



Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Κατηγοριοποίηση εικόνων με λίγα ή καθόλου παραδείγματα

Απόστολος Παναγιωτόπουλος

Επιβλέποντες: Χρήστος Δίου
Αναστάσιος Ντελόπουλος

Θεσσαλονίκη, Ιούνιος 2020

Περίληψη

Η παρούσα διπλωματική εργασία μελετάει την δυνατότητα υλοποίησης ενός ταξινομητή εικόνων που έχει εκπαιδευτεί σε ένα μόνο υποσύνολο των κατηγοριών των συνολικών εικόνων που καλείται να ταξινομήσει, πρόβλημα γνωστών στην βιβλιογραφία ως zero-shot learning. Η προσέγγιση στηρίζεται στο γεγονός πως οι κατηγορίες των εικόνων δεν δίνονται ως απλές κατηγορίες, αλλά περιέχουν πληροφορία για την εκάστοτε κατηγορία· πληροφορία που μπορεί να συσχετιστεί με την οπτική πληροφορία της εικόνας για την ταξινόμηση εικόνων από άγνωστες κατηγορίες ή κατηγορίες με λίγα παραδείγματα.

Προς την επίτευξη αυτού του σκοπού, χρησιμοποιούνται διανυσματικές αναπαραστάσεις τόσο των εικόνων όσο και των κατηγοριών των εικόνων. Για τις εικόνες χρησιμοποιείται το προτελευταίο επίπεδο του ResNet-101 ως διάνυσμα αναπαράστασης, αν και θα μπορούσε να χρησιμοποιηθεί οποιοδήποτε άλλον αντίστοιχο νευρωνικό δίκτυο. Για τις κατηγορίες των εικόνων στα μικρά dataset έχουν προταθεί διανύσματα κατασκευασμένα από ανθρώπους, ενώ για μεγάλα dataset όπως το ImageNet μπορούμε να χρησιμοποιήσουμε μεθόδους όπως το word2vec.

Έπειτα, αφού βρεθούν διανυσματικές αναπαραστάσεις τόσο των εικόνων όσο και των κατηγοριών τους, συγχρονίζονται 2 Variational Autoencoders (VAE) ώστε να κωδικοποιήσουν τις διανυσματικές αυτές αναπαραστάσεις σε έναν χώρο κοινής ενσωμάτωσης (embedding space). Ο συγχρονισμός γίνεται έτσι ώστε μια εικόνα να αντιστοιχηθεί στην ίδια περιοχή που αντιστοιχήθηκε το διάνυσμα της κατηγορίας της, ενώ οι περιοχές από τις διάφορες κατηγορίες καταλήγουν να είναι καλά διαχωρίσιμες και οι αποστάσεις τους να είναι αντιστρόφως ανάλογες της σημασιολογικής τους συσχέτισης.

Τέλος εκπαιδεύεται ένας απλός ταξινομητής, ώστε να ταξινομεί τις περιοχές του χώρου κοινής ενσωμάτωσης με βάση τις περιοχές κωδικοποίησης στον χώρο αυτό. Έτσι μετά την εκπαίδευση του μοντέλου, οι εικόνες κωδικοποιούνται στον χώρο κοινής ενσωμάτωσης, στην ίδια περιοχή με την κατηγορία τους, και έπειτα ο ταξινομητής υποδεικνύει την κατηγορία που ανήκουν.

Το προτεινόμενο μοντέλο εξετάζεται σε αρκετά σύνολα δεδομένων, τόσο στο πρόβλημα της κατηγοριοποίησης εικόνων χωρίς καθόλου παραδείγματα (zero-shot learning) όσο και στο πρόβλημα της κατηγοριοποίησης με λίγα παραδείγματα (few-shot learning). Γίνονται ποσοτικά και ποιοτικά πειράματα αξιολόγησης, για την καλύτερη κατανόηση της συμπεριφοράς του μοντέλου.

Abstract

This thesis studies the feasibility of an image classifier trained only on a subset of the categories of the dataset, while being able to classify all the categories in the dataset. For this to be possible, image categories are not provided as pure categorical data but contain some information for each category instead; information which can correlate to the optical information of the image for classifying image from seen and unseen categories.

More specifically, both image and their categories are represented by vectors. For images the vector representation is the penultimate layer of ResNet-101, although any similar architecture could be used to extract features. For image categories small datasets provide handcrafted vector descriptions, while for big datasets like ImageNet we can use methods like word2vec.

After vector representation are extracted for both modalities (image and text), two Variational Autoencoders (VAE) are aligned to encode these vector representations in a common embedding space. The alignment of the VAEs is such that an image is encoded in the same region where its category has been encoded, while regions from different categories are well separated and their distance is inversely proportional to their semantical correlation.

Finally, a simple classifier model is trained to classify the regions in the embedding space according to the category encodings. Hence, after the training of the proposed model, images are encoded in the embedding space, hopefully in the same region where their category has been encoded, and then the classifier points out the category they belong.

The proposed model is tested in various datasets, both in the zero-shot and the few-shot setting. The thesis presents quantitative and qualitative experiments to better understand how the proposed method works.

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω όσους βοήθησαν για την επίτευξη αυτής της εργασίας.

Αρχικά τον επιβλέποντα καθηγητή μου, Χρήστο Δίου, που μέσα από τις πολύωρες συναντήσεις μας με βοήθησε να καταλάβω περίπλοκες ιδέες από τον χώρο της μηχανικής μάθησης. Τον ευχαριστώ επίσης για την εμπιστοσύνη που έδειξε για τις ιδέες μου, σύντομα μετά την έναρξη της διπλωματικής.

Επίσης την οικογένεια μου, που επωμίστηκε τις ευθύνες της καθημερινότητας προκειμένου να μπορώ να αφοσιωθώ σε αυτήν την εργασία.

Τέλος, θα ήθελα να ευχαριστήσω κάποιους από τους φίλους μου, που μέσα από τις συζητήσεις μας με ενέπνευσαν όταν το χρειαζόμουν και μου έδωσαν ερεθίσματα που μετεξελίχθηκαν σε ιδέες.

Χωρίς όλους αυτούς η παρούσα εργασία δε θα ήταν αυτή που είναι, αλλά κι εγώ δε θα ήμουν το ίδιο άτομο.

Ακρωνύμια και συντομογραφίες

Από εδώ και στο εξής θα χρησιμοποιούνται τα παρακάτω ακρωνύμια στην έκταση του κειμένου:

ZSL	Zero-Shot Learning
GZSL	Generalized Zero-Shot Learning
FSL	Few-Shot Learning
GFSL	Generalized Few-Shot Learning
CNN	Convolutional Neural Network
MLP	Multi-layer Perceptron
SGD	Stochastic Gradient Descent
GAN	Generative Adversarial Network
VAE	Variational Auto-Encoder
CVAE	Conditional Variational Auto-Encoder
SVM	Support Vector Machine
GMM	Gaussian Mixture Model
ELBO	Evidence Lower Bound

Περιεχόμενα

Περίληψη	iii
Abstract	v
Ευχαριστίες	vii
Ακρωνύμια και συντομογραφίες	ix
1 Εισαγωγή	1
1.1 Διατύπωση του προβλήματος	1
1.2 Σπουδαιότητα	2
1.3 Προτεινόμενη αντιμετώπιση	3
1.4 Συνεισφορά της εργασίας	4
1.5 Δομή της εργασίας	4
2 Βιβλιογραφική επισκόπηση	5
2.1 Μοντέλα	5
2.1.1 Περιγραφική ταξινόμηση	5
2.1.2 Συναρτήσεις συμβατότητας	6
2.1.3 Υθριδικά μοντέλα	8
2.1.4 Παραγωγικά μοντέλα	9
2.1.5 Χώροι κοινής ενσωμάτωσης	13
2.2 Σύνολα δεδομένων	16
2.3 Σύγκριση μοντέλων	17
3 Προτεινόμενη αντιμετώπιση	19
3.1 Η ιδέα	19
3.2 Το κίνητρο	20
3.3 Το μοντέλο VaDE	21
3.3.1 Η παραγωγική διαδικασία	22
3.3.2 Η συνάρτηση σφάλματος	22
3.3.3 Ενδεικτικό παράδειγμα - MNIST	25
3.3.4 Τροποποίηση για ημι-επιβλεπόμενη μάθηση	26
3.4 Η αρχιτεκτονική	27
3.4.1 Εξαγωγέας χαρακτηριστικών	27
3.4.2 Κωδικοποιητές	28
3.4.3 Αποκωδικοποιητές	28
3.4.4 Ταξινομητής	28
3.5 Η μεθοδολογία εκπαίδευσης	28

3.5.1 Εκπαίδευση των VaDE	28
3.5.2 Εκπαίδευση του ταξινομητή	29
4 Πειραματική αξιολόγηση	31
4.1 Μεθοδολογία αξιολόγησης	31
4.1.1 Μετρικές αξιολόγησης	31
4.1.2 Σύνολα δεδομένων αξιολόγησης	31
4.2 Διερεύνηση παραμέτρων αρχιτεκτονικής	32
4.2.1 Αριθμός διαστάσεων του χώρου κοινής ενσωμάτωσης	32
4.2.2 Αρχιτεκτονική των VaDE	33
4.2.3 Μοντέλο ταξινομητή	36
4.3 Διερεύνηση παραμέτρων εκπαίδευσης	37
4.3.1 Παράμετροι του αλγορίθμου Adam	37
4.3.2 Αναλογία γνωστών και άγνωστων κλάσεων στην εκπαίδευση του ταξινομητή	39
4.4 Αποτελέσματα	42
4.4.1 Κατηγοριοποίησης εικόνων χωρίς καθόλου παραδείγματα	42
4.4.2 Κατηγοριοποίησης εικόνων με λίγα παραδείγματα	42
4.5 Σχολιασμός αποτελεσμάτων	43
5 Συμπεράσματα και μελλοντικές προεκτάσεις	49

1. Εισαγωγή

1.1 Διατύπωση του προβλήματος

Την τελευταία δεκαετία η επιστήμη της μηχανικής μάθησης έχει κάνει τεράστια άλματα τόσο λόγω ριζοσπαστικών ανακαλύψεων στον τομέα, όσο και λόγω της αύξησης της διαθέσιμης υπολογιστικής ισχύς στα εργαστήρια. Μια από τις περιοχές έρευνας της μηχανικής μάθησης που ακολούθησε την παραπάνω τάση είναι η κατηγοριοποίηση (ή αλλιώς ταξινόμηση)¹ εικόνων.

Συγκεκριμένα, αν και σχετική έρευνα γίνονταν από τις απαρχές της μηχανικής μάθησης, ήταν στην δεκαετία του 2000 που υπήρξαν τα πρώτα ενθαρρυντικά αποτελέσματα σε φυσικές εικόνες [20], [19]. Το 2009 δημοσιεύτηκε το πρώτο τεράστιο σύνολο δεδομένων πάνω σε κατηγοριοποίηση εικόνων, το ImageNet [6], ενώ το 2012 ήρθε το πρώτο επαναστατικό νευρωνικό δίκτυο ταξινόμησης για το ImageNet, το AlexNet [15]. Έκτοτε προτάθηκαν νέα μοντέλα, που πρόσφατα ξεπέρασαν τις ανθρώπινες επιδόσεις.

Ωστόσο ο τρόπος που αυτά τα νευρωνικά δίκτυα μαθαίνουν είναι λίγο ανορθόδοξος για τον άνθρωπο. Όλες αυτές οι ριζοσπαστικές μεθοδολογίες χρειάζονται εκατοντάδες ή χιλιάδες δείγματα ανά κατηγορία, ενώ αντίθετα ο άνθρωπος μόλις σε νεαρή ηλικία μάθει τα βασικά αντικείμενα που υπάρχουν, μπορεί εύκολα να μάθει καινούριες κατηγορίες αντικειμένων. Μπορεί να χρειαστεί να δεις μόλις μερικά δείγματα. Πολλές φορές είναι δυνατόν να φανταστεί ένα αντικείμενο μόλις με την λεκτική περιγραφή του.

Αυτή η ιδέα γέννησε το πρόβλημα του zero-shot learning, δηλαδή της κατηγοριοποίησης χωρίς καθόλου παραδείγματα, και του few-shot learning, δηλαδή της κατηγοριοποίησης με λίγα παραδείγματα (τυπικά 1-10). Αν και έχουν υπάρξει κάποιες παραλλαγές του προβλήματος, θα δώσουμε εδώ τον κλασσικό ορισμό του:

Έστω \mathcal{Y} ένα σύνολο κατηγοριών και $\mathcal{Y} = \mathcal{Y}_s \cup \mathcal{Y}_u$ μια διαμέριση του. Η κλάση \mathcal{Y}_s συμβολίζει τις γνωστές κατηγορίες (seen), ενώ η κλάση \mathcal{Y}_u τις άγνωστες κατηγορίες (unseen). Έστω επίσης $\mathcal{X} = \mathcal{X}_s \cup \mathcal{X}_u$ σύνολο με δείγματα εικόνων από τις αντίστοιχες κατηγορίες. Για κάθε κατηγορία $y_i \in \mathcal{Y}$ υπάρχει μια "περιγραφή" $a_i = \pi(y_i)$ (που στα πλαίσια αυτής της εργασίας μπορεί να θεωρηθεί ως ένα διάνυσμα). Κατά την εκπαίδευση του μοντέλου δίνονται μόνο δείγματα εικόνων που ανήκουν στο \mathcal{X}_s , αλλά περιγραφές a_i για κάθε κατηγορία $y_i \in \mathcal{Y}$. Στόχος είναι το μοντέλο να μπορεί να ταξινομεί και εικόνες που ανήκουν στο \mathcal{X}_u μετά την εκπαίδευση του.

¹Σε αυτό το κείμενο οι όροι ταξινόμηση και κατηγοριοποίηση θα χρησιμοποιούνται εναλλακτικά



Σχήμα 1.1: Το προτεινόμενο μοντέλο δεν έχει δει πότε εικόνα αλόγου. Ωστόσο έχει δει εικόνες από ζέβρες και από την περιγραφή (διάνυσμα) της κατηγορίας "ζέβρα" και "άλογο" θα μπορεί να διακρίνει πως ένα άλογο μοιάζει με μια ζέβρα αλλά δεν θα έχει ρίγες.

Στο παραπάνω πρόβλημα υποτίθεται η γνώση πως οι εικόνες του συνόλου \mathcal{X}_u ανήκουν στο \mathcal{Y}_u , και αναφέρεται από εδώ και πέρα ως ZSL (από το Zero-Shot Learning). Όμως υπάρχει και η γενίκευση του, το GZSL (από το Generalized Zero-Shot Learning), στην οποία τόσο για τις εικόνες του \mathcal{X}_s όσο και για τις εικόνες του \mathcal{X}_u θα θεωρείται πως ανήκουν στο \mathcal{Y} . Το τελευταίο σενάριο είναι πιο ρεαλιστικό, αλλά κι πιο δύσκολο καθώς ο ταξινομητής έχει περισσότερες επιλογές ανάθεσης για κάθε εικόνα που δέχεται ως είσοδο. Αυτή η εργασία επικεντρώνεται κυρίως σε αυτήν την προσέγγιση, καθώς στην πράξη δεν γίνεται να γνωρίζει κάποιος εκ των προτέρων, μόλις δει μια εικόνα, σε ποιο από τα 2 υποσύνολα του \mathcal{Y} ανήκει.

Επιπλέον, δίνοντας κάποια δείγματα εικόνων από το σύνολο \mathcal{X}_u κατά την διάρκεια της εκπαίδευσης δημιουργούνται αντίστοιχα τα προβλήματα FSL (Few-Shot Learning) και GFSL (Generalized Few-Shot Learning). Τέλος υπάρχει μια ακόμα παραλλαγή του προβλήματος, με την οποία δεν ασχολείται η παρούσα διπλωματική, αλλά αναφέρεται για λόγους πληρότητας, το Transductive ZSL/GZSL. Σε αυτήν την παραλλαγή κατά την διάρκεια της εκπαίδευσης το μοντέλο έχει πρόσθαση τόσο στις εικόνες του συνόλου \mathcal{X}_u όσο και στις περιγραφές του συνόλου \mathcal{Y}_u , χωρίς ωστόσο να έχει την γνώση για το ποια είναι η περιγραφή / κατηγορία της κάθε εικόνας $x \in \mathcal{X}_u$.

1.2 Σπουδαιότητα

Αν και το πρόβλημα του GZSL δεν είναι τόσο ώριμο όσο το πρόβλημα της απλής ταξινόμησης εικόνων, μπορεί να προσφέρει αρκετά παραπάνω στην κοινότητα της μηχανικής μάθησης.

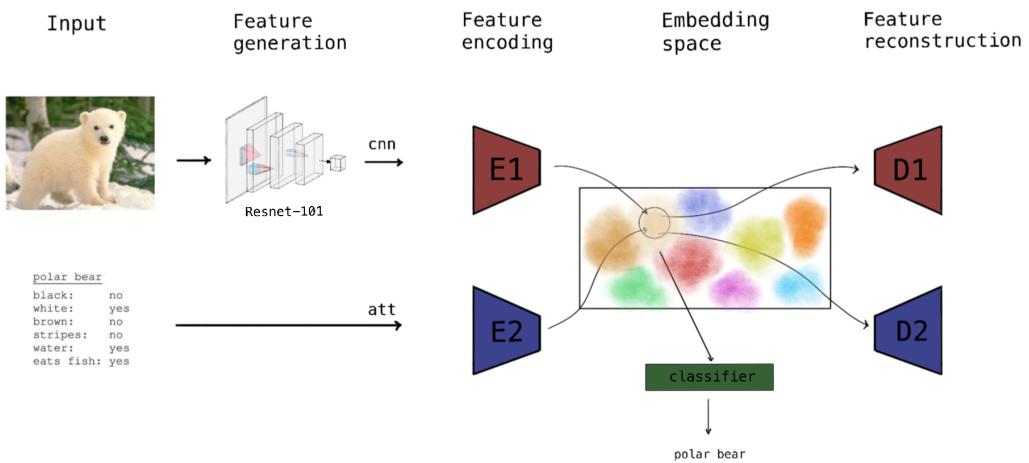
Μας δίνει την δυνατότητα να αντλήσουμε πληροφορία από μια οντότητα (π.χ. κείμενο) και να την προσαρμόσουμε σε μια άλλη (π.χ. εικόνα). Μπορεί να υπάρχει ξεχωριστή έρευνα προς αυτήν την κατεύθυνση², ωστόσο είναι άρρηκτα δεμένη με την φύση του GZSL, και η εργασία αυτή αποτελεί ένα παράδειγμα. Η δυνατότητα να χειριζόμαστε 2 και παραπάνω οντότητες μαζί είναι ένα απαραίτητο βήμα προς την επίτευξη ανώτερων στόχων της τεχνητής νοημοσύνης, καθώς και ο άνθρωπος βασίζεται σε 5 διαφορετικές οντότητες εισόδου, τις

αισθήσεις του.

Επίσης, ίσως με μια μικρή παραλλαγή του προβλήματος να μπορούσαμε να προσθέτουμε συνεχώς καινούριες κατηγορίες στο σύστημα μας, ώστε να συνεχίσει να μαθαίνει. Αυτή είναι άλλη μια δυνατότητα που μας φέρνει ένα βήμα πιο κοντά σε μια μορφή ευφυΐας παρόμοια με αυτή του ανθρώπου.

1.3 Προτεινόμενη αντιμετώπιση

Το προτεινόμενο σύστημα αρχικά θα πρέπει να μπορεί να φέρει δυο πολύ διαφορετικές οντότητες σε παρόμοια μορφή, την εικόνα και την κατηγορία της. Όπως ήδη αναφέρθηκε, συνήθως δίνεται η περιγραφή της κατηγορίας με την μορφή ενός διανύσματος $a_i \in \mathbb{R}^n$, ωστόσο κάποιες φορές αυτό δεν είναι εφικτό και θα χρειαστεί να χρησιμοποιηθούν διανυσματικές αναπαραστάσεις λέξεων, όπως το word2vec [21]³. Αντίστοιχα πρέπει να εξαχθούν κάποια χαρακτηριστικά από τις εικόνες, για να έρθουν στην ίδια διανυσματική μορφή με τις περιγραφές. Γι' αυτό χρησιμοποιείται το τελευταίο επίπεδο του νευρωνικού δικτύου ResNet-101 [12] που, αν και θα μπορούσε να χρησιμοποιηθεί οποιαδήποτε παρόμοια αρχιτεκτονική. Πλέον το μοντέλο θα χρησιμοποιεί τα διανύσματα $\phi(x)$ (cnn χαρακτηριστικά) και $\pi(y)$ (att χαρακτηριστικά) που αποτελούν χαρακτηριστικά της εικόνας και της κατηγορίας αντίστοιχα.



Σχήμα 1.2: Σχηματική επεξήγηση της λειτουργίας του μοντέλου.

Αυτά θα πρέπει να κωδικοποιηθούν σε έναν κοινό χώρο, έτσι ώστε κάθε εικόνα να εμπίπτει στην περιοχή που κωδικοποιείται η κατηγορία της, αλλά διαφορετικές κατηγορίες να εμπίπτουν σε διαφορετικές περιοχές. Επιπλέον θέλουμε κάθε εικόνα να έχει μια μικρή απόκλιση σε σχέση με την κατηγορία της, η οποία θα κωδικοποιεί τα ιδιαίτερα χαρακτηριστικά της εικόνας σε σχέση με μια τυπική εικόνα της κατηγορίας της. Το σύστημα περιγράφεται σχηματικά στο Σχήμα 1.2

²Η περιοχή ονομάζεται Multi-modal Learning

³Δηλαδή θα χρησιμοποιείται ως περιγραφή a_i το $word2vec(y_i)$

Για την κωδικοποίηση χρησιμοποιούνται δύο Variational Auto-Encoders [14] οι οποίοι πρέπει να εκπαιδευτούν με κάποιον τρόπο ώστε να εξυπηρετούν την παραπάνω απαίτηση. Στο πρόσφατα δημοσιευμένο μοντέλο CADA-VAE [27] χρησιμοποιείται ο κλασσικός κωδικοποιητής-αποκωδικοποιητής που περιγράφεται στην δουλειά των Kingma και Welling με κάποιες εμπειρικές τροποποιήσεις στην συνάρτηση σφάλματος, αλλά εδώ χρησιμοποιήθηκε ένα μοντέλο που μοιάζει με το VaDE [13] καθώς μοντελοποιεί καλύτερα την απαίτηση του συγχρονισμού.

Τέλος, εκπαιδεύεται ένας απλός γραμμικός ταξινομητής ώστε να ταξινομεί τις κωδικοποιημένες περιοχές του χώρου κοινής ενσωμάτωσης (embedding space). Έτσι άμα το μοντέλο κωδικοποιητή-αποκωδικοποιητή πετύχει καλό συγχρονισμό θα μπορεί να κατηγοριοποιεί ακόμα και εικόνες από το σύνολο \mathcal{X}_u .

1.4 Συνεισφορά της εργασίας

Η συνεισφορά της παρούσας διπλωματικής είναι τριπλή:

- Αρχικά προσφέρει μια σύγχρονη βιβλιογραφική επισκόπηση της τελευταίας δεκαετίας, ομαδοποιώντας παρόμοιες προσεγγίσεις και τονίζοντας τα κοινά τους χαρακτηριστικά.
- Προτείνει ένα καινούριο μοντέλο για την προσέγγιση των προβλημάτων GZSL και GFSL, που είναι συγκρίσιμο με τις πλέον σύγχρονες μεθόδους.
- Παρουσιάζει μια ποιοτική ανάλυση της προτεινόμενης μεθοδολογίας, προσφέροντας χρήσιμες μελλοντικές κατευθύνσεις για την βελτίωση της.

1.5 Δομή της εργασίας

Ο κύριος κορμός της διπλωματικής χωρίζεται στα επόμενα 4 κεφάλαια:

- Στο **κεφάλαιο 2** δίνεται μια σύντομη βιβλιογραφική επισκόπησή για το πρόβλημα του GZSL. Αρχικά περιγράφονται τα μοντέλα που έχουν προταθεί μέχρι και σήμερα, έπειτα αναλύονται τα σύνολα δεδομένων που αφορούν την παρούσα διπλωματική και τέλος αντιπαρατίθενται τα αποτελέσματα των μοντέλων στα σύνολα δεδομένων.
- Στο **κεφάλαιο 3** γίνεται λεπτομερής περιγραφή του προτεινόμενου συστήματος. Αναλύονται τα επιμέρους τμήματά του και οι υπερπαραμέτροι από τις οποίες εξαρτώνται. Περιγράφεται η συνάρτηση σφάλματος καθώς και η μεθοδολογία εκπαίδευσης.
- Στο **κεφάλαιο 4** εκτείνεται το πειραματικό σκέλος. Αυτό είναι τόσο χρήσιμο για την εκτίμηση των υπερπαραμέτρων όσο και για την αξιολόγηση του μοντέλου μας. Παρατίθενται αναλυτικά αποτελέσματα, καθώς και γίνεται επιθεώρηση του χώρου κοινής ενσωμάτωσης.
- Στο **κεφάλαιο 5** είναι συγκεντρωμένα τα συμπεράσματα της εργασίας και προτείνονται μελλοντικές επεκτάσεις.

2. Βιβλιογραφική επισκόπηση

2.1 Μοντέλα

Σε αυτήν την ενότητα θα παρουσιαστούν τα πιο γνωστά και αξιοσημείωτα μοντέλα που προτάθηκαν την τελευταία δεκαετία για τα προβλήματα ZSL και GZSL. Όλα χρησιμοποιούν ως είσοδο χαρακτηριστικά $\phi(x) \in \mathbb{R}^N$ της εικόνας x και χαρακτηριστικά (ή αλλιώς περιγραφές) $\pi(y) \in \mathbb{R}^M$ των κατηγοριών y . Ο τρόπος εξαγωγής των χαρακτηριστικών $\phi(x)$ είναι διαφορετικός από μοντέλο σε μοντέλο, και εξαρτάται σημαντικά από την χρονολογία της δημοσίευσης. Για τις κατηγορίες y μπορούν να χρησιμοποιηθούν είτε περιγραφές που έχουν δημιουργεί από ανθρώπους είτε γλωσσικά μοντέλα όπως το word2vec¹. Σε κάθε περίπτωση τα μοντέλα εξαγωγής χαρακτηριστικών ϕ και π εκπαιδεύονται σε άλλους τομείς² και χρησιμοποιούνται προ-εκπαιδευμένα, αν και μπορεί να γίνει ταυτόχρονα fine-tuning πάνω στο πρόβλημα του ZSL.

2.1.1 Περιγραφική ταξινόμηση

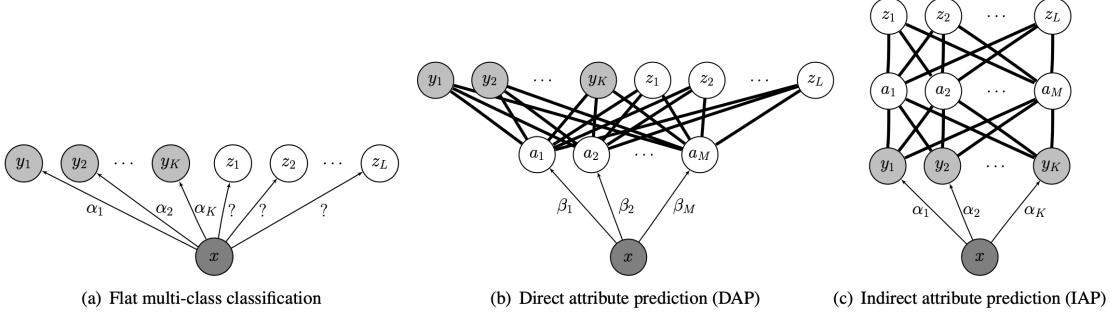
Η πρώτη απόπειρα λύσης του προβλήματος ήταν η περιγραφική ταξινόμηση που περιγράφεται στην δουλεία του Lampert και άλλων [17] το 2009, με τα μοντέλα DAP και IAP.

'Οπως φαίνεται στα αριστερά του Σχήματος 2.1, ένας απλός ταξινομητής αποτελείται από έναν νευρώνα με είσοδο μια κάποια χαρακτηριστικά $\phi(x)$ της εικόνας x και έξοδο τις κατηγορίες y_i ³. Αυτό όμως δε μας δίνει την δυνατότητα να αξιοποιήσουμε την γνώση που μας δίνουν οι περιγραφές $\pi(y_i)$ για τις κατηγορίες y_i , κι συνεπώς είναι αδύνατο να κατηγοριοποιηθούν εικόνες για τις οποίες δεν έχουμε καθόλου παραδείγματα.

¹Σε αυτήν την περίπτωση $a_i = \pi(y_i) = \text{word2vec}(y)$

²Σε τομείς εικόνας και κειμένου αντίστοιχα. Για την εκπαίδευση του ϕ χρησιμοποιείται πολύ συχνά το ImageNet ενώ του π μεγάλες πηγές κειμένων όπως η Wikipedia.

³η έξοδος είναι στον νευρώνα y_i είναι η πιθανότητα $p(y_i|x)$



Σχήμα 2.1: Τα μοντέλα DAP και IAP σε σύγκριση με έναν απλό ταξινομητή εικόνας.

Το **DAP** (Direct Attribute Prediction) είναι μια παραλλαγή του απλού ταξινομητή εικόνας. Αντί να εξάγει την κατηγορία y από τα χαρακτηριστικά μιας δοσμένης εικόνας x , μαθαίνει την περιγραφή $\pi(y) = \{\pi_1, \pi_2, \dots, \pi_M\}$ της εικόνας. Έπειτα θα χρησιμοποιήσει αυτήν την περιγραφή για να κατηγοριοποιήσει την εικόνα x στο σύνολο $\mathcal{Y}_s \cup \mathcal{Y}_u$.

Το **IAP** (Indirect Attribute Prediction) χρησιμοποιεί ένα επίπεδο για να ταξινομήσει την εικόνα στις γνωστές κατηγορίες, ένα ακόμα επίπεδο για να βρει την περιγραφή $\pi(y) = \{\pi_1, \pi_2, \dots, \pi_M\}$ με βάση την ταξινόμηση στα $y_1, y_2, \dots, y_{|\mathcal{Y}_u|}$ και τέλος ένα επίπεδο για την πρόβλεψη των $y_i \in \mathcal{Y}_u$. Οι συγγραφείς υποστηρίζουν πως αυτός ο ανορθόδοξος τρόπος βοηθάει στην άνιση αντιμετώπιση των γνωστών και άγνωστων κατηγοριών \mathcal{Y}_s και \mathcal{Y}_u .

Αν και τα παραπάνω μοντέλα δεν έφεραν καλά αποτελέσματα, έδωσαν την ένδειξη ότι είναι όντως δυνατό να χρησιμοποιηθεί παράπλευρη γνώση για να κατηγοριοποιηθούν εικόνες από άγνωστες κατηγορίες.

2.1.2 Συναρτήσεις συμβατότητας

Μετά από κάποια χρόνια, την περίοδο 2013-2016, πολλοί ερευνητές δοκίμασαν να εισάγουν μια συνάρτηση συμβατότητας $F(x, y; \theta)$, όπου θ είναι εκπαιδεύσιμες παράμετροι. Το μοντέλο εκτιμάει τις παραμέτρους θ από το σύνολο δεδομένων εκπαίδευσης $(\mathcal{X}_s, \mathcal{Y}_s)$ και στην συνέχεια αντιστοιχεί την εικόνα x στην κατηγορία $\hat{y} = \arg \max_{y \in \mathcal{Y}} F(x, y; \theta)$.

Στην βιβλιογραφία δοκιμάστηκαν τόσο γραμμικές όσο και μη-γραμμικές συναρτήσεις.

Γραμμικές Συναρτήσεις συμβατότητας

Οι γραμμικές συναρτήσεις συμβατότητας έχουν την μορφή:

$$F(x, y; W) = \phi(x)^T W \pi(y)$$

Στόχος είναι η εκμάθηση του πίνακα W , και οι διαφορές μεταξύ των μοντέλων που αναφέρονται παρακάτω είναι κυρίως στον τρόπο εκπαίδευσης.

Παρακάτω χρησιμοποιούνται οι ακόλουθοι μαθηματικοί συμβολισμοί:

- $[a]_+ \triangleq \max\{a, 0\}$
- $\Delta(a, b) = 1$ αν $a = b$, αλλιώς 0
- $\mathbb{1}(a) = 1$ αν a αληθής αλλιώς 0

Όπως φαίνεται στην συνέχεια, κάθε συνάρτηση σφάλματος εμπεριέχει την ποσότητα:

$$\Delta(y_k, y) + F(x_k, y; W) - F(x_k, y_k; W)$$

Η ποσότητα αυτή είναι μηδενική για $y = y_k$ ενώ μεγαλώνει όσο μεγαλύτερη γίνεται η διαφορά $F(x_k, y; W) - F(x_k, y_k; W)$ για $y \neq y_k$. Στην περίπτωση που $F(x_k, y; W) > F(x_k, y_k, W)$ η παράμετρος W του μοντέλου, και κατά συνέπεια η συνάρτηση συμβατότητας F , υποδεικνύει πως η εικόνα x_k είναι πιθανότερο να αντιστοιχεί στην κατηγορία $y \neq y_k$ παρά στην πραγματική της κατηγορία y_k . Είναι επιθυμητό λοιπόν για $y \neq y_k$ η συνάρτηση συμβατότητας να έχει μικρότερη τιμή (κατά μια ποσότητα $\Delta(y, y_k)$ από την ποσότητα $F(x_k, y_k; W)$, και αν δεν συμβαίνει το σφάλμα του μοντέλου μεγαλώνει.

Το **DeViSE** [8] ελαχιστοποιεί την ακόλουθη συνάρτηση σφάλματος:

$$L(W) = \sum_{k=1}^K \sum_{y \in \mathcal{Y}^{tr}} [\Delta(y_k, y) + F(x_k, y; W) - F(x_k, y_k; W)]_+$$

η οποία είναι κυρτή και βελτιστοποιείται με SGD. Για την συμπίεση της αναπαράστασης των εικόνων $\phi(x)$ χρησιμοποιείται το τελευταίο επίπεδο AlexNet ενώ για τα διανύσματα των κατηγοριών εκπαιδεύεται ένα μοντέλο word2vec. Έτσι $\phi(x) \in \mathbb{R}^{4096}$ ενώ για τις περιγραφές $\pi(y)$ βρέθηκε πως η καλύτερη επίδοση επιτυγχάνεται για διανύσματα 500 ή 1000 διαστάσεων.

Το **ALE** [1] ελαχιστοποιεί την ίδια συνάρτηση σφάλματος, αλλά σταθμισμένη:

$$L(W) = \sum_{k=1}^K \sum_{y \in \mathcal{Y}^{tr}} \frac{l_{r_\Delta(x_k, y_k)}}{r_\Delta(x_k, y_k)} [\Delta(y_k, y) + F(x_k, y; W) - F(x_k, y_k; W)]_+$$

όπου για τα σταθμά ισχύει $r_\Delta(x_k, y_k) = \sum_{y \in \mathcal{Y}^{tr}} \mathbb{1}[F(x_k, y; W) + \Delta(y_k, y) \geq F(x_k, y_k; W)]$ και $l_k = \sum_{i=1}^k 1/k$. Αυτήν την φορά οι αναπαραστάσεις ϕ και θ βρίσκονται με ένα ευριστικό τρόπο, η περιγραφή του οποίου ξεφεύγει από τον σκοπό της παρούσας διπλωματικής.

Το **SJE** [2] δίνει βαρύτητα μόνο στην κατηγορία y με το μεγαλύτερο σφάλμα:

$$L(W) = \sum_{k=1}^K [\max_{y \in \mathcal{Y}^{tr}} \Delta(y_k, y) + F(x_k, y; W) - F(x_k, y_k; W)]_+$$

Αυτή την φορά δοκιμάζονται ως χαρακτηριστικά εικόνων $\phi(x)$ τα χαρακτηριστικά που χρησιμοποιήθηκαν στα DeViSE και ALE, αλλά και τα CNN χαρακτηριστικά του GoogLeNet [31], το οποίο έχει τις καλύτερες επιδόσεις. Ως χαρακτηριστικά των κατηγοριών χρησιμοποιούνται τόσο χαρακτηριστικά που έφτιαξαν ανθρώπινες ομάδες όσο και τα μοντέλα word2vec και glove [25], αλλά και συνδυασμοί των τριών.

Όλα τα παραπάνω μοντέλα κάνουν μια πολύ απλή μοντελοποίηση του προβλήματος. Οι παράμετροι των μοντέλων ϕ και π είναι τάξεις μεγέθους παραπάνω από το μέγεθος της παραμέτρου W . Ωστόσο γίνεται αντιληπτό πως με κατάλληλα μοντέλα εξαγωγής χαρακτηριστικών γίνεται καλύτερη εκμετάλλευση πληροφοριών για κατηγορίες που λείπουν κατά την εκπαίδευση.

Μη-Γραμμικές Συναρτήσεις συμβατότητας

Το μοντέλα σε αυτήν την ενότητα προσπαθούν να γενικεύσουν λίγο την μοντελοποίηση του προβλήματος εισάγοντας μη γραμμικότητες και μεγαλύτερο αριθμό παραμέτρων. Αναφέρονται δύο χαρακτηριστικά μοντέλα.

Το **LATEM** [35] μοντελοποιεί την συνάρτηση συμβατότητας ως:

$$F(x, y; W_i) = \max_{1 \leq i \leq K} \phi(x)^T W_i \pi(y)$$

όπου κάθε πίνακας W_i προσπαθεί να αιχμαλωτίσει ένα διαφορετικό οπτικό χαρακτηριστικό της εικόνας. Η τιμή K είναι υπερπαράμετρος του μοντέλου που ρυθμίζεται πάνω στο σύνολο επικύρωσης (validation set). Για την εκπαίδευση του μοντέλου χρησιμοποιείται πάλι η συνάρτηση

$$L(W) = \sum_{k=1}^K \sum_{y \in \mathcal{Y}^{tr}} [\Delta(y_k, y) + F(x_k, y; W) - F(x_k, y_k; W)]_+$$

Αν και η συνάρτηση σφάλματος αυτήν την φορά δεν είναι κυρτή ως προς όλα τα W_i η ελαχιστοποίηση της γίνεται ξανά με SGD, χωρίς να παρουσιάζονται προβλήματα.

Το **CMT** [29], αν και παλαιότερο απ' όλα τα μοντέλα που παρουσιάζονται στην ενότητα, χρησιμοποιεί μια διαφορετική προσέγγιση με ένα απλό νευρωνικό δίκτυο. Το νευρωνικό δύο επιπέδων εκπαιδεύεται ώστε να αντιστοιχεί τα χαρακτηριστικά μιας εικόνας $\phi(x)$ σε μια κλάση $\pi(y)$ ελαχιστοποιώντας το ακόλουθο σφάλμα:

$$L(\Theta) = \sum_{y \in \mathcal{Y}^{tr}} \sum_{x \in \mathcal{X}_y} \|\pi(y) - \theta_2 f(\theta_1 \phi(x))\|^2$$

Για την κατηγοριοποίηση χρησιμοποιείται η κατηγορία \hat{y} που η περιγραφή της $\pi(\hat{y})$ είναι πιο κοντά στην περιγραφή $\pi(y)$ που δίνει ως έξοδο το μοντέλο. Επίσης χρησιμοποιούνται 2 μηχανισμοί για να τον διαχωρισμό εικόνων από γνωστές και άγνωστες κατηγορίες, ώστε να γίνεται πιο εύκολα η κατηγοριοποίηση τους στην γενικότερη περίπτωση του GZSL. Τα αποτελέσματα δείχνουν πως αυτοί οι μηχανισμοί βελτιώνουν σημαντικά την επίδοση του μοντέλου.

2.1.3 Υβριδικά μοντέλα

Το 2014-2017 προτάθηκαν κάποια υβριδικά μοντέλα που αντλούν το όνομα τους από το γεγονός ότι προσπαθούν να λύσουν το πρόβλημα μοντελοποιώντας ενδιάμεσες αναπαραστάσεις κλάσεων.

Το **ConSE** [23] εκπαιδεύει έναν απλό ταξινομητή για να υπολογίσει την πιθανότητα $p_s(y|\phi(x))$, όπου $y \in \mathcal{Y}_s$ για κάθε εικόνα στο σύνολο εκπαίδευσης. Έπειτα δεδομένης μια εικόνας x στο σύνολο ελέγχου, υπολογίζει τις T πιο πιθανές γνωστές κλάσεις που ανήκει, έστω s_1, s_2, \dots, s_T . Έτσι προσεγγίζει την περιγραφή της εικόνας x ως:

$$\hat{\pi}(x) = \frac{1}{Z} \sum_{t=1}^T p_s(s_t|x) \cdot \pi(s_t)$$

όπου $Z = \sum_{t=1}^T p(s_t|x)$ και T υπερπαράμετρος του μοντέλου. Τέλος το πρόβλημα του ZSL λύνεται υπολογίζοντας την ομοιότητα συνημιτόνου (cosine similarity) μεταξύ των άγνωστων

περιγραφής και της εκτιμώμενης περιγραφών, και διαλέγοντας την πιο όμοια κλάση:

$$\hat{y} = \arg \max_{y \in \mathcal{Y}_u} \cos(\pi(y), \pi(\hat{x}))$$

To **SYNC** [5] εισάγει R συνθετικές κλάσεις z_1, z_2, \dots, z_R που δεν έχουν φυσικές αναπαραστάσεις. Υπενθυμίζουμε πως οι περιγραφές των πραγματικών κατηγοριών είναι $a_1, a_2, \dots, a_{|\mathcal{Y}|}$ και θα συμβολίζουμε με b_1, b_2, \dots, b_R τις περιγραφές των συνθετικών κατηγοριών. Οι πραγματικές και οι συνθετικές κατηγορίες σχηματίζουν ένα διμερή γράφο με βάρη

$$s_{yz} = \frac{\exp(-d(a_y, b_z))}{\sum_{z=1}^R \exp(-d(a_y, b_z))}$$

όπου η συσχέτιση d μεταξύ κλάσεων ορίζεται ως $d(a_y, b_z) = (a_y - b_z)^T \Sigma^{-1} (a_y - b_z)$ και η παράμετρος Σ υπολογίζεται από τα δεδομένα.

Υποθέτουμε πως κάθε κλάση (πραγματική και συνθετική) χρησιμοποιεί έναν απλό ταξινομητή (w_y ή u_z) προκειμένου να ταξινομεί την εικόνα x ως $\hat{y} = \arg \max_{y \in \mathcal{Y}} w_y^T \phi(x)$ ή $\hat{z} = \arg \max_{z \in \mathcal{Z}} u_z^T \phi(x)$. Μετά, με σκοπό την ελαχιστοποίηση του σφάλματος παραμόρφωσης $\|w_y - \sum_{z=1}^R s_{yz} u_z\|_2^2$ οι ταξινομητές w_y γίνονται $w_y = \sum_{z=1}^R s_{yz} u_z \forall y \in \mathcal{Y}$.

Υπό αυτήν την συνθήκη ελαχιστοποιείται το σφάλμα

$$\mathcal{L}(u_1, \dots, u_R) = \sum_{k=1}^K \sum_{y \in \mathcal{Y}_s} [1 \pm w_y^T \phi(x_n)]_+^2 + \frac{\lambda}{2} \sum_{y \in \mathcal{Y}_s} \|w_y\|_2^2$$

όπου το πρόσημο είναι $+$ αν $y = y_n$ και $-$ διαφορετικά. Έτσι μαθαίνονται οι ταξινομητές u_z από το σύνολο \mathcal{Y}_s και έπειτα μέσω της σχέσης $w_c = \sum_{z=1}^R s_{cz} u_z$ υπολογίζονται οι ταξινομητές w_c για όλες τις πραγματικές κλάσεις (γνωστές και άγνωστες).

To **GFZSL** [33] εισάγει δύο συναρτήσεις f_μ και f_σ , που μοντελοποιούνται με νευρωνικά δίκτυα, και θεωρεί πως η πιθανότητα $f(x|y)$ ακολουθεί την κανονική κατανομή $N(f_\mu(a), f_\sigma(a))$. Οι συναρτήσεις f_μ και f_σ υπολογίζονται με εκπαίδευση στο σύνολο $(\mathcal{X}_s, \mathcal{Y}_s)$ και στην συνέχεια για κάθε $x \in \mathcal{X}_u$ το μοντέλο διαλέγει την κλάση $\hat{y} = \arg \max_{y \in \mathcal{Y}} f(x|y)$.

2.1.4 Παραγωγικά μοντέλα

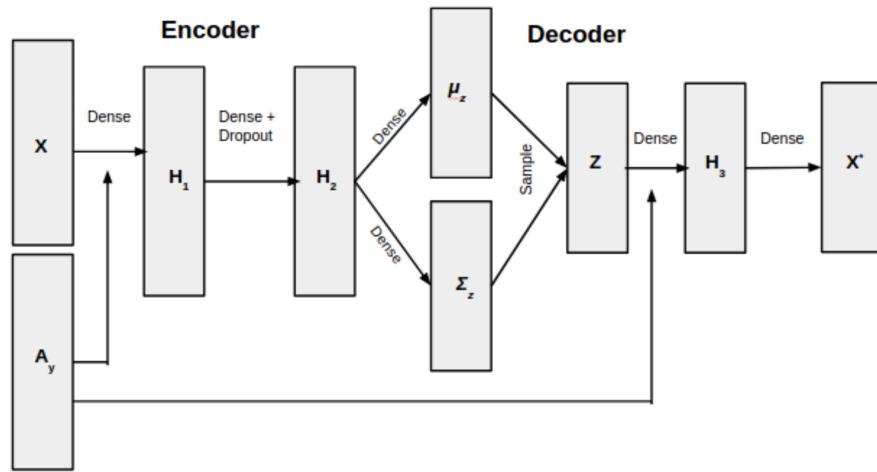
Το 2017 ξεκίνησαν οι πρώτες προσπάθειες επίλυσης του προβλήματος με την χρήση παραγωγικών μοντέλων. Τα παραγωγικά μοντέλα εκπαιδεύονται ώστε να μπορούν να δημιουργούν αντιπροσωπευτικά χαρακτηριστικά $\phi(x)$ δεδομένης μίας κατηγορίας y ή μιας περιγραφής $\pi(y)$. Έτσι μπορούν να αντιμετωπίσουν το πρόβλημα του ZSL δημιουργώντας συνθετικά δεδομένα για τις κατηγορίες που έχουν λίγα ή καθόλου παραδείγματα. Έπειτα εκπαιδεύεται ένας απλός ταξινομητής $f : \phi(\mathcal{X}) \rightarrow \mathcal{Y}$ για την ταξινόμηση των εικόνων του συνόλου δεδομένων ελέγχου. Παρακάτω αναφέρονται τρία χαρακτηριστικά μοντέλα αυτής της κατηγορίας.

To **CVAE-ZSL** [22] χρησιμοποιεί έναν CVAE [30] για να δημιουργήσει δείγματα εικόνων $\phi(x)$ βάσει των περιγραφών τους $\pi(y)$. Το μοντέλο περιγράφεται στο Σχήμα 2.6 και αποτελείται από έναν δεσμευμένο κωδικοποιητή και έναν δεσμευμένο αποκωδικοποιητή. Ο κωδικοποιητής δέχεται σαν όρισμα τα χαρακτηριστικά της εικόνας και της κατηγορίας $\phi(x)$ και $\pi(y)$ και δίνει ως έξοδο ένα διάνυσμα μ και έναν πίνακα Σ που χρησιμοποιούνται ώστε να δειγματοληπτηθεί μια αναπαράσταση $z \sim \mathcal{N}(\mu, \Sigma)$. Η αναλυτική περιγραφής της λειτουργίας

των VAE και των παραλλαγών τους ξεφεύγει από τους σκοπούς αυτής της ενότητας αλλά περιγράφεται στο Παράρτημα ;;.

Με αυτόν τον τρόπο η πληροφορία του χαρακτηριστικού $\phi(x)$ διαμοιράζεται στην περιγραφή $\pi(y)$ (που κωδικοποιεί τα γενικά χαρακτηριστικά της κατηγορίας) και στην εσωτερική αναπαράσταση z (που κωδικοποιεί τα ιδιαίτερα χαρακτηριστικά της εικόνας x μέσα στην κατηγορία y).

Έπειτα η εσωτερική αναπαράσταση z και το χαρακτηριστικό $\pi(y)$ εισάγονται στον δεσμευμένο αποκωδικοποιητή για την ανακατασκευή του $\phi(x)$. Το μοντέλο δηλαδή αρχικά εκπαιδεύεται στην ανακατασκευή της χαρακτηριστικών $\phi(x)$ και όχι στο πρόβλημα του ZSL. Όταν το μοντέλο του κωδικοποιητή-αποκωδικοποιητή εκπαιδευτεί, ο κωδικοποιητής θα είναι σε θέση να δημιουργεί αντιπροσωπευτικά δείγματα για κάθε κατηγορία δεδομένης μιας εσωτερικής αναπαράστασης $z \sim \mathcal{N}(0, I)$ και μιας περιγραφής $\pi(y)$.



Σχήμα 2.2: Η αρχιτεκτονική του μοντέλου CVAE-ZSL

Βρίσκοντας 100 τυχαίες τιμές $z_i \sim \mathcal{N}(0, I)$ και δεδομένης της περιγραφής μιας κατηγορίας y το CVAE-ZSL μπορεί να δημιουργήσει 100 δείγματα $\phi(x_i)$ για αυτήν την κατηγορία. Έπειτα μπορεί να χρησιμοποιηθεί ένας απλός γραμμικός ταξινομητής ή ένα SVM για την ταξινόμηση των εικόνων με βάση τα χαρακτηριστικά τους ϕ .

Η συνάρτηση σφάλματος του μοντέλου είναι ίδια με αυτήν ενός VAE:

$$\mathcal{L}(x, y; E, D) = -\text{D}_{\text{KL}}(q_E(z|x, a) || p_D(z|a)) + \mathbb{E}_{q_E(z|a)}[\log p_D(x|z, a)]$$

όπου E, D οι παράμετροι του κωδικοποιητή, αποκωδικοποιητή και q_E, p_D τα πιθανοτικά τους μοντέλα.

Το **SE** [16] στηρίζεται στην ίδια φιλοσοφία αλλά χρησιμοποιεί έναν μηχανισμό ανάδρασης. Η αρχιτεκτονική του μοντέλου φαίνεται στο Σχήμα 2.3. Αυτήν την φορά η περιγραφή $\pi(y)$ χρησιμοποιείται μόνο στον αποκωδικοποιητή, ενώ η εσωτερική αναπαράσταση z έχει τις ίδιες ιδιότητες με το μοντέλο CVAE-ZSL. Ο παλινδρομητής έχει στόχο την ανακατασκευή της περιγραφής $\pi(y)$, και χρησιμεύει τόσο στην χρήση εικόνων χωρίς ετικέτα (transductive setting) όσο και στην βελτίωση την παραγώμενων ανακατασκευών του $\phi(x)$.

Σε αυτό το μοντέλο η συνάρτηση σφάλματος είναι πιο σύνθετη.

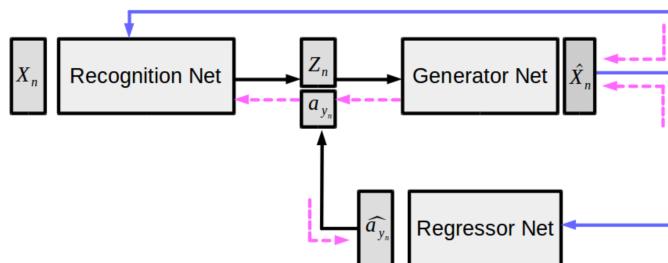
$$\mathcal{L}(x, y; E, D, R) = \mathcal{L}_{VAE} + \lambda_c \cdot \mathcal{L}_c + \lambda_E \cdot \mathcal{L}_E + \lambda_{reg} \cdot \mathcal{L}_{reg}$$

όπου:

- $\mathcal{L}_{VAE} = \text{D}_{\text{KL}}(p_E(z|\phi(x))||p(z)) - \mathbb{E}_{p_E(z|\phi(x)), p(a|\phi(x))}[\log p_D(\phi(x)|z, a)]$
- $\mathcal{L}_c = -\mathbb{E}_{p_D(\hat{x}|z, a)p(z)p(a)}[p_R(a|\hat{x})]$
- $\mathcal{L}_E = -\mathbb{E}_{p_D(\hat{x}|z, a)}[\text{D}_{\text{KL}}(p_E(z|\hat{x})||p(z))]$
- $\mathcal{L}_{reg} = -\mathbb{E}_{p(z)p(a)}[\log p_D(\hat{x}|z, a)]$

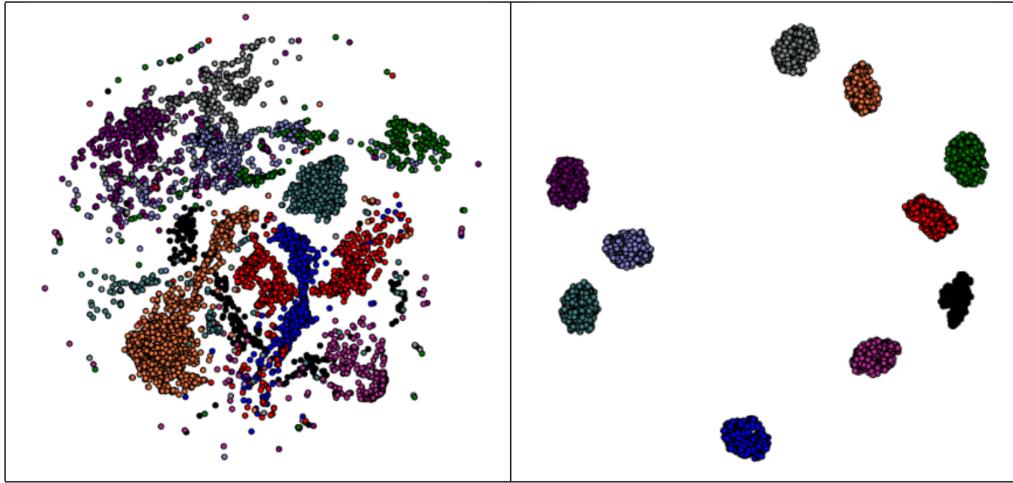
Με E, D, R συμβολίζονται οι παράμετροι του κωδικοποιητή, αποκωδικοποιητή και παλινδρομητή αντίστοιχα ενώ με p_E, p_D και p_R τα πιθανοτικά τους μοντέλα. Οι σταθερές $\lambda_c, \lambda_E, \lambda_R, \lambda_{reg}$ αποτελούν υπερπαραμέτρους του συστήματος.

Η ακριβής περιγραφή των παραπάνω σχέσεων, καθώς και το πως αυτές μπορούν να απλοποιηθούν στα πλαίσια του προβλήματος ξεφεύγει από τους σκοπούς της διπλωματικής. Ωστόσο δίνονται κάποιοι παραλληλισμοί με τους κλασσικούς VAE. Το σφάλμα του κωδικοποιητή-αποκωδικοποιητή \mathcal{L}_{VAE} είναι το κλασσικό σφάλμα ενός VAE / CVAE (προσοχή στ' ότι είναι μία μίξη των δύο καθώς στην κωδικοποιήση δε χρησιμοποιείται η περιγραφή a , αλλά στην αποκωδικοποίηση χρησιμοποιείται). Το σφάλμα του παλινδρομητή \mathcal{L}_c μπορεί να θεωρηθεί ως σφάλμα ανακατασκευής του a δεδομένου του $\phi(x)$, και χρησιμοποιείται για να παραγάγει ο κωδικοποιητής αναπαραστάσεις z που μπορούν να γίνουν αντιληπτές από τον παλινδρομητή. Τέλος τα σφάλματα \mathcal{L}_{reg} και \mathcal{L}_E χρησιμοποιούνται για να εξασφαλίσουν ότι τα συνθετικά δείγματα \hat{x} είναι εξίσου "πιθανά" (δηλαδή εξίσου ποιοτικά) με τα πραγματικά και να κρατήσουν τις αναπαραστάσεις z κοντά στην *a priori* κατανομή τους, $p(z)$.

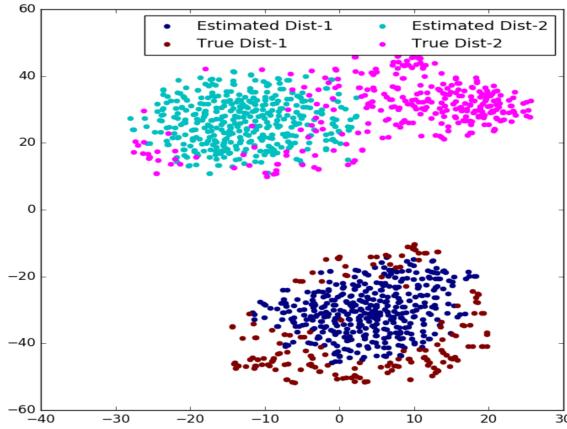


Σχήμα 2.3: Η αρχιτεκτονική του μοντέλου SE. Με κόκκινο φαίνεται η πρόσθια τροφοδότηση και με μπλε η οπισθοδιάδοση.

Παρακάτω φαίνονται τα ανακατασκευασμένα χαρακτηριστικά $\hat{\phi}(x)$ σε σύγκριση με τα αρχικά $\phi(x)$ για τα μοντέλα CVAE-ZSL και SE. Πιο συγκεκριμένα στο Σχήμα 2.4 φαίνεται ο μετασχηματισμός t-SNE εικόνων από 10 άγνωστες κατηγορίες, στα αριστερά των πραγματικών δειγμάτων αυτών των κατηγοριών και στα αριστερά συνθετικών δειγμάτων. Στο Σχήμα 2.5 βλέπουμε πάλι τον μετασχηματισμό t-SNE, αυτήν την φορά 2 άγνωστων κατηγοριών. Τα χαρακτηριστικά των πραγματικών και συνθετικών εικόνων υπερτίθενται. Όπως φαίνεται το SE συνθέτει καλύτερα δείγματα για τις άγνωστες κατηγορίες, γεγονός που αιτιολογείται από την ύπαρξη του μηχανισμού ανάδρασης.



Σχήμα 2.4: Σύγκριση χαρακτηριστικών από συνθετικές και πραγματικές εικόνες για το μοντέλο CVAE-ZSL.



Σχήμα 2.5: Σύγκριση χαρακτηριστικών από συνθετικές και πραγματικές εικόνες για το μοντέλο SE.

To **f-CLSWGAN** [37] χρησιμοποιεί GANs [10] για την παραγωγή συνθετικών χαρακτηριστικών $\hat{\phi}(x)$. Συνοπτικά τα GANs αποτελούνται από έναν παραγωγό (generator), που παράγει συνθετικά δείγματα από μια αυθαίρετη κατανομή, και έναν διευκρινιστή (discriminator), που προσπαθεί να διακρίνει τα πραγματικά δείγματα από τα συνθετικά. Αυτή η συνεχόμενη “διαμάχη” μεταξύ παραγωγού και διευκρινιστή, αναγκάζει τον παραγωγό να παράγει χαρακτηριστικά εικόνων όσο το δυνατόν πιο αληθοφανή. Στην προκειμένη περίπτωση, επειδή θέλουμε να παράγουμε χαρακτηριστικά εικόνων αναλόγως την κατηγορία, τόσο ο παραγωγός όσο και ο διευκρινητής πρέπει να είναι δεσμευμένοι ως προς την κλάση y (Conditional GAN). Η συνάρτηση σφάλματος σε αυτήν την περίπτωση είναι:

$$\mathcal{L}_{GAN}(x, y; G, D) = \mathbb{E}[\log D(\phi(x), a)] + \mathbb{E}[\log(1 - D(\hat{\phi}(x), a))]$$

όπου $\hat{\phi}(x) = G(z, a)$, $z \sim \mathcal{N}(0, I)$ είναι συθετικά χαρακτηριστικά που δημιουργεί ο παραγωγός, $\phi(x)$ είναι τα πραγματικά χαρακτηριστικά της εικόνας x και G, D οι παράμετροι του παραγωγού και του διευκρινιστή αντίστοιχα. Αντικειμενικός στοίχος της εκπαίδευσης είναι η ελαχιστοποίηση του σφάλματος στον παραγωγό και η μεγιστοποίηση του σφάλματος στον

διευκρινιστή:

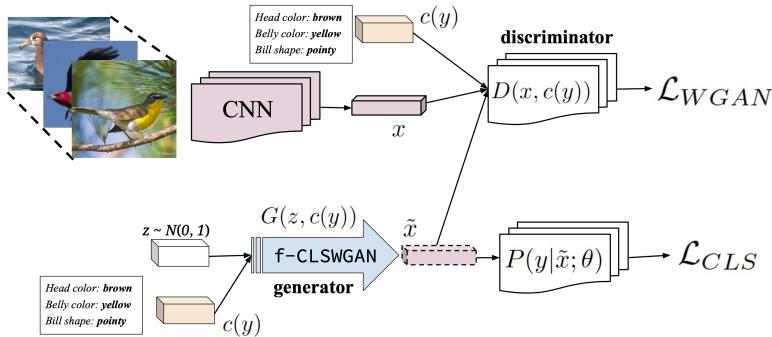
$$\min_G \max_D \mathcal{L}_{GAN}$$

Αξίζει αν τονιστεί πως η εκπαίδευση του GAN γίνεται μόνο στο \mathcal{Y}_s αλλά το μοντέλο μπορεί να παράγει αληθοφανή χαρακτηριστικά για τις κατηγορίες που ανήκουν στο \mathcal{Y}_u . Ως χαρακτηριστικά εικόνων χρησιμοποιούνται τα 2048 CNN χαρακτηριστικά του τελευταίου επιπέδου του ResNet-101.

Λόγω των διάφορων μειονεκτημάτων των GAN όμως, οι συγγραφείς εξετάζουν μια παραλλαγή του WGAN [3] στην θέση του κλασιστικού GAN. Τέλος, επειδή κανένα από τα μοντέλα GAN και WGAN δε εγγυάται ότι τα παραγόμενα χαρακτηριστικά θα είναι ευδιάκριτα μεταξύ τους, οι συγγραφείς ενσωματώνουν το σφάλμα ενός απλού ταξινομητή στο σφάλμα του WGAN προκειμένου τα παραγόμενα χαρακτηριστικά $\phi(x)$ να είναι ευδιάκριτα. Πλέον ο αντικειμενικός στόχος του συστήματος είναι:

$$\min_G \max_D \mathcal{L}_{WGAN} + \beta \mathcal{L}_{CLS}$$

όπου β υπερπαράμετρος του μοντέλου.



Σχήμα 2.6: Η αρχιτεκτονική του μοντέλου f-CLSWGAN

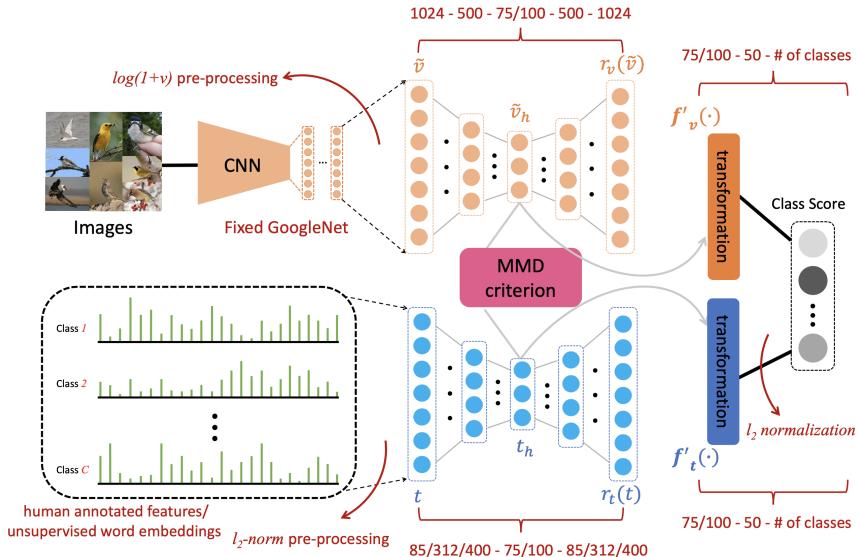
2.1.5 Χώροι κοινής ενσωμάτωσης

Τα μοντέλα που ανήκουν σε αυτήν την κατηγορία προσπαθούν να προβάλουν τα χαρακτηριστικά των εικόνων $\phi(x)$ και των κατηγοριών $\pi(y)$ σε ένα χώρο κοινής ενσωμάτωσης (embedding space). Η κεντρική ιδέα είναι πως εικόνες από την ίδια κατηγορία πρέπει να προβάλλονται στην ίδια περιοχή, στην οποία θα προβάλλεται και η περιγραφή της κατηγορίας, ενώ εικόνες από διαφορετικές κατηγορίες πρέπει να προβάλλονται σε διαφορετικές περιοχές. Έτσι δημιουργείται ένας διαμερισμός του χώρου κοινής ενσωμάτωσης σε $|\mathcal{Y}|$ περιοχές. Επειτα κάποια απλή αρχιτεκτονική, όπως για παράδειγμα ένας απλός ταξινομητής, χρησιμοποιείται για την εκμάθηση αυτή του διαχωρισμού. Εκπαιδεύεται δηλαδή ώστε να ταξινομεί σημεία του χώρου κοινής ενσωμάτωσης στις $|\mathcal{Y}|$ πιθανές κλάσεις. Τέλος, την ώρα της αξιολόγησης του μοντέλου, οι εικόνες (είτε ανήκουν σε γνωστές είτε σε άγνωστες κλάσεις) κωδικοποιούνται στον χώρο κοινής ενσωμάτωσης και στην συνέχεια ταξινομούνται σε μία από τις διαθέσιμες κατηγορίες. Το μυστικό της επιτυχίας αυτών των προσεγγίσεων είναι πως η προθολή στον χώρο κοινής ενσωμάτωσης γίνεται με τέτοιο τρόπο, που δεν συγχρονίζονται μόνο οι περιοχές των γνωστών κατηγοριών που είναι διαθέσιμες κατά την διάρκεια της εκπαίδευσης, αλλά και των άγνωστων κατηγοριών, παρόλο που το μοντέλο δεν τις έχει δει ποτέ.

Το **ReViSE** [32] ήταν το πρώτο μοντέλο που δημοσιεύτηκε σε αυτήν την κατεύθυνση, το 2017. Η αρχιτεκτονική του μοντέλου φαίνεται στο σχήμα 2.7. Τα χαρακτηριστικά των εικόνων εξάγονται από το τελευταίο επίπεδο του GoogLeNet που έχει 1024 μονάδες, και έπειτα υπόκεινται στον μετασχηματισμό $v \rightarrow \log(1 + v)$. Για τις κατηγορίες χρησιμοποιούνται είτε οι περιγραφές που έχουν δημιουργηθεί από τους δημιουργούς των συνόλων δεδομένων είτε τα μοντέλα word2vec και GloVe. Έπειτα 2 auto-encoders κωδικοποιούν και αποκωδικοποιούν αυτά τα χαρακτηριστικά, σε ένα χώρο κοινών διαστάσεων. Τα διανύσματα \tilde{v}_h και t_h είναι οι αναπαραστάσεις στον χώρο αυτό, και ακολουθούν τις επιθυμητές ιδιότητας που αναφέρθηκαν στην εισαγωγή της υποενότητας.

Για να συγχρονιστούν οι περιοχές του χώρου κοινής ενσωμάτωσης μεταξύ των εικόνων και των περιγραφών τους, χρησιμοποιείται η μετρική Maximum Mean Discrepancy (MMD) [11] στην συνάρτηση σφάλματος του μοντέλου. Αν και η αναλυτική περιγραφή της \mathcal{L}_{MMD} ξεφεύγει από τον οκοπό της ενότητας, η ελαχιστοποίηση της ποσότητας αυτής συγχρονίζει τις κατανομές πιθανοτήτων των εικόνων και των κατηγοριών τους στον χώρο κοινής ενσωμάτωσης.

Τέλος αξίζει να σημειωθεί πως για τις εικόνες χρησιμοποιείται μια παραλλαγή του κλασσικού auto-encoder, που περιγράφεται εδώ [26], ενώ για το κείμενο χρησιμοποιείται ένας απλός auto-encoder. Οι συναρτήσεις f'_v και f'_t λειτουργούν ως ταξινομητές του χώρου κοινής ενσωμάτωσης, και αποτελούνται από ένα MLP με ένα κρυφό επίπεδο.



Σχήμα 2.7: Η αρχιτεκτονική του μοντέλου ReViSE

Το **CADA-VAE** [27] χρησιμοποιεί το τελευταίο επίπεδο της αρχιτεκτονικής ResNet-101, με 2048 μονάδες, για την εξαγωγή χαρακτηριστικών εικόνων. Για τις κατηγορίες χρησιμοποιούνται είτε οι ανθρώπινες περιγραφές είτε το μοντέλο word2vec.

Σε αντίθεση με το ReViSE εδώ χρησιμοποιούνται Variational Auto-Encoders για την προβολή των χαρακτηριστικών σε ένα κοινό χώρο, όπως φαίνεται στο σχήμα 2.8. Όπως περιγράφεται και στο Παράρτημα ;; οι VAE κωδικοποιούν τα CNN χαρακτηριστικά $\phi(x)$ σε μια πιθανοτική κατατομή, που εδώ μοντελοποιείται σαν κανονική, άρα εξαρτάται από μια μέση τιμή και έναν πίνακα διασποράς. Έπειτα ένα σημείο του χώρου κοινής ενσωμάτωσης $z \sim \mathcal{N}(\mu, \Sigma)$ δειγ-

ματοληπτείται και αποκωδικοποιείται. Το σφάλμα των 2 VAE θα προστεθεί στην συνολική συνάρτηση σφάλματος του μοντέλου:

$$\begin{aligned}\mathcal{L}_{VAE} &= \mathcal{L}_{VAE_1} + \mathcal{L}_{VAE_2} \\ &= \beta D_{KL}(p_{E_1}(z|\phi(x))||p(z)) - \mathbb{E}_{p_{E_1}(z|\phi(x))}[\log p_{D_1}(\phi(x)|z)] + \\ &\quad \beta D_{KL}(p_{E_2}(z|\pi(y))||p(z)) - \mathbb{E}_{p_{E_2}(z|\pi(y))}[\log p_{D_2}(\pi(y)|z)]\end{aligned}$$

όπου β υπερπαράμετρος του μοντέλου, E_1, D_1 ο κωδικοποιητής και αποκωδικοποιητής των CNN χαρακτηριστικών και E_2, D_2 ο κωδικοποιητής και αποκωδικοποιητής των περιγραφών.

Το προηγούμενο σφάλμα φροντίζει οι εικόνες και οι περιγραφές να αποκωδικοποιούνται σωστά, και να σχηματίζουν περιοχές στον χώρο κοινής ενσωμάτωσης, αλλά ο συγχρονισμός των κατανομών (Distribution Alignment) τους γίνεται ελαχιστοποιώντας την Wasserstein απόσταση [9].

$$\mathcal{L}_{DA} = \sqrt{\|\mu_1 - \mu_2\|_2^2 + \|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_{\text{Frobenius}}^2}$$

Αν και σε αυτό το σημείο το μοντέλο δουλεύει ικανοποιητικά, χρησιμοποιείται επιπρόσθετα η τεχνική της διασταυρωμένου συγχρονισμού (Cross Alignment) όπως στα [18], [28].

$$\mathcal{L}_{CA} = \|\phi(x) - D_1(E_2(\pi(y)))\| + \|\pi(y) - D_2(E_1(\phi(x)))\|$$

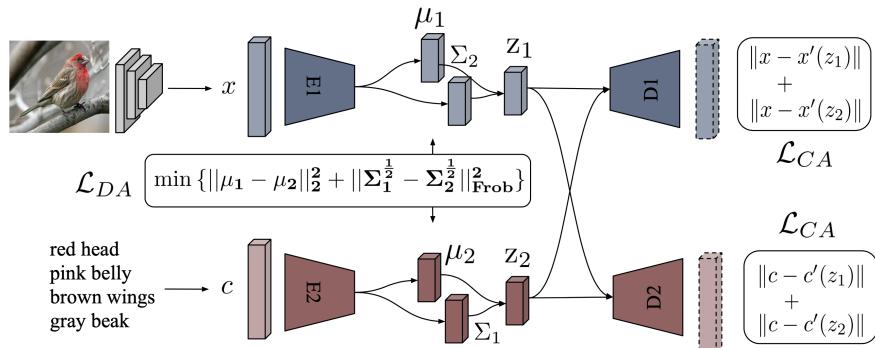
Δηληδή επιβάλλεται στον αποκωδικοποιητή των περιγραφών να μπορεί να αποκωδικοποιήσει εικόνες και στον αποκωδικοποιητή των εικόνων να μπορεί να αποκωδικοποιήσει περιγραφές.

Έτσι η συνολική συνάρτηση σφάλματος του μοντέλου είναι:

$$\mathcal{L}(x, y; E, D) = \mathcal{L}_{VAE} + \gamma \mathcal{L}_{CA} + \delta \mathcal{L}_{DA}$$

και β, γ, δ υπερπαράμετροι του μοντέλου.

Τέλος χρησιμοποιείται ένας απλός ταξινομητής Softmax για την ταξινόμηση του χώρου κοινής ενσωμάτωσης στις $|\mathcal{Y}|$ κατηγορίες.



Σχήμα 2.8: Η αρχιτεκτονική του μοντέλου CADA-VAE

Αξίζει να σημειωθεί πως η συγκεκριμένη ιδεά λειτουργεί ακόμα και με παραπάνω από δύο οντότητες (εικόνα, περιγραφή) ή και με υβριδικές οντότητες, όπως για παράδειγμα το 50% των περιγραφών να προέρχονται από το word2vec και το άλλο 50% από το glove.

2.2 Σύνολα δεδομένων

Υπάρχουν 5 σύνολα δεδομένων που είναι κατάλληλα για την εκπαίδευση και αξιολόγηση του προβλήματος του ZSL / GZSL:

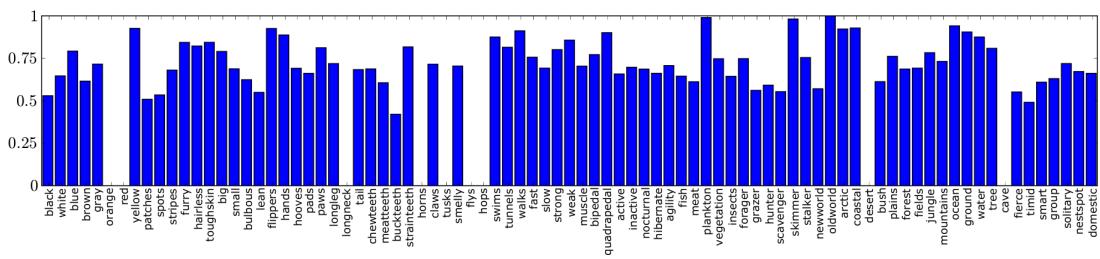
To Attribute Pascal and Yahoo ή αλλιώς **aPY** [7] είναι ένα μικρό σύνολο δεδομένων εκπαιδευσης με μόλις 15K εικόνες και 32 κατηγορίες αντικείμενων. Για κάθε κατηγορία υπάρχει μια περιγραφή με 64 δυαδικές τιμές, που αντιστοιχούν σε μέγεθος, σχήμα, υλικό και εξαρτήματα. Ένα παράδειγμα φαίνεται στο παρακάτω σχήμα.

Το Animals with Attributes 1 ή αλλιώς **AWA1** [17] είναι ενα μεσαίου μεγέθου σύνολο δεδομένων με 30K εικόνες και 50 κατηγορίες ζώων. Για κάθε κατηγορία υπάρχουν 85 πραγματικές τιμές που καθορίζουν την περιγραφή της όπως φαίνεται στο Σχήμα 2.9.

To Animals with Attributes 2 ή αλλιώς **AWA2** [36] δημιουργήθηκε ως προς αντικατάσταση του AWA1 λόγω των προβλημάτων που είχε με τα πνευματικά δικαιώματα των εικόνων. Αντικαθιστά τις εικόνες που δεν επιτρέπεται να χρησιμοποιηθούν λόγω πνευματικών δικαιωμάτων καθώς και προσθέτει μερικές ακόμα.

To Caltech UCSD Birds 200-2011 ή αλλιώς **CUB** [34] είναι κι αυτός μεσαίου μεγέθους σύνολο δεδομένων με 14K εικόνες και 200 κατηγορίες από είδη πουλιών. Οι κατηγορίες περιγράφονται με 312 δυαδικές τιμές που αφορούν κυρίως οπτικά χαρακτηριστικά του πουλιού όπως μέγεθος, χρώμα, σχήμα κτλ.

Τέλος το **SUN** [24] αποτελείται από 12K εικόνες και 717 κατηγορίες τοπίων όπως για παράδειγμα συνοστισμός, φωτιά, κάμπινγκ, γραφείο και καταδύσεις. Κάθε σκηνή περιγράφεται με 102 τιμές που αφορούν διαφορετικά χαρκτηριστικά της.



Σχήμα 2.9: Παράδειγμα περιγραφής μίας κατηγορίας στο σύνολο δεδομένων AWA.

Παρακάτω συγκεντρώνονται και αντιπαρατίθενται τα χαρακτηριστικά των συνόλων δεδομένων. Ο διαχωρισμός των συνόλων δεδομένων σε σύνολα εκπαίδευσης (training set), επικύρωσης (validation set) και ελέγχου (test set) είναι σύμφωνος με την πρόσφατη επισκόπιση [36] του ZSL. Τα σύνολα εκπαίδευσης και επικύρωσης χρησιμοποιούνται για την εκπαίδευση του μοντέλου και την εκτίμηση υπερπαραμέτρων, ενώ το σύνολο ελέγχου για την αξιολόγηση των άγνωστων κλάσεων. 'Όμως θέλουμε να αξιολογήσουμε και την επίδοση του μοντέλου στις γνωστές κλάσεις, και γιάυτό τον σκόπο συλλέγονται δείγματα από τα σύνολα εκπαίδευσης και επικύρωσης.

Σύνολο Δεδομένων	Μέγεθος Περιγραφής	γ	γ^{train}	γ^{val}	γ^{test}	\mathcal{X}	$\mathcal{X}^{\text{train}}$	\mathcal{X}^{val}	$\mathcal{X}^{\text{test}}$
aPY	64	32	15	5	12	15339	5932	1483	7924
AWA1	85	50	27	13	10	30475	19832	4958	5685
AWA2	85	50	27	13	10	37322	23527	5882	7913
CUB	312	200	100	50	50	11788	7057	1764	2967
SUN	102	717	580	65	72	14340	14340	2580	1440

Πίνακας 2.1: Σύγκριση των σύνολων δεδομένων

Επιπλέον, τα τελευταία χρόνα, με την συνεχή βελτίωση των αποτελεσμάτων χρησιμοποιείται συχνά και το ImageNet ως μεγέλο σύνολο δεδομένων. Το ImageNet αποτελείται από 14 εκατομμύρια εικόνες από 21 χιλιάδες κατηγορίες, που είναι ιεραρχικά δομημένες. Αποτελεί φυσική επιλογή για το πρόβλημα, καθώς υπάρχει μεγάλη ανισορροπία δειγμάτων εικόνων μεταξύ των κλάσεων. Η κλάση με τα περισσότερα παραδείγματα αποτελείται από 3047 δείγματα εικόνων την στιγμή που υπάρχουν κλάσεις με μόλις ένα δείγμα. Ένα υποσύνολο 1000 κατηγοριών, με περίπου 1000 δείγματα ανά κατηγορία, χρησιμοποιείται ως σύνολο εκπαίδευσης (200 από τις οποίες είναι σύνολο επικύρωσης), ενώ οι υπόλοιπες 20K κατηγορίες (ή διάφορα υποσύνολα τους όπως π.χ. οι 500 κατηγορίες με τα περισσότερα δείγματα) χρησιμοποιούνται ως σύνολο ελέγχου.

2.3 Σύγκριση μοντέλων

Μέχρι πρότινος δεν υπήρχε κάποιος κοινός τρόπος αξιολόγησης των μοντέλων του ZSL. Αυτό είχε ως αποτέλεσμα διαφορετικές δημοσιεύσεις να μην είναι πάντα συγκρίσιμες. Στο [36] προτείνεται ως μετρική αξιολόγησης η μέση top-1 ακρίβεια ανά κατηγορία, επιλογή που φαίνεται να ακολούθησαν οι συγγραφείς των μετέπειτα δημοσιεύσεων. Πιο συγκεκριμένα στην περίπτωση του ZSL βρίσκουμε την top-1 ακρίβεια των εικόνων $x \in \mathcal{X}_u$, όπου πιθανές κατηγορίες αποτελούν μόνο οι $|\mathcal{Y}_u|$ άγνωστες κλάσεις. Ο μέσος όρους τους αποτελεί την ακρίβεια, με βάση την οποία αξιολογείται το μοντέλο.

Στην πιο γενική περίπτωση του GZSL, βρίσκουμε την top-1 ακρίβεια ανά κατηγορία για κάθε εικόνα x σε ένα υποσύνολο του $\mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{val}}$, όπου αν και $y \in \mathcal{Y}_s$ δεν λαμβάνουμε υπόψιν μας αυτήν την γνώση. Ο μέσος όρος από αυτές τις ακρίβειες συμβολίζεται με S και δείχνει την επιτυχία του μοντέλου στις γνωστές κλάσεις. Αντίστοιχα βρίσκουμε την top-1 ακρίβεια ανά κατηγορία για κάθε εικόνα $x \in \mathcal{X}_{\text{test}}$, όπου πάλι δεν λαμβάνουμε υπόψιν την γνώση πως $y \in \mathcal{Y}_u$. Ο μέσος όρων από αυτές τις ακρίβειες συμβολίζεται με U και δείχνει την επιτυχία του μοντέλου στις άγνωστες κλάσεις. Η ποσότητα U διαφέρει από το ZSL σκορ στ' ότι το τελευταίο θεωρεί πιθανές κατηγορίες μόνο τις \mathcal{Y}_u και συνεπώς έχει μεγαλύτερη τιμή ακρίβειας. Για το στάθμισμα των S και U , και προκειμένου να βρεθεί μια ποσότητα που αντανακλά την επίδοση του μοντέλου, χρησιμοποιείται ο αρμονικός μέσος όρος τους H . Έτσι στην περίπτωση που οι τιμές S και U διαφέρουν σημαντικά, η τελική αξιολόγηση του μοντέλου H θα είναι πολύ πιο κοντά στην χαμηλότερη από τις 2 μετρικές.

Στην παρούσα διπλωματική ενδιαφερόμαστε για το σενάριο του GZSL. Γιάυτό και στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα σε αυτό το σενάριο.

Μοντέλο	AWA			CUB			SUN		
	S	U	H	S	U	H	S	U	H
DAP	84.7	0	0	67.9	1.7	3.3	25.1	4.2	7.2
IAP	87.6	0.9	1.8	72.8	0.4	0.2	37.8	1.0	1.8
DeViSE	74.7	17.1	27.8	53.0	23.8	32.8	27.4	16.9	20.9
ALE	81.8	14.0	23.9	62.8	23.7	34.4	33.1	21.8	26.3
SJE	73.9	8.0	14.4	59.2	23.5	33.6	30.5	14.7	19.8
LATEM	77.3	11.5	20.0	71.7	7.3	13.3	28.8	14.7	19.5
CMT	89.0	8.7	15.9	60.1	4.7	8.7	28.0	8.7	13.3
ConSE	90.6	0.5	1.0	72.2	1.6	3.1	39.9	6.8	11.6
SYNC	90.5	10.0	18.0	70.9	11.5	19.8	43.3	7.9	13.4
GFZSL	80.1	2.5	4.8	45.7	0.0	0.0	39.6	0.0	0.0
CVAE-ZSL	-	-	51.2	-	-	34.5	-	-	26.7
SE	68.1	58.3	62.8	53.3	41.5	46.7	30.5	40.9	34.9
ReViSE	39.7	46.4	42.8	28.3	37.6	32.3	20.1	24.3	22.0
CADA-VAE	75.0	55.8	63.9	53.5	51.6	52.4	47.2	35.7	40.6

Πίνακας 2.2: Σύγκριση των μοντέλων που παρουσιάστηκαν για το πρόβλημα του GZSL

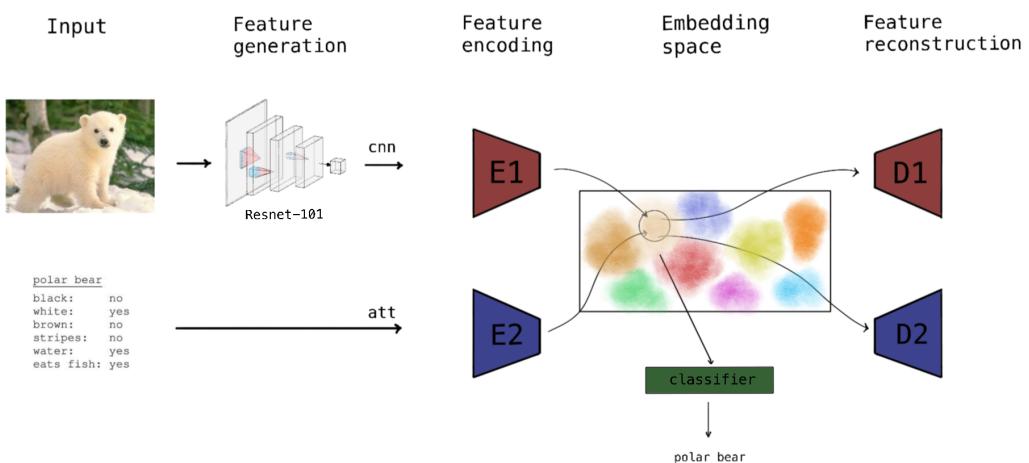
Παρατηρούμε πως τα τελευταία χρόνια, με την χρήση σύνθετων αρχιτεκτονικών όπως οι VAE και τα GAN, έχει υπάρξει μεγάλη άνοδος στα αποτελέσματα. Ειδικά το μοντέλο CADA-VAE, από το οποίο εμπνέεται η προτεινόμενη αντιμετώπιση του προβλήματος, επιτυγχάνει τα καλύτερα αποτελέσματα σε όλα τα σύνολα δεδομένων.

3. Προτεινόμενη αντιμετώπιση

3.1 Η ιδέα

Βάσει της βιβλιογραφικής επισκόπησης του προηγούμενου κεφαλαίου μπορεί εύκολα να αντληθεί το συμπέρασμα πως το μοντέλο CADA-VAE [27] αποτελεί την πιο επιτυχημένη και επεκτάσιμη μεθοδολογία που παρουσιάστηκε. Γι' αυτό και ο τρόπος αντιμετώπισης που ακολουθείται αποτελεί μια παραλλαγή αυτής της μεθόδου. Συγκεκριμένα αντί των κλασικών VAE για καδικοποίηση των οντοτήτων στο χώρο κοινής ενσωμάτωσης χρησιμοποιείται το μοντέλο VaDE [13]. Η παραλλαγή αυτή εξυπηρετεί την καλύτερη μοντελοποίηση του προβλήματος και γενικότερα των προβλημάτων κατηγοριοποίησης, όπως θα γίνει κατανοητό στην συνέχεια. Το προτεινόμενο μοντέλο ονομάζεται **CSVaDE** (Common Space Variational Deep Embedding).

Το CSVaDE, που απεικονίζεται στο Σχήμα 3.1, δέχεται σαν είσοδο χαρακτηριστικά εικόνων και κατηγοριών. Καθώς η εξαγωγή χαρακτηριστικών των εικόνων γίνεται από τα ResNet-101, το προτεινόμενο μοντέλο απλώς χρησιμοποιεί τα χαρακτηριστικά $\phi(x)$ των εικόνων x . Οπότε, για λόγους ευκολίας, θα συμβολίζονται από εδώ και πέρα με x_i τα CNN χαρακτηριστικά του τελευταίου επιπέδου του ResNet-101 της i -οστής εικόνας. Όσο αφορά τα χαρακτηριστικά των κατηγοριών, χρησιμοποιούνται οι περιγραφές των συνόλων δεδομένων $a_i = \pi(y_i) \forall y \in \mathcal{Y}$.



Σχήμα 3.1: Η αρχιτεκτονική του προτεινόμενου μοντέλου.

Για την κωδικοποίηση τόσο των CNN χαρακτηριστικών όσο και των περιγραφών χρησιμοποιούνται 2 κωδικοποιητές E_1 και E_2 . Οι κωδικοποιητές είναι πιθανοτικοί, όπως ακριβώς και στο μοντέλο CADA-VAE, καθώς κωδικοποιούν μια είσοδο x σε μια κατανομή πιθανοτήτων $p(z|x)$ αντί για ένα απλό σημείο z του χώρου κοινής ενσωμάτωσης. Ως κατανομή πιθανοτήτων χρησιμοποιείται η κανονική κατανομή, οπότε η έξοδος του κωδικοποιητή δεν είναι τίποτα άλλο παρά μια μέση τιμή $\mu_1 \in \mathbb{R}^d$ και ένας πίνακας μεταβλητήτης $\Sigma_1 \in \mathbb{R}^{d \times d}$, όπου d ο αριθμός διαστάσεων του χώρου κοινής ενσωμάτωσης. Για να απλοποιηθεί το πρόβλημα ο πίνακας Σ_1 θεωρείται διαγώνιος και έτσι μπορεί να αναπαρασταθεί μόνο με τα στοιχεία της κύριας διαγώνιου του $\sigma_1 \in \mathbb{R}^d$. Αντίστοιχα οι περιγραφές a κωδικοποιούνται σε μια κανονική κατανομή $\mathcal{N}(\mu_2, \Sigma_2)$ με τον πίνακα Σ_2 να είναι πάλι διαγώνιος.

Οι αποκωδικοποιητές D_1, D_2 είναι επίσης πιθανοτικοί. Δέχονται σαν είσοδο σημεία του χώρου κοινής ενσωμάτωσης και δίνουν ως έξοδο τις παραμέτρους των κατανομών των CNN χαρακτηριστικών των εικόνων και των περιγραφών, αντίστοιχα. Έτσι, διαλέγοντας τα διανύσματα \tilde{x} και \tilde{a} που μεγιστοποιούν τις κατανομές $p(x|z)$ και $p(a|z)$ που αποκωδικοποιήσαν οι D_1 και D_2 , μπορούμε να ανακατασκευάσουμε τις αρχικές οντότητες x και a .

Κατά την διάρκεια εκπαίδευσης του μοντέλου, οι οντότητες κωδικοποιούνται και αποκωδικοποιούνται, με αποτέλεσμα οι κωδικοποιητές να σχηματίζουν περιοχές στον χώρο κοινής ενσωμάτωσης που αντιστοιχούν σε εικόνες / περιγραφές της ίδιας κατηγορίας, και οι αποκωδικοποιητές να μπορούν να ανακατασκευάσουν τα CNN χαρακτηριστικά των εικόνων και τις περιγραφές τους από αυτές τις περιοχές.

Η βασική ιδέα για την επίλυση του προβλήματος GZSL είναι ίδια με την υποενότητα 2.1.5. Εκπαιδεύοντας τους 2 VaDE μόνο με τα δεδομένα από τις γνωστές κλάσεις, μπορούμε να διαμορφώσουμε κατάλληλες περιοχές στον χώρο κοινής ενσωμάτωσης τόσο για γνωστές όσο και για άγνωστες κλάσεις. Για παράδειγμα αν στις γνωστές κατηγορίες υπάρχουν ποδήλατα και αυτοκίνητα, όταν στον κωδικοποιητή E_1 δοθούν τα CNN χαρακτηριστικά μιας εικόνας μηχανής θα την κωδικοποιήσει κάπου ανάμεσα. Αν οι περιγραφές του αυτοκινήτου, του ποδηλάτου και της μηχανής είναι λεπτομερείς και ακριβείς, ο κωδικοποιητής περιγραφών E_2 θα πρέπει να κωδικοποιεί την περιγραφή της μηχανής ανάμεσα στους αυτοκινήτου και του ποδηλάτου, ιδανικά εκεί που θα κωδικοποιηθούν και τα CNN χαρακτηριστικά. Έτσι η εικόνα μιας άγνωστης κατηγορίας („μηχανή“) κωδικοποιήθηκε σε μια γνωστή περιοχή, καθώς η περιγραφή της μηχανής ήταν γνωστή κατά την εκπαίδευση του μοντέλου.

Για την κατηγοριοποίηση χρησιμοποιείται ένας απλός ταξινομητής, όπως μια Softmax ή μία Μηχανή Διανυσμάτων Υποστήριξης (Support Vector Machine - SVM), για το διαμερισμό του χώρου σε $|\mathcal{Y}|$ κλάσεις. Συγκεκριμένα, υποθέτοντας ότι οι κωδικοποιητές E_1 και E_2 είναι συγχρονισμένοι, δημιουργούνται N σημεία $z_y^{(i)} \sim \mathcal{N}(\mu_y, \Sigma_y)$ $i = 1, 2, \dots, N$, όπου $\mu_y, \Sigma_y = E_2(a_y)$. Ο ταξινομητής δέχεται σαν είσοδο τα σημεία $z_y^{(i)}$ και έξοδο τις κατηγορίες $y \in \mathcal{Y}$. Κατά την διάρκεια της αξιολόγησης του μοντέλου τα CNN χαρακτηριστικά μιας εικόνας δίνονται στον κωδικοποιητή E_1 και η κωδικοποίηση τους εισάγεται στον ταξινομητή του χώρου κοινής ενσωμάτωσης, για την εύρεση της κατηγορίας στην οποία ανήκουν.

3.2 Το κίνητρο

Η κύρια διαφορά της προτεινόμενης αντιμετώπισης με αυτήν που παρουσιάζεται στην δημοσίευση του CADA-VAE είναι η αντικατάσταση των κλασσικών VAE με το μοντέλο VaDE. Οι VAE κάνουν την υπόθεση πως όλα τα αντικείμενα που αναπαρίστανται στον λανθάνων χώρο (latent space) προέρχονται από την ίδια περιοχή, από την ίδια γκαουσιανή κατανομή

$N(0, I)$. Αυτό δεν φαίνεται να εξυπηρετεί τις ανάγκες του προβλήματος του GZSL, καθώς υπάρχουν αντικείμενα από πολλές κατηγορίες. Θα ήταν πολύ λογικό τα αντικείμενα της ίδιας κατηγορίας να προέρχονται από την ίδια περιοχή, κι συγκεκριμένα από την ίδια κατανομή, αλλά αντικείμενα διαφορετικών κατηγοριών να ανήκουν σε διαφορετικές κατανομές.

Πιο συγκεκριμένα, θεωρούμε πως οι κλάσεις του συνόλου δεδομένων μπορούν να περιγραφούν σε έναν χώρο d διαστάσεων, έτσι ώστε να είναι οργανωμένες σε συμπλέγματα (clusters). Κάθε σύμπλεγμα αντιστοιχεί σε μία μόνο κλάση, και το σημείο στο κέντρο του αναπαριστά την πιο τυπική μορφή μιας εικόνας από την κλάση αυτή. Τα σημεία γύρω από το κέντρο, που ανήκουν στο σύμπλεγμα, θεωρούμε πως αναπαριστούν εικόνες της κατηγορίας με 1-διαίτερα οπτικά χαρακτηριστικά, που προσδιορίζονται από την σχετική θέση του σημείου ως προς το κέντρο. Γίνεται η υπόθεση πως όπως πολλά πράγματα στην φύση, έτσι και τα σημεία αυτού του χώρου ακολουθούν κανονική κατανομή ανά κατηγορία, σχηματίζοντας έτσι μια Μίξη Γκαουσιανών (Gaussian Mixture Model). Έτσι γίνεται σαφές γιατί ότι ένας κλασσικός VAE, που μοντελοποιεί τα σημεία στον χώρο κοινής ενσωμάτωσης ώστε να ακολουθούν κανονική κατανομή $\mathcal{N}(0, I_d)$, δεν είναι η καταλληλότερη επιλογή τόσο για ένα πρόβλημα κατηγοριοποίησης όσο και για την πιο ιδιαίτερη θεώρηση που περιγράφεται εδώ.

Για να ικανοποιηθεί η παραπάνω απαίτηση πέρα από τα μοντέλα E_1, E_2, D_1, D_2 εισάγονται και οι παράμετροι M, Σ . Ο πίνακας $M = [\mu_1, \mu_2, \dots, \mu_{|\mathcal{Y}|}]$ περιέχει τις μέσες τιμές των κατανομών των κλάσεων (που ταυτίζονται με τις συντεταγμένες των κέντρων των συστάδων), ενώ ο $\Sigma = [\sigma_1, \sigma_2, \dots, \sigma_{|\mathcal{Y}|}]$ τις τυπικές αποκλίσεις των κατανομών. Αυτοί οι δύο πίνακες φροντίζουν να συγχρονίσουν τους δύο VaDE μεταξύ τους χωρίς την ανάγκη επιπλέον όρων στην συνάρτηση σφάλματος.

Ουσιαστικά και το CADA-VAE προσπαθεί να πετύχει την παραπάνω ιδέα επιβαρύνοντας κι άλλο την συνολική συνάρτηση σφάλματος. Προσπαθεί να κρατήσει το κομμάτι των VAE για την δημιουργία ενός λανθάνων χώρου με βολικές ιδιότητες, αλλά ταυτόχρονα επιβάλει Distribution Alignment ώστε οι κατανομές CNN χαρακτηριστικών και οι περιγραφές από αντικείμενα που ανήκουν στην ίδια κλάση να έχουν ίδια κατανομή. Επιπλέον επιβάλει Cross-Aligment, που είναι μια ευρύτερη τακτική στον χώρο της μεταφοράς γνώσης (transfer learning), ώστε να πετύχει καλύτερο συγχρονισμό κατανομών μέσα στην ίδια κατηγορία.

Αντιθέτως το μοντέλο VaDE υποθέτει πως τα αντικείμενα προέρχονται από μια Μίξη Γκαουσιανών (Guassian Mixture), και αυτό επιτρέπει τον χωρισμό αντικείμενων από διαφορετικές κατηγορίες σε διαφορετικές Γκαουσιανές. Αν και το μοντέλο που παρουσιάζεται στο [13] χρησιμοποιείται για επίλυση προβλημάτων ομαδοποίησης, εδώ χρησιμοποιείται μια ελαφρώς τροποποιημένη εκδοχή του, προσαρμοσμένη για προβλήματα ημι-επιβλεπώμενης μάθησης.

3.3 Το μοντέλο VaDE

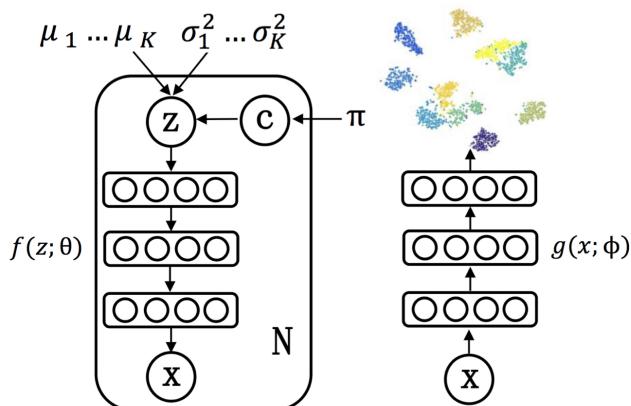
Στην προηγούμενη ενότητα περιγράφηκε η κεντρική διαφορά μεταξύ του VaDE και του VAE, πάνω στην οποία στηρίζεται το CSVaDE. Ωστόσο λείπει ο μαθηματικός φορμαλισμός που αποδεικνύει την καλή λειτουργία του VaDE, ο οποίος αναλύεται σε αυτήν την ενότητα.

3.3.1 Η παραγωγική διαδικασία

Καθώς το VaDE είναι παραγωγικό μοντέλο η θεωρία του στηρίζεται στην διαδικασία με την οποία παράγεται ένα δείγμα x από τον λανθάνων χώρο των z , ως εξής:

- Επιλέγεται μια συστάδα (κατηγορία) $y \in \mathcal{Y}$ με πιθανότητα $p(y) \sim \text{Cat}(\pi)$, όπου Cat είναι η κατηγορική κατανομή.
 - Επιλέγεται ένα διάνυσμα $z \sim \mathcal{N}(\mu_y, \sigma_y^2 I)$ του λανθάνοντος χώρου.
 - Επιλέγεται ένα δείγμα x . Εδώ διακρίνουμε 2 περιπτώσεις:
 - Αν το x λαμβάνει δυαδικές τιμές, υπολογίζουμε την μέση τιμής της κατανομής του $\mu_x = f(z)$ και διαλέγουμε ένα δείγμα $x \sim \text{Ber}(\mu_x)$
 - Αν το x λαμβάνει πραγματικές τιμές, υπολογίζουμε την μέση τιμή και την διασπορά του $[\mu_x, \log \sigma_x^2] = f(z)$ και διαλέγουμε ένα δείγμα $x \sim \mathcal{N}(\mu_x, \sigma_x^2 I)$

Δηλαδή η συνάρτηση f λειτουργεί ως πιθανοτικός αποκωδικοποιητής, και μπορεί να προσεγγιστεί με την χρήση νευρωνικών δικτύων όπως φαίνεται στο Σχήμα 3.2. Αντίστοιχα βλέπουμε πως η συνάρτηση g του σχήματος λειτουργεί σαν πιθανοτικός κωδικοποιητής. Δηλαδή ο κωδικοποιητής g με είσοδο x έχει ως έξοδο τα διανύσματα μ_z και σ_z . Με βάση αυτά μπορεί να δειγματοληπτηθεί ένα σημείο $z \sim \mathcal{N}(\mu_z, \sigma_z)$ του λανθάνοντος χώρου.



Σχήμα 3.2: Σχηματική αναπαράσταση της παραγωγικής διαδικασίας του VaDE

3.3.2 Η συνάρτηση σφάλματος

Αντικειμενικός στόχος είναι η εύρεση των συναρτήσεων f και g καθώς και των παραμέτρων $M = [\mu_1, \mu_2, \dots, \mu_{|\mathcal{Y}|}]$ και $\Sigma = [\sigma_1, \sigma_2, \dots, \sigma_{|\mathcal{Y}|}]$, για να είναι δυνατή η μετατροπή από τον χώρο των δειγμάτων x στον λανθάνων χώρο τον z και αντίστροφα. Αν και η αναλυτική εύρεση των f και g δεν είναι εφικτή, μπορούν να προσεγγιστούν από νευρωνικά έχοντας ικανό αριθμό δειγμάτων x . Συγκεκριμένα, έχοντας επιλέξει μια αρχιτεκτονική για τον κωδικοποιητή E και τον αποκωδικοποιητή D , πρέπει να βρεθούν οι βέλτιστες παράμετροι τους θ και ϕ , ώστε $f(z) \approx E(z; \theta)$ και $g(x) \approx D(x; \phi)$. Αυτό, όπως σε πολλά προβλήματα της μηχανικής μάθησης, μπορεί να επιτευχθεί με την μέθοδο της μέγιστης πιθανοφάνειας.

Σύμφωνα με την προηγούμενη υποενότητα $p(x, z, y) = p(x|z)p(z|y)p(y)$ όπου

$$\begin{aligned} p(y) &= \text{Cat}(y|\pi) \\ p(z|y) &= \mathcal{N}(z|\mu_y, \sigma_y^2 I) \\ p(x|z) &= \text{Ber}(x|\mu_x) \dot{\wedge} \mathcal{N}(x|\mu_x, \sigma_x^2 I) \end{aligned} \quad (3.1)$$

οπότε

$$p(x) = \int_z \sum_y p(x, z, y) = \int_z \sum_y p(x|z)p(z|y)p(y)$$

Άρα αρκεί να μεγιστοποιηθεί η ποσότητα

$$\mathcal{L}(x) = \log p(x) = \log \int_z \sum_y p(x, z, y)$$

Η τιμή αυτή είναι υπολογιστικά απρόσιτη (intractable), και γι' αυτό χρησιμοποιούνται μεταβολικές μέθοδοι προσέγγισης (variational inference).

Αφού

$$\mathcal{L}(x) = \log \int_z \sum_y p(x, z, y) = \log \int_z \sum_y q(z, y|x) \frac{p(x, z, y)}{q(z, y|x)} = \log \mathbb{E}_{q(z, y|x)} \left[\frac{p(x, z, y)}{q(z, y|x)} \right]$$

Σύμφωνα με την ανισότητα Jensen θα ισχύει:

$$\mathcal{L}(x) \geq \mathbb{E}_{q(z, y|x)} \left[\log \frac{p(x, z, y)}{q(z, y|x)} \right] \triangleq \mathcal{L}_{\text{ELBO}}(x)$$

όπου $q(z, y|x)$ είναι η προσέγγιση της πραγματικής κατανομής $p(z, y|x)$ ¹. Έτσι μπορεί να μεγιστοποιηθεί η ποσότητα $\mathcal{L}_{\text{ELBO}}$ αντί της \mathcal{L} , οδηγώντας ίσως όχι σε βέλτιστα, αλλά πιθανότατα καλά αποτελέσματα. Υπό την υπόθεση πως η ποσότητα $q(z, y|x)$ είναι προσέγγιση μέσου πεδίου, μπορεί να παραγοντοποιηθεί ως:

$$q(z, y|x) = q(z|x)q(y|x) \quad (3.2)$$

Σύμφωνα με τις εξισώσεις 3.1 και 3.2 το μεταβολικό κάτω φράγμα (variational lower bound) μπορεί να γραφεί ως:

$$\mathcal{L}_{\text{ELBO}} = E_{q(z, y|x)} \left[\log \frac{p(x, z, y)}{q(z, y|x)} \right] \quad (3.3)$$

$$= \mathbb{E}_{q(z, y|x)} [\log p(x, z, y) - \log q(z, y|x)] \quad (3.4)$$

$$= \mathbb{E}_{q(z, y|x)} [\log p(x|z) + \log p(z|y) + \log p(y) - \log q(z|x) - \log q(y|x)] \quad (3.5)$$

¹Η ανισότητα ισχύει για οποιαδήποτε κατανομή q , απλά για $q \approx p$ η ποσότητα $\log \frac{p(x, z, y)}{q(z, y|x)} \approx p(x) = \text{const}$ και συνεπώς το μεταβολικό κάτω φράγμα EBLO προσεγγίζει καλύτερα την πραγματική συνάρτηση σφάλματος.

Οι κατανομές $p(x|z)$, $p(z|y)$ και $p(y)$ δίνονται από την σχέση 3.1 και συνεπώς μένει να υπολογιστούν οι $q(z|x)$ και $q(y|x)$.

- Η $q(z|x)$ μοντελοποιείται με μία κανονική κατανομή $\mathcal{N}(z; \mu_z, \sigma_z^2 I)$ όπου $[\mu_z, \sigma_z] = E(x; \phi)$.

Δηλαδή χρησιμοποιείται ένα νευρωνικό με παραμέτρους ϕ για την εκτίμηση των παραμέτρων μ_z, σ_z της κανονικής κατανομής $q(z|x)$

Έτσι

$$q(z|x) = \mathcal{N}(z; \mu_z, \sigma_z^2 I) \quad (3.6)$$

- Ο υπολογισμός της $q(y|x)$ είναι λίγο πιο περίπλοκος και απαιτεί μερικές υποθέσεις. Αρχικά το μεταβολικό κάτω φράγμα μπορεί να ξαναγραφεί ως:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} &= \mathbb{E}_{q(z,y|x)} \left[\log \frac{p(x, z, y)}{q(z, y|x)} \right] \\ &= \int_z \sum_y q(y|x) q(z|x) \left[\log \frac{p(x|z)p(z)}{q(z|x)} + \log \frac{p(y|z)}{q(y|x)} \right] \\ &= \int_z q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)} - \int_z q(z|x) D_{\text{KL}}(q(y|x)||p(y|z)) \end{aligned}$$

Στα πλαίσια μεγιστοποίησης του, και καθώς ο πρώτος όρος δεν έχει σχέση με την κατηγορία y ενώ ο δεύτερος είναι μη αρνητικός, θα πρέπει:

$$D_{\text{KL}}(q(y|x)||p(y|z)) = 0$$

Έτσι

$$q(y|x) = p(y|z) = \frac{p(y)p(z|y)}{\sum_{y'=1}^{|Y|} p(y')(z|y')} \triangleq \gamma_y \quad (3.7)$$

Χρησιμοποιώντας τις εξισώσεις 3.1 και 3.5 - 3.7 η $\mathcal{L}_{\text{ELBO}}$ γράφεται:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} &= \frac{1}{L} \sum_{l=1}^L \sum_{i=1}^D \left[x_i \log \mu_x^{(l)}|_i + (1 - x_i) \log(1 - \mu_x^{(l)}|_i) \right] \\ &\quad - \frac{1}{2} \sum_{y=1}^K \gamma_y \sum_{j=1}^J \left[\log \sigma_y^2|_j + \frac{\sigma_z^2|_j}{\sigma_y^2|_j} + \frac{(\mu_z|_j - \mu_y|_j)^2}{\sigma_y^2|_j} \right] \\ &\quad + \sum_{y=1}^{|Y|} \gamma_y \log \frac{\pi_y}{\gamma_y} + \frac{1}{2} \sum_{j=1}^J (1 + \log \sigma_z^2|_j) \end{aligned}$$

αν τα δείγμα x ακολουθούν κατανομή Bernoulli ενώ αν ακολουθούν Γκαουσιανή κατανομή:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = & \frac{1}{L} \sum_{l=1}^L -\frac{1}{2} \sum_{i=1}^D \left[2 \log \sigma_x^{(l)}|_i + \frac{(x - \mu_x^{(l)})^2|_i}{\sigma_x^{(l)2}|_i} + \log(2\pi) \right] \\ & - \frac{1}{2} \sum_{y=1}^{|Y|} \gamma_y \sum_{j=1}^d \left[\log \sigma_y^2|_j + \frac{\sigma_z^2|_j}{\sigma_y^2|_j} + \frac{(\mu_z|_j - \mu_y|_j)^2}{\sigma_y^2|_j} \right] \\ & + \sum_{y=1}^{|Y|} \gamma_y \log \frac{\pi_y}{\gamma_y} + \frac{1}{2} \sum_{j=1}^d (1 + \log \sigma_z^2|_j)\end{aligned}$$

όπου L ο αριθμός των δειγμάτων Monte Carlo, D ο αριθμός διαστάσεων των δειγμάτων x και d ο αριθμός διαστάσεων του λανθάνοντος χώρου. Επίσης με $a|_i$ συμβολίζεται το i -οστό στοιχείο του διανύσματος a . Οι πράξεις για των υπολογισμό των παραπάνω συναρτήσεων βρίσκονται στα παραρτήματα του [13].

Αξίζει να σημειωθεί πως στην πράξη όταν τα δεδομένα παίρνουν πραγματικές τιμές, χάριν απλότητας, χρησιμοποιείται η ποσότητα $-\|x_i - \mu_x^{(l)}\|$ αντί της $\log \mathcal{N}(x; \mu_x, \sigma_x)$. Έτσι σαν μεταβολικό κάτω όριο χρησιμοποιείται η ποσότητα

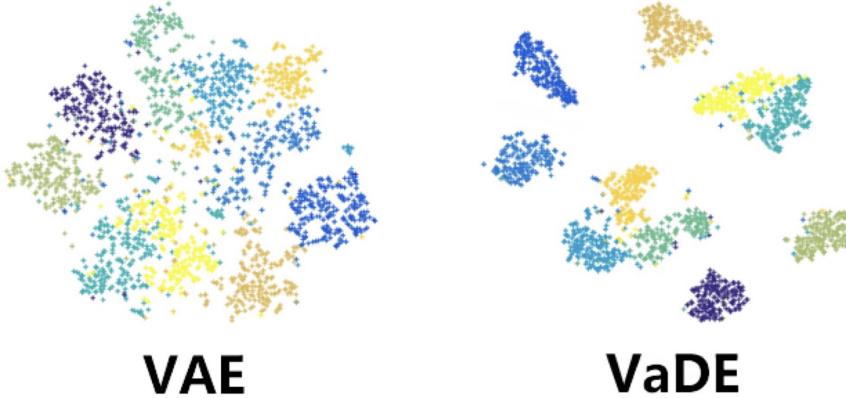
$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = & \frac{1}{L} \sum_{l=1}^L -\|x_i - \mu_x^{(l)}\| \\ & - \frac{1}{2} \sum_{y=1}^{|Y|} \gamma_y \sum_{j=1}^d \left[\log \sigma_y^2|_j + \frac{\sigma_z^2|_j}{\sigma_y^2|_j} + \frac{(\mu_z|_j - \mu_y|_j)^2}{\sigma_y^2|_j} \right] \\ & + \sum_{y=1}^{|Y|} \gamma_y \log \frac{\pi_y}{\gamma_y} + \frac{1}{2} \sum_{j=1}^d (1 + \log \sigma_z^2|_j)\end{aligned}$$

3.3.3 Ενδεικτικό παράδειγμα - MNIST

Για την καλύτερη κατανόηση του τρόπου λειτουργίας των VaDE και την αντιπαράθεση τους με τους αυτοκωδικοποιητές VAE χρησιμοποιείται το σύνολο δεδομένων MNIST. Το σύνολο αποτελείται από ασπρόμαυρες εικόνες χειρόγραφων ψηφίων από το 0 έως το 9, και μπορεί να χρησιμοποιηθεί είτε για ταξινόμηση των χειρόγραφων ψηφίων είτε για ομαδοποίηση. Στην περίπτωση της συγκεκριμένης σύγκρισης γίνεται ομαδοποίηση στον λανθάνων χώρο του μοντέλου.

Στην εικόνα 3.3 βλέπουμε τον λανθάνοντα χώρο 2 διαστάσεων των 2 μοντέλων. Είναι προφανές πως το VaDE μοντελοποιεί καλύτερα το πρόβλημα, αφού τα ψηφία μπορούν να κατανεμηθούν το καθένα σε διαφορετική κατανομή. Αντίθετα στο μοντέλου του VAE υποτίθεται πως όλα ανήκουν σε μια κοινή κατανομή, κι έτσι είναι λιγότερο ευδιάκριτα μεταξύ τους.

Τέλος, μπορεί να παρατηρηθεί πως οι κατανομές στο σχήμα του VaDE μπορεί να είναι κοντά μεταξύ τους (για ψηφία που μοιάζουν όπως το 4 και το 9) ή να είναι απομονωμένες (για ψηφία όπως το 0). Έτσι ο λανθάνων χώρος αντικατοπτρίζει καλύτερα τις ιδιομορφίες των χειρόγραφων ψηφίων, ενώ ταυτόχρονα πετυχαίνει και καλύτερη ομαδοποίηση. Τέλος οι κωδικοποιήσεις στον λανθάνων χώρο μπορούν να χρησιμοποιηθούν για ταξινόμηση των ψηφίων.



Σχήμα 3.3: Σύγκριση του VAE και του VaDE στο σύνολο δεδομένων MNIST.

3.3.4 Τροποποίηση για ημι-επιβλεπόμενη μάθηση

Το μοντέλο του VaDE που παρουσιάστηκε μέχρι στιγμής χρησιμοποιείται σε προβλήματα ομαδοποίησης. Καθώς το πρόβλημα του GZSL, με το οποίο ασχολείται η διπλωματική, είναι πρόβλημα ημι-επιβλεπόμενης μάθησης, πρέπει να γίνουν κάποιες τροποποιήσεις στο πρωτότυπο μοντέλο του VaDE.

Αρχικά, εφόσον κατά την διάρκεια της εκπαίδευσης το νευρωνικό δίκτυο τροφοδοτείται με ζευγάρια (x, y) η πιθανότητα $q(y'|x)$ γίνεται:

$$\gamma_y = q(y'|x) = \begin{cases} 1, & \text{αν } y' = y \\ 0, & \text{αλλιώς} \end{cases}$$

Έτσι η συνάρτηση $\mathcal{L}_{\text{ELBO}}$ μετατρέπεται στην

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} = & -\frac{1}{L} \sum_{l=1}^L \|x - \mu_x^{(l)}\|_2^2 - \frac{1}{2} \sum_{j=1}^d \left[\log \sigma_y^2|_j + \frac{\sigma_z^2|_j}{\sigma_y^2|_j} + \frac{(\mu_z|_j - \mu_y|_j)^2}{\sigma_y^2|_j} \right] \\ & + \log \pi_y + \frac{1}{2} \sum_{j=1}^d (1 + \log \sigma_z^2|_j) \end{aligned}$$

Επίσης, σύμφωνα με το [14] για ικανοποιητικά μεγάλο μέγεθος παρτίδας (batch size) μπορεί να χρησιμοποιηθεί $L = 1$ στην δειγματοληψία Monte-Carlo. Τέλος, αφού όπως αναφέρθηκε και προηγουμένως το μοντέλο τροφοδοτείται με ζευγάρια (x, y) , το διάνυσμα π είναι γνωστές και σταθερές, άρα ο όρος $\log \pi_y$ μπορεί να απαλειφθεί. Έτσι το κάτω μεταβολικό φράγμα γίνεται:

$$\mathcal{L}_{\text{ELBO}} = -\|x - \mu_x\| + \frac{1}{2} \sum_{j=1}^d \left[1 + (\log \sigma_z^2|_j - \log \sigma_y^2|_j) - \frac{\sigma_z^2|_j}{\sigma_y^2|_j} - \frac{(\mu_z|_j - \mu_y|_j)^2}{\sigma_y^2|_j} \right]$$

και η τελική συνάρτηση σφάλματος του μοντέλου VaDE που χρησιμοποιείται στα πειράματα

είναι:

$$\mathcal{L} = -\mathcal{L}_{\text{ELBO}} = \underbrace{\|x - \mu_x\|}_{\text{Σφάλμα ανακατασκευής}} - \underbrace{\frac{1}{2} \sum_{j=1}^d \left[1 + (\log \sigma_z^2|_j - \log \sigma_y^2|_j) - \frac{\sigma_z^2|_j}{\sigma_y^2|_j} - \frac{(\mu_z|_j - \mu_y|_j)^2}{\sigma_y^2|_j} \right]}_{\text{Σφάλμα κανονικοποίησης}}$$

3.4 Η αρχιτεκτονική

Όπως φαίνεται από την Εικόνα 3.1 η αρχιτεκτονική του CSVaDE αποτελείται από 6 επιμέρους τμήματα:

- Ένα νευρωνικό δίκτυο που λειτουργεί ως εξαγωγέας χαρακτηριστικών, στην περίπτωση μας το ResNet-101.
- Δύο νευρωνικά δίκτυα E_1 και E_2 που λειτουργούν ως κωδικοποιητές, ένα για τα CNN χαρακτηριστικά των εικόνων και ένα για τις περιγραφές
- Δύο νευρωνικά δίκτυα D_1 και D_2 που λειτουργούν ως αποκωδικοποιητές, αντίστοιχα
- Μία μονάδα ταξινόμησης

3.4.1 Εξαγωγέας χαρακτηριστικών

Ως εξαγωγέας χαρακτηριστικών χρησιμοποιείται το νευρωνικό δίκτυο ResNet-101 [12], που αποτελεί μια τυπική επιλογή στην βιβλιογραφία τα τελευταία χρόνια. Σαν είσοδο δέχεται έναν $M \times N \times 3$ πίνακα, που αναπαριστά μια RGB εικόνα, και στην έξοδο έχει 2048 χαρακτηριστικά που περιγράφουν την εικόνα. Μια τέτοια συμπίεση της πληροφορίας είναι πολύ χρήσιμη καθώς μειώνει δραματικά των όγκο των δεδομένων που μεταφέρονται στο υπόλοιπο μοντέλο και συνεπώς την διάρκειας της προσοτροφοδότησης και της οπίσθιας τροφοδότησης του μοντέλου.

Προκειμένου να μειωθεί κι άλλο ο συνολικός χρόνος εκπαίδευσης χρησιμοποιούνται απευθείας ως είσοδοι στο υπόλοιπο σύστημα τα CNN χαρακτηριστικά που έχουν εξαχθεί σε προηγούμενο χρόνο από προ-εκπαίδευμένο μοντέλο². Έτσι το σύστημα φορτώνει πιο γρήγορα τα δεδομένα, μπορεί να επεξεργαστεί μεγαλύτερα μεγέθη παρτίδας, και χρειάζεται λιγότερο χρόνο για πρόσθια και οπίσθια τροφοδότηση. Τα μοντέλο που χρησιμοποιήθηκε για την εξαγωγή των χαρακτηριστικών είναι εκπαίδευμένο σε υποσύνολο 1000 εικόνων του ImageNet, σύμφωνα με τον προτεινόμενο διαχωρισμό (Proposed Split - PS) του [36]. Επιπλέον, μετά το πέρας της εκπαίδευσης, μπορεί να προσαρτηθεί το νευρωνικό ResNet στο σύστημα και ως είσοδοι να χρησιμοποιούνται εικόνες αντί για CNN χαρακτηριστικά. Εκπαίδευοντας το συνολικό μοντέλο με αυτόν τον τρόπο για μερικές ακόμα επαναλήψεις θα προσαρμόσει (fine-tune) το ResNet και θα παράγει τα βέλτιστα αποτελέσματα. Βέβαια, καθώς αυτή η τεχνική δεν έχει χρησιμοποιηθεί στην βιβλιογραφία, και δίνεται να βελτιώσει ελαφρώς τα αποτελέσματα, ξεφεύγει του ενδιαφέροντος της εργασίας.

Τέλος, θα μπορούσε να χρησιμοποιηθεί οποιοδήποτε άλλο μοντέλο της οικογενείας ResNet ή και εκτός, αλλά η διερεύνηση του εξαγωγέα χαρακτηριστικών δεν είναι στους σκοπούς της παρούσας διπλωματικής εργασίας.

²Τα CNN χαρακτηριστικά που χρησιμοποιήθηκαν βρίσκονται εδώ <https://www.dropbox.com/sh/btoc495ytfbnbat/AAaurkoKnnk0uV-swqF-gdSa?dl=0>

3.4.2 Κωδικοποιητές

Ο κωδικοποιητής αποτελούνται από ένα πλήρως συνδεδεμένο πολυστρωματικό νευρωνικό δίκτυο με $k \geq 2$ επίπεδα. Μετά από κάθε επίπεδο, εκτός του τελευταίου, χρησιμοποιείται η συνάρτηση ReLU ως μη-γραμμικότητα. Το τελευταίο επίπεδο χωρίζεται σε δύο πανομοιότυπα τμήματα, που αποτελούνται από d νευρώνες το καθένα, και οι έξοδοι τους προσεγγίζουν την μέση τιμή μ_z και την τυπική απόκλιση σ_z του δείγματος στον λανθάνοντα χώρο. Έπειτα γίνεται δειγματοληψία για την επιλογή ενός δείγματος του λανθάνων χώρου $z \sim \mathcal{N}(\mu, \sigma)$.

3.4.3 Αποκωδικοποιητές

Οι αποκωδικοποιητές αποτελούνται επίσης ένα πολυστρωματικό νευρωνικό δίκτυο με $l \geq 2$ επίπεδα. Στο πρώτο επίπεδο δέχεται ως είσοδο την αναπαράσταση z του λανθάνοντος χώρου και έχει ως έξοδο τις εκτιμήσεις των παραμέτρων της κατανομής x . Καθώς τα CNN χαρακτηριστικά έχουν πραγματικές τιμές, υποθέτουμε πως οι ανακατασκευές τους ακολουθούν κανονική κατανομή, και συνεπώς η έξοδος θα έχει την μορφή $[\mu_x, \sigma_x] = D(z; \theta)$. Ωστόσο στην συνάρτηση σφάλματος εμπεριέχεται μόνο η ποσότητα μ_x , και συνεπώς δε χρειάζεται να υπολογίσουμε την σ_x . Επιπλέον, ακόμα και αν μας ενδιέφερε η ανακατασκευή \tilde{x} του x , θα επιλέγαμε την τιμή \tilde{x} της κατανομής $\mathcal{N}(\mu_x, \sigma_x)$ με την μεγαλύτερη πιθανοφάνεια, κι αυτή είναι η μ_x . Οπότε σε κάθε περίπτωση ο πιθανοτικός κωδικοποιητής δίνει ως έξοδο μόνο την ποσότητα μ_x , η οποία καταχρηστικά σε κάποια διαγράμματα μπορεί να εμφανίζεται ως \tilde{x} .

3.4.4 Ταξινομητής

Ο ταξινομητής δέχεται σαν είσοδο ένα σημείο του χώρου κοινής ενσωμάτωσης και το ταξινομεί σε μία από τις $|\mathcal{Y}|$ διαθέσιμες κατηγορίες. Αν και στο [37] δοκιμάζονται αρκετά μοντέλα ως πιθανοί ταξινομητές, οι συγγραφείς καταλήγουν πως ο πιο αποτελεσματικός είναι μια απλή συνάρτηση Softmax. Στα πειράματα του επόμενου κεφαλαίου θα δοκιμαστεί τόσο η συνάρτηση Softmax λόγω της ευρίας χρήσης της στην βιβλιογραφία, όσο και ένας SVM ταξινομητής με rbf kernel (radial basis function kernel) λόγω της γκαουσιανής φύσης του χώρου κοινής ενσωμάτωσης. Παρουσιάζεται σύγκριση των δύο ταξινομητών στην υποενότητα 4.2.3.

3.5 Η μεθοδολογία εκπαίδευσης

Η εκπαίδευση χωρίζεται σε 2 τμήματα. Αρχικά εκπαιδεύονται συγχρόνως οι 2 VaDE με βάση την συνάρτηση σφάλματος της υποενότητας 3.3.2 με σκοπό την δημιουργία ενός "καλού" χώρου κοινής ενσωμάτωσης. Έπειτα εκπαιδεύεται ο ταξινομητής με συνθετικά στοιχεία αυτού του χώρου.

3.5.1 Εκπαίδευση των VaDE

Το μοντέλα αυτοκωδικοποιητών VaDE για τα CNN χαρακτηριστικά και τις περιγραφές εκπαιδεύονται πάνω στο σύνολο εκπαίδευσης, που αποτελείται μόνο από τις γνωστές κλάσεις.

'Οπως αναφέρθηκε στην ενότητα 3.3 η συνάρτηση σφάλματος για την κάθε οντότητα μπορεί να γραφεί ως:

$$\mathcal{L} = \mathcal{L}_{\text{reconstruction}} + \mathcal{L}_{\text{KLD}}$$

Έτσι η συνολική συνάρτηση σφάλματος του μοντέλου στο πρώτο στάδιο της εκπαίδευσης είναι:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{cnn}} + \mathcal{L}_{\text{att}} \\ &= \mathcal{L}_{\text{cnn-reconstruction}} + \mathcal{L}_{\text{cnn-KLD}} + \mathcal{L}_{\text{att-reconstruction}} + \mathcal{L}_{\text{att-KLD}}\end{aligned}$$

Ωστόσο το [27] και το [4] χρησιμοποιούν ένα συντελεστή β για τον όρο \mathcal{L}_{KLD} ο οποίος ξεκινάει από το 0 και αυξάνει γραμμικά για έναν συγκεκριμένο αριθμό επαναλήψεων. Ακολουθώντας αυτήν την τεχνική η συνάρτηση σφάλματος γίνεται:

$$\mathcal{L} = (\mathcal{L}_{\text{cnn-reconstruction}} + \mathcal{L}_{\text{att-reconstruction}}) + \beta \cdot (\mathcal{L}_{\text{cnn-KLD}} + \mathcal{L}_{\text{att-KLD}})$$

Διαισθητικά, αυτός ο τρόπος εξυπηρετεί την εξεύρεση ενός χώρου κοινής ενσωμάτωσης με “καλές” αναπαραστάσεις πριν αυτός ο χώρος αποκτήσει την μορφή μείζης γκαουσιανών. Μπορεί να θεωρηθεί σαν μια ομαλή μετάβαση από ένα κλασσικό αυτοκωδικοποιητή σε έναν VAE / VaDE, και αν και στην αρχή δεν ελαχιστοποιείται το πραγματικό μεταβολικό κάτω φράγμα, έχει παρατηρηθεί καλύτερη βελτιστοποίηση του στο τέλος της εκπαίδευσης.

Στο CADA-VAE η εποχή που η υπερπαράμετρος β ξεκινάει να αυξάνεται, η εποχή που σταματάει να αυξάνεται, καθώς κι η τελική της τιμή αποτελούν υπερπαραμέτρους του μοντέλου. Ωστόσο στο προτεινόμενο μοντέλο η τελική τιμή είναι $\beta = 1$ για να μην υπάρχουν αποκλίσεις από το μεταβολικό κάτω φράγμα, ενώ χρησιμοποιούμε την ίδια αναλογία με το [4] και κρατάμε την μετάβαση από $\beta = 0$ σε $\beta = 1$ στο πρώτο 10% των εποχών εκπαίδευσης.

Η μη ρύθμιση της β καθώς και η έλλειψη των παραμέτρων γ και δ του CA και του DA κάνουν το προτεινόμενο μοντέλο πολύ πιο απλό σε σχέση με το αρχικό.

3.5.2 Εκπαίδευση του ταξινομητή

Όπως περιγράφηκε στην προηγούμενη ενότητα ο ταξινομητής ταξινομεί σημεία του χώρου κοινής ενσωμάτωσης στις κλάσεις του συνόλου εκπαίδευσης. Συνεπώς για την εκπαίδευση του απαιτείται ένα σύνολο σημείων του χώρου κοινής ενσωμάτωσης μαζί με την κλάση στην οποία ανήκουν. Επιπλέον αυτό το σύνολο σημείων πρέπει να περιέχει σημεία που ανήκουν σε όλες τις κλάσεις.

Καθώς ένα τέτοιο σύνολο σημείων είναι αδύνατο να παρέχεται στο αρχικό σύνολο δεδομένων, πρέπει να κατασκευαστεί. Υποθέτοντας πως η εκπαίδευση του προηγούμενου βήματος έχει δημιουργήσει έναν αρκετά “καλό” χώρο κοινής ενσωμάτωσης, μπορεί να δημιουργηθεί ένα σύνολο δεδομένων εκπαίδευσης για τον ταξινομητή δειγματοληπτώντας σημεία γύρω από τις κωδικοποιήσεις των περιγραφών των κατηγοριών.

Πιο συγκεκριμένα καθώς οι περιγραφές και των γνωστών και των άγνωστων κατηγοριών είναι διαθέσιμες, μπορούν να δειγματοληπτηθούν σημεία από τις κατανομές των κατηγοριών που δίνει ο κωδικοποιητής των περιγραφών. Με αυτά θα εκπαιδευτεί ο ταξινομητής ώστε να μάθει ένα διαχωρισμό του χώρου κοινής ενσωμάτωσης. Έτσι όταν κωδικοποιηθούν CNN χαρακτηριστικά από εικόνες του συνόλου ελέγχου, είτε ανήκουν στις γνωστές είτε στις άγνωστες κλάσεις, θα κωδικοποιηθούν κοντά στις κωδικοποιήσεις των περιγραφών, και θα ταξινομηθούν στις ανάλογες κατηγορίες.

Ο αριθμός των σημείων που δειγματοληπτούνται είναι ίδιος για κατηγορίες από γνωστές κλάσεις, και ίδιος για κατηγορίες από άγνωστες κλάσεις. Η αναλογία αυτών των δύο αποτελεί υπερπαράμετρο που ρυθμίζεται στο επόμενο κεφάλαιο.

4. Πειραματική αξιολόγηση

Στόχος αυτού του κεφαλαίου είναι η αξιολόγηση του προτεινόμενου μοντέλου και η αντιπαράθεση του σε σχέση με το CADA-VAE. Αρχικά αναλύεται η πειρατική διάταξη, δηλαδή σε ποια σύνολα δεδομένων θα γίνει η αξιολόγηση και υπό ποιες υποθέσεις. Έπειτα να βρεθούν οι υπερπαράμετροι που δίνουν την βέλτιστη απόδοση του μοντέλου και με βάση αυτές να γίνουν τα τελικά πειράματα αξιολόγησης. Στην τελευταία ενότητα υπάρχει σχολιασμός για την διαισθητικότερη ερμήνευση των αποτελεσμάτων.

4.1 Μεθοδολογία αξιολόγησης

4.1.1 Μετρικές αξιολόγησης

Ως μετρική αξιολόγησης ενός μοντέλου, όπου μοντέλο εννοείται μια αρχιτεκτονική CSVaDE με συγκεκριμένες υπερπαραμέτρους, χρησιμοποιείται κυρίως ο αρμονικός μέσος H των S (ακρίβεια για τις γνωστές κλάσεις) και U (ακρίβεια για τις άγνωστες κλάσεις), καθώς το σφάλμα του VADE:

- εξαρτάται από την αρχιτεκτονική του μοντέλου, και
- δεν μας δίνει μια άμεση διαίσθησή όσο αφορά την επίδοση στο πρόβλημα του GZSL

Ωστόσο αξίζει να σημειωθεί πως η συνάρτηση αξιολόγησης του VaDE καθώς και το σφάλμα ανακατασκευής θα μπορούσαν να αποτελούν χρήσιμες μετρικές αξιολόγησης σε άλλες προβλήματα, όπως αυτό την κατασκευής συνθετικών εικόνων.

Σε περιπτώσεις ισοθαθμίας της μετρικής H μεταξύ μοντέλων, όσο και στο κεφάλαιο 4.3 που γίνεται η τελική αξιολόγηση του συστήματος, αναφέρονται και οι μετρικές S και U για πληρότητα των αποτελεσμάτων. Καλύτερη επιλογή αποτελούν προφανώς το μοντέλα για τα οποία το S και το H είναι το δυνατό πιο κοντά, καθώς θα είχαν μεγαλύτερη επιτυχία σας ένα ρεαλιστικό σύστημα.

4.1.2 Σύνολα δεδομένων αξιολόγησης

Όπως αναφέρθηκε και στο Κεφάλαιο 2 υπάρχουν 5 γνωστά σύνολα δεδομένων που είναι κατάλληλα για το πρόβλημα του Zero-Shot Learning:

- aPY
- AWA1
- AWA2

- CUB
- SUN

Από αυτά στην παρούσα εργασία χρησιμοποιούνται τα 3 τελευταία καθώς:

- το aPY είναι πολύ μικρό και πλέον δυσεύρετο στην σύγχρονη βιβλιογραφία, ενώ
- το AWA1 έχει αντικατασταθεί από το AWA2 που είναι σχεδόν ίδιο αλλά δεν έχει προβλήματα πνευματικών δικαιωμάτων.

Αν και η τελική αξιολόγηση του μοντέλου θα γίνει και στα 3 σύνολα δεδομένων, η ρύθμιση των παραμέτρων θα γίνει μόνο στο AWA2, καθώς και το [27] χρησιμοποιεί μόνο ένα σύνολο δεδομένων για ρύθμιση. Έτσι τα αποτελέσματα είναι συγκρίνονται πιο δίκαια.

Για την ρύθμιση των παραμέτρων χρειάζεται ένα σύνολο επικύρωσης, που περιέχει κλάσεις ξένες με το σύνολο εκπαίδευσης. Ωστόσο, αφού τα πειράματα είναι πάνω στο γενικότερο GZSL, το μοντέλο πρέπει να αξιολογηθεί και στις γνωστές κλάσεις (αυτές που ανήκουν στο σύνολο εκπαίδευσης). Συνεπώς το σύνολο εκπαίδευσης πρέπει να χωριστεί σε ένα καινούριο σύνολο εκπαίδευσης (80%) και ένα σύνολο ελέγχου γνωστών κλάσεων (20%). Αντίστοιχα για την τελική αξιολόγηση χρειάζεται ένα σύνολο ελέγχου με τις άγνωστες κλάσεις. Ωστόσο θα χρειαστεί και ένα σύνολο ελέγχου από γνωστές. Γ' αυτό ενώνονται τα σύνολα εκπαίδευσης και επικύρωσης σε ένα μεγαλύτερο σύνολο, το 80% του οποίου χρησιμοποιείται για εκπαίδευση και το άλλο 20% για έλεγχο στις γνωστές κλάσεις.

Για τον τελευταίο διαχωρισμό, δηλαδή τον διαχωρισμό του συνόλου εκπαίδευσης και επικύρωσης σε ένα καινούριο σύνολο εκπαίδευσης και ένα σύνολο ελέγχου γνωστών κλάσεων, χρησιμοποιήθηκε ο ίδιος διαχωρισμός με το CADA-VAE σε όλα τα σύνολα δεδομένων. Ωστόσο δε βρέθηκε αντίστοιχος διαχωρισμός για το σύνολο επικύρωσης, οπότε για την ρύθμιση των υπερπαραμέτρων τα δεδομένα χωρίστηκαν τυχαία με αναλογία 80%-20%.

4.2 Διερεύνηση παραμέτρων αρχιτεκτονικής

Σε αυτήν την ενότητα ακολουθείται μια συστηματική μέθοδος ρύθμισης των υπερπαραμέτρων του μοντέλου που αφορούν την αρχιτεκτονική. Πιο συγκεκριμένα η απόδοση του μοντέλου εξαρτάται σε μεγάλο βαθμό από τις ακόλουθες υπερπαραμέτρους:

- τον αριθμό διαστάσεων του χώρου κοινής ενσωμάτωσης d ,
- την αρχιτεκτονική των κωδικοποιητών και αποκωδικοποιητών E_1, E_2, D_1, D_2 ,
- την χρήση Softmax ή SVM στον ταξινομητή.

Για την ορθότερη ρύθμιση θα έπρεπε να εξεταστούν όλες οι παραπάνω παράμετροι ταυτόχρονα, στον ίδιο χώρο αναζήτησης. Καθώς όμως μια τέτοια διερεύνηση δημιουργεί έναν τεράστιο χώρο αναζήτησης, είναι υπολογιστικά ανέφικτη. Έτσι, η ρύθμιση ξεκινάει με τις τιμές που δίνονται στο [27] για τα αντίστοιχα τμήματα του μοντέλου, και κάθε φορά ρυθμίζουμε μια παράμετρο. Στο επόμενο πείραμα χρησιμοποιούμε τις τιμές που βρέθηκαν στα προηγούμενα.

4.2.1 Αριθμός διαστάσεων του χώρου κοινής ενσωμάτωσης

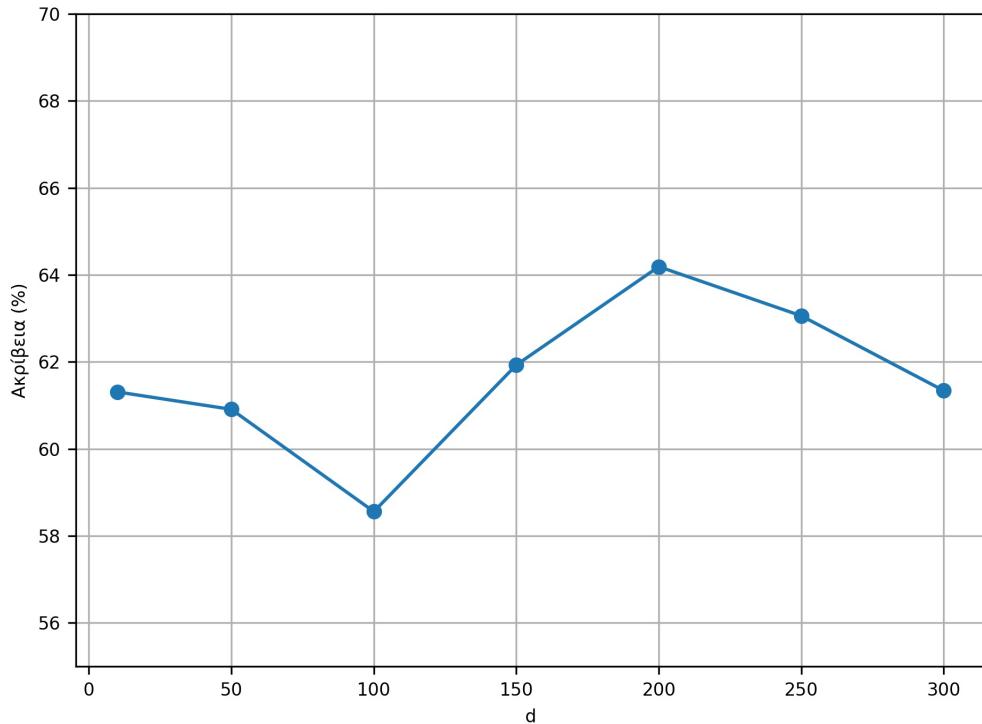
Η πιο βασική παράμετρος που πρέπει να ρυθμιστεί είναι ο αριθμός των διαστάσεων του χώρου κοινής ενσωμάτωσης, d . Δεδομένου του μεγέθους των συνόλων δεδομένων που έχουμε,

εξετάζονται τιμές $10 \leq d \leq 300$, και πιο συγκεκριμένα για να μειώσουμε περαιτέρω τον χώρο αναζήτησης εξετάζονται οι τιμές $d \in \{10, 50, 100, 200, 250, 300\}$.

Για την αρχιτεκτονική των VaDE χρησιμοποιούνται πλήρως συνδεδεμένα (fully connected) νευρωνικά δίκτυα με 1 κρυφό επίπεδο. Οι διαστάσεις του επιπέδου εισόδου και εξόδου δίνονται προφανώς από τις διαστάσεις των οντοτήτων (CNN χαρακτηριστικά και περιγραφές) και την υπερπαράμετρο d αντίστοιχα. Ο αριθμός νευρώνων του κρυφού επιπέδου στα E_1 , E_2 , D_1 , D_2 είναι 1560, 1460, 1660, 665 αντίστοιχα.

Επιπλέον χρησιμοποιείται ταξινομητής Softmax, για την εκπαίδευση του οποίου δημιουργούνται 200 δείγματα για κάθε γνωστή κλάση και 400 για κάθε άγνωστη.

Για την ελαχιστοποίηση της συνάρτησης σφάλματος των VaDE χρησιμοποιείται ο αλγόριθμος Adam με AMSGrad και ρυθμό μάθησης 0.00015. Για τον ταξινομητή χρησιμοποιείται πάλι Adam με τις προκαθορισμένες τιμές της βιβλιοθήκης PyTorch με εξαίρεση τις παραμέτρους beta που έχουν την τιμή $[0.5, 0.999]$.



Σχήμα 4.1: Ρύθμιση του αριθμού διαστάσεων του χώρου κοινής ενσωμάτωσης

Τα αποτελέσματα φαίνονται στο σχήμα 4.1 και η βέλτιστη τιμή για την παράμετρο d είναι η τιμή $d = 200$.

4.2.2 Αρχιτεκτονική των VaDE

Μια εξίσου σημαντική παράμετρος είναι η αρχιτεκτονική των κωδικοποιητών και αποκωδικοποιητών. Όπως ήδη αναφέρθηκε χρησιμοποιούνται πλήρως συνδεδεμένα πολυστρωματικά νευρωνικά δίκτυα για την μοντελοποίηση των κωδικοποιητών και των αποκωδικοποιητών. Στο [27] δοκιμάστηκαν ένα και δύο κρυφά επίπεδα, ενώ στο [13] 3, οπότε τα πειράματα

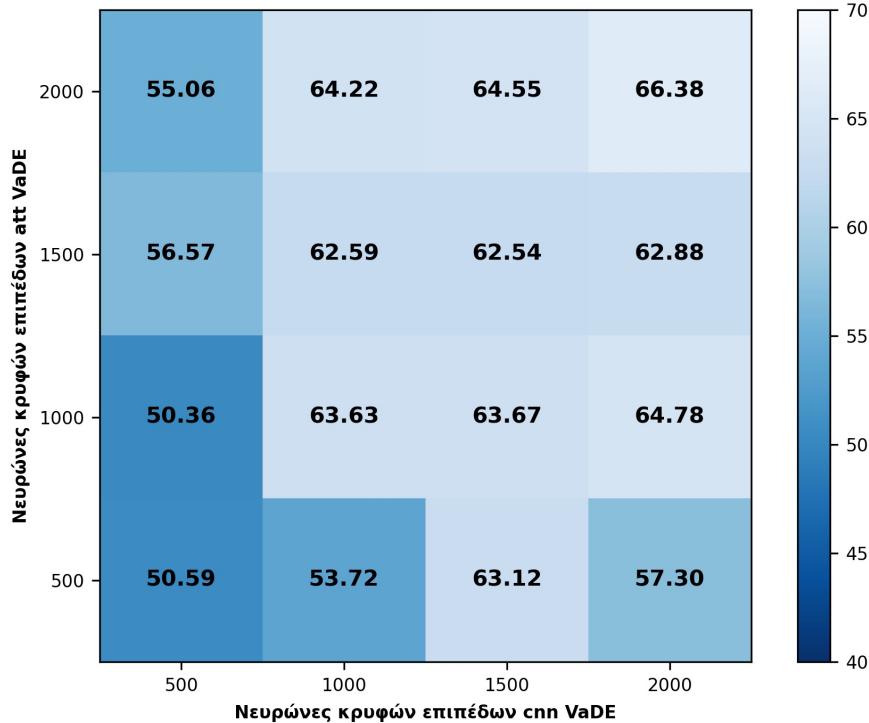
της ενότητας ξεκινούν με ένα κρυφό επίπεδο μέχρι να υπάρξει πιώση στην επίδοση του μοντέλου.

Επίσης για να μειωθεί ο χώρος αναζήτησης χρησιμοποιούμε αντίστροφες αρχιτεκτονικές στον κωδικοποιητή και τον αποκωδικοποιητή, δηλαδή αν ο κωδικοποιητές έχει k κρυφά επίπεδα με v_1, v_2, \dots, v_k νευρώνες τότε ο κωδικοποιητής θα έχει επίσης k κρυφά επίπεδα με v_k, v_{k-1}, \dots, v_1 νευρώνες.

Χρησιμοποιείται χώρος κοινής ενσωμάτωσης με $d = 200$ διαστάσεις ενώ οι υπόλοιπες παράμετροι παραμένουν ίδιες με το πείραμα της προηγούμενης ενότητας.

Αρχιτεκτονικές CNN με 1 επίπεδο και att με 1 επίπεδο

Αρχικά δοκιμάζονται αρχιτεκτονικές με ένα κρυφό επίπεδο τόσο για τους VaDE των CNN χαρακτηριστικών όσο και για τους VaDE των περιγραφών. Και για τις 2 κατηγορίες εξετάζονται 500, 1000, 1500 και 2000 κρυφοί νευρώνες οδηγώντας σε 16 πειράματα που φαίνονται στο σχήμα 4.2. Για νευρώνες με μέγεθος μεγαλύτερο από 2000 δε παρουσιάστηκε κάποια βελτίωση στην επίδοση. Από τα αποτελέσματα μπορούμε να καταλάβουμε πως για επίπεδα με τουλάχιστον 1000 νευρώνες δεν υπάρχει ουσιαστική διαφορά. Καλύτερο αποτέλεσμα ωστόσο δίνει η αρχιτεκτονική με 2000 νευρώνες και στα 4 νευρωνικά.

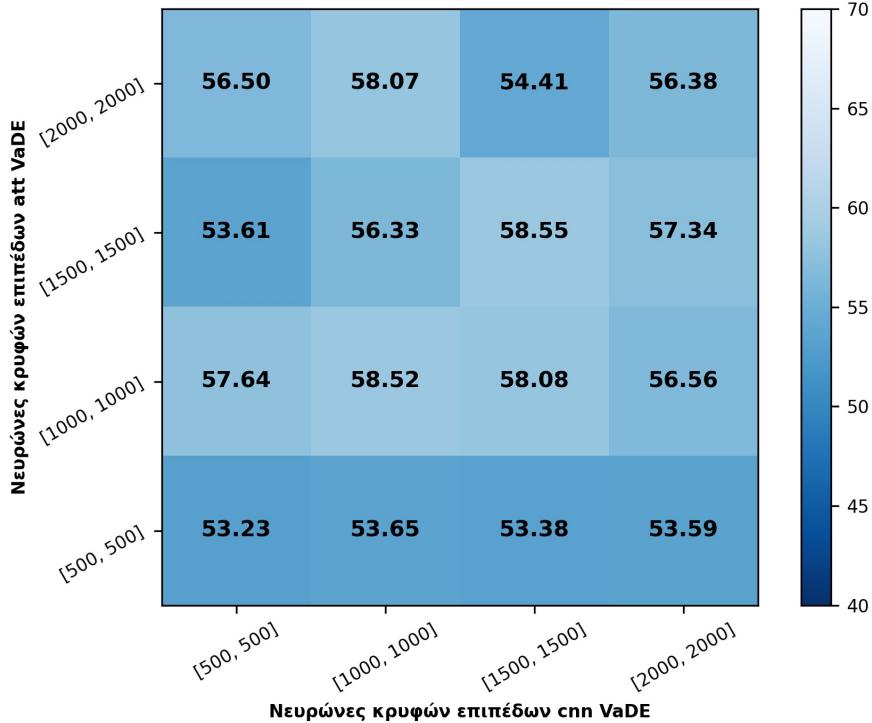


Σχήμα 4.2: Ακρίβεια H για μεταβλητό αριθμό νευρώνων στο κρυφό επίπεδο

Αρχιτεκτονικές CNN με 2 επίπεδα και att με 2 επίπεδα

Έπειτα δοκιμάζονται αρχιτεκτονικές που αποτελούνται από 2 επίπεδα τόσο στον VaