ανάποδος**
  - Θέλει να κάνει τον κριτή να παράξει υψηλές τιμές για τα δείγματα που παράγει

Περιορισμένο εύρος αποτελεσμάτων

- Συχνά ο γεννήτορας καταλήγει να επαναδημιουργεί ένα συγκεκριμένο μόνο εύρος των δεδομένων εισόδου
  - Αν το **μοναδικό** του **κριτήριο** είναι να «**κοροϊδέψει**» τον κριτή, τότε μπορεί να το πετύχει δημιουργώντας **δεδομένα** από τις πιο *πυκνές περιοχές* της κατανομής τους

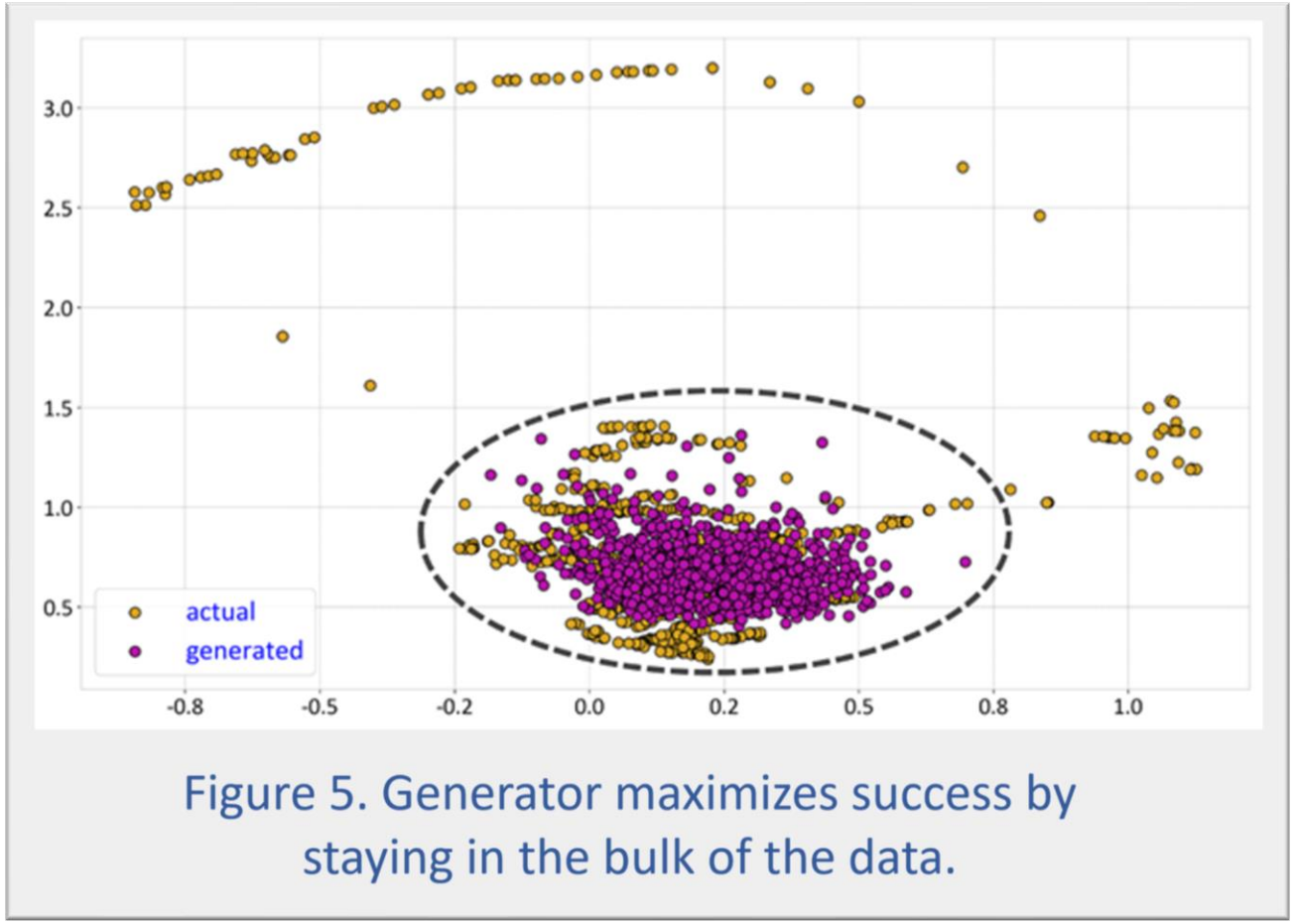


Figure 5. Generator maximizes success by staying in the bulk of the data.

Αύξηση εύρους αποτελεσμάτων

- Λειτουργία του κριτή σε **μικρο-δέσμες εισόδων** (*minibatch discrimination*)
  - Ο κριτής συλλέγει στατιστικές πληροφορίες για τα δείγματα που του παρέχει ο γεννήτορας
  - Αν η κατανομή των παραγόμενων δειγμάτων είναι πολύ διαφορετική από αυτή των πραγματικών, ο κριτής μπορεί να ζητήσει ο γεννήτορας να μεταβάλλει το εύρος του
- Παροχή και στον κριτή και στον γεννήτορα **επιπλέον μεταβλητών**
  - Έχουν τον ρόλο της **υπό συνθήκη πληροφορίας** (*conditional information*)
  - Ενημερώνουν τα δίκτυα για το περιβάλλον στο οποίο λειτουργούν, κατευθύνοντας τις εξόδους τους

Βελτίωση εκπαίδευσης των WGANs: Βελτιστοποιητές (optimizers)

- Η έξοδος του κριτή είναι πλέον **μη-σταθερή** (*non-stationary*)
  - Δεν είναι στο [0,1]
- Χρήση βελτιστοποιητών που δεν περιέχουν **παράγοντα ορμής** (*momentum*)
  - *RMSProp*, *Adam* / *SGD* χωρίς ορμή
  - Ο παράγοντας ορμής «**συμπιέζει**» (*squashes*) την έξοδο στο [0,1]

Βελτίωση εκπαίδευσης των WGANs: Ποινή κλίσης (Gradient penalty)

- Αποτρέπει τις κλίσεις της γραμμικής εξόδου από το να γίνουν πολύ «απότομες»
- **Διαδικασία**
  1. Δημιουργία με τυχαίο τρόπο των πραγματικών δεδομένων από τα δεδομένα του γεννήτορα
  2. Υπολογισμός της κλίσης, μετρώντας την μεταβολή της εξόδου του κριτή
  3. Υπολογισμός της νόρμας της κλίσης και δημιουργία παράγοντα ποινής, ανάλογα με το πόσο αυτή απέχει από το 1
  4. Προσθήκη του όρου ποινής του προηγούμενου βήματος στη συνάρτηση απώλειας του κριτή
- Πλεονέκτημα
  - Αύξηση της ευστάθειας των GANs
- Μειονέκτημα
  - Αύξηση χρόνου εκπαίδευσης για τον υπολογισμό του όρου ποινής

Εξέλιξη των GANs

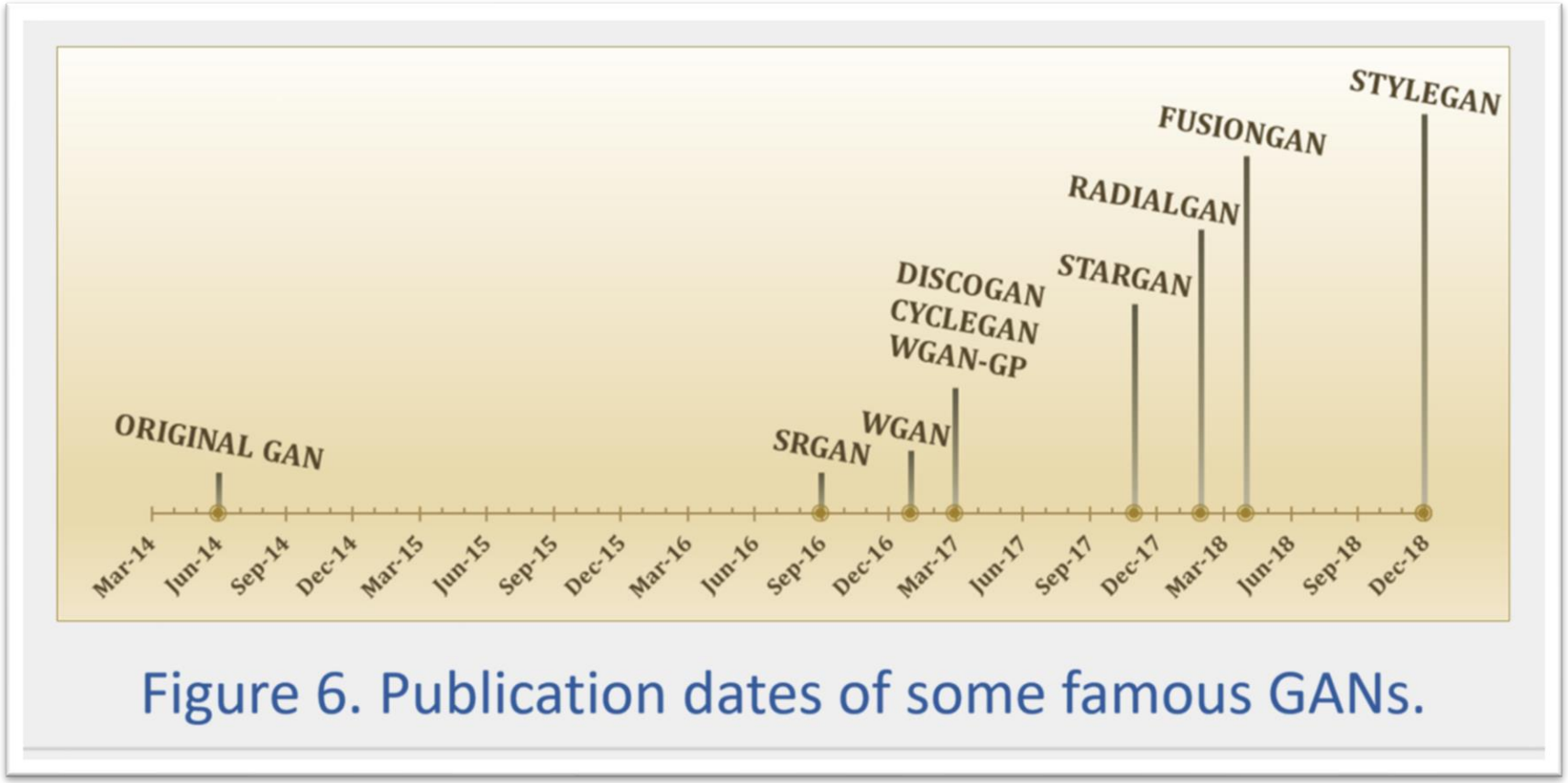


Figure 6. Publication dates of some famous GANs.

Αύξηση των ερευνητικών εργασιών σε GANs μετά την εμφάνιση του WGAN

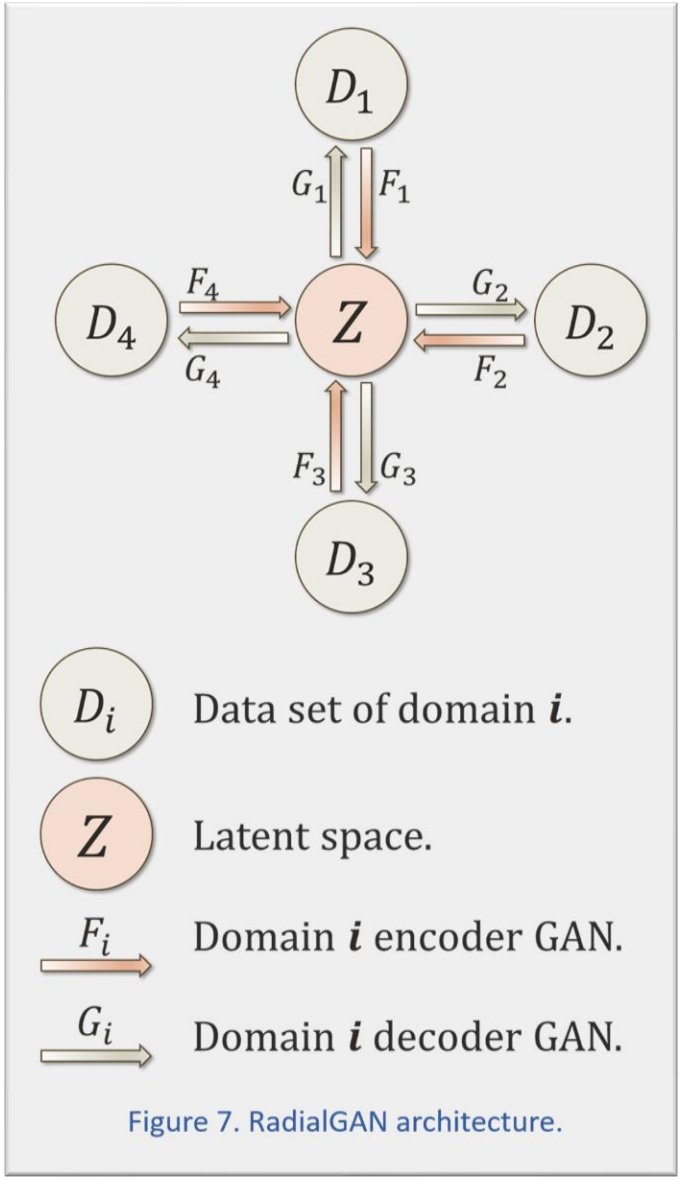


Figure 7. RadialGAN architecture.

1ο Παράδειγμα: RadialGAN

- Κατάλληλο για συνδυασμό δεδομένων από πολλαπλά **πεδία** (*domains*)
  - **Μεταφορά πεδίου** (*domain transfer*)
- Μετασχηματισμός σε **λανθάνοντα χώρο** (*latent space*)
- Κάθε σύνολο δεδομένων έχει δικό του:
  1. Δίκτυο **κωδικοποίησης** (*encoder network*)
  2. Δίκτυο **αποκωδικοποίησης** (*decoder network*)
     - Ο διευκρινιστής επιβεβαιώνει ότι η πληροφορία που έρχεται από το λανθάνοντα χώρο ταιριάζει με τις ιδιότητες του συγκεκριμένου πεδίου
- **Κυκλική συνέπεια** (*cycle consistency*)
  - Μεταφορά πληροφορίας μεταξύ του λανθάνοντα χώρου και των πεδίων και προς τις δύο κατευθύνσεις
- Παράλληλη εκπαίδευση όλων των δικτύων οδηγεί στη δημιουργία *επαυξημένου συνόλου δεδομένων*

2ο Παράδειγμα: StyleGAN

- Συνδυασμός **Progressive GAN** (ProGAN) και **μεταφοράς στυλ μέσω νευρώνων** (neural style transfer)
- **ProGAN**
  - Αναπτύσσει αρχική εικόνα μικρής διάστασης (πχ 4×4 ή 8×8 pixel) μέχρι αυτή να θεωρηθεί ρεαλιστική από τον διευκρινιστή
  - Στην επόμενη φάση, προστίθεται ένα επίπεδο υψηλότερης ανάλυσης και επαναλαμβάνεται η εκπαίδευση
  - Η διαδικασία ολοκληρώνεται όταν φτάσουμε στο επιθυμητό επίπεδο ανάλυσης (πχ 1024×1024 pixel)
- **StyleGAN**
  - Μεταφορά στυλ μέσω της προσαρμογής των βαρών του επιπέδου του στυλ στο επίπεδο της εικόνας

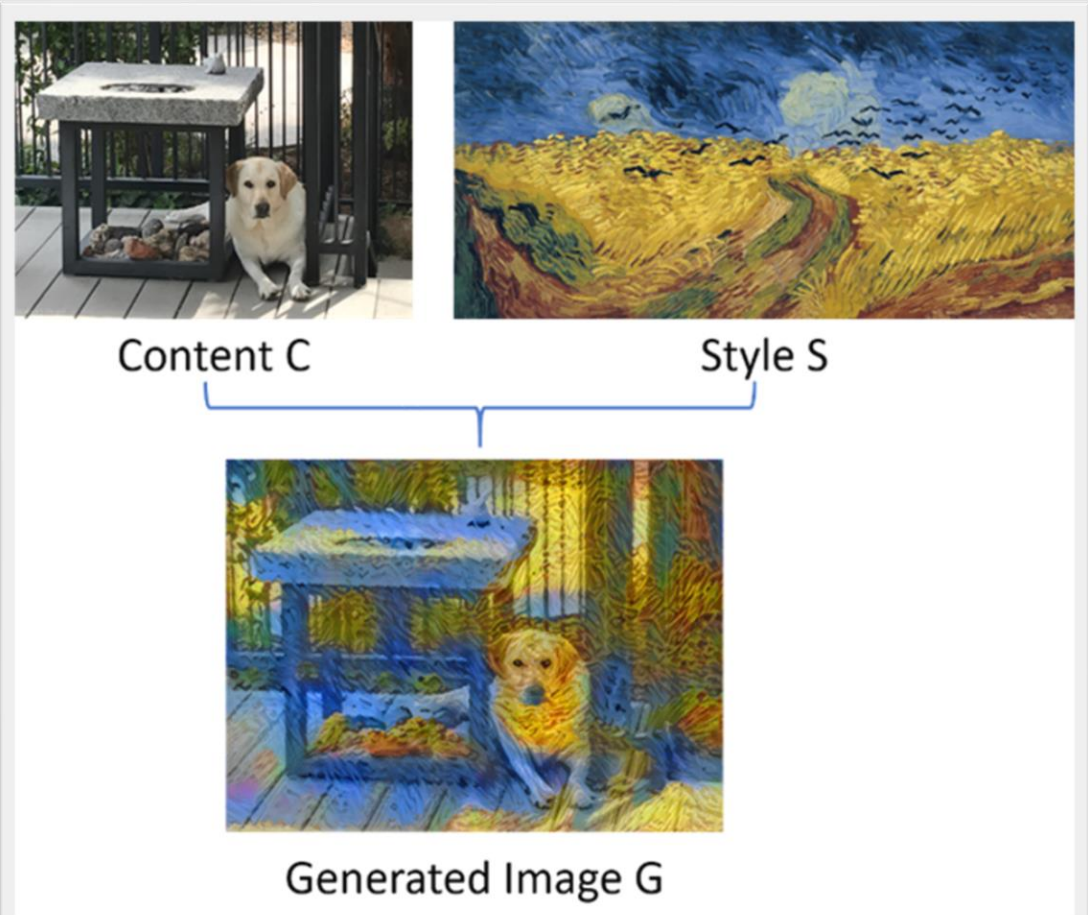


Figure 10. Neural style transfer redraws a content image in the design of a style image.

AdaIN: Adaptive Instance Normalization

- Τεχνική μεταφοράς στυλ που χρησιμοποιείται στο StyleGAN
- Δεν περιλαμβάνει βελτιστοποίηση και συνεπώς είναι μια γρήγορη τεχνική
- Εφαρμογή του AdaIN σε συνελικτικό επίπεδο εικόνας
  1. Κανονικοποίηση του επιπέδου
     - Αφαίρεση μέσης τιμής και διαίρεση με τυπική απόκλιση
  2. Κλιμάκωση του κανονικοποιημένου επιπέδου έτσι ώστε να προσαρμοστεί στην τυπική απόκλιση του επιπέδου του στυλ
  3. **Ολίσθηση** (*shift*) των βαρών του κανονικοποιημένου και του κλιμακούμενου επιπέδου μέσω της προσθήκης της μέσης τιμής του επιπέδου του στυλ

This person does not exist

https://www.thispersondoesnotexist.com/

Βιβλιογραφία για αυτοκωδικοποιητές

- Ian Goodfellow, Yoshua Bengio, Aaron Courville “Deep Learning” – MIT Press (https://www.deeplearningbook.org/)
  - Αυτοκωδικοποιητές (Κεφ. 14)
    - Εισαγωγή (§14.1)
    - Ομαλοποιημένοι Αυτοκωδικοποιητές (§14.2, §14.5, §14.7)
    - Στοχαστικοί Αυτοκωδικοποιητές (§14.4)

Βιβλιογραφία για GANs

- Επιστημονικά άρθρα
  - Ian J. Goodfellow et al. Generative Adversarial Networks
  - Martin Arjovsky et al. Wasserstein GAN
  - Alec Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
  - Tim Salimans et al. Improved Techniques for Training GANs
  - Ishaan Gulrajani et al. Improved Training of Wasserstein GANs
  - Lars Mescheder et al. Which Training Methods for GANs do Actually Converge?
  - Tero Karras et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation
  - Leon A. Gatys et al. A Neural Algorithm of Artistic Style
- Άρθρα στον Παγκόσμιο Ιστό
  - A Beginner's Guide to Generative Adversarial Networks (GANS)
  - A Leap into the Future: Generative Adversarial Networks
  - Understanding Generative Adversarial Networks (GANs)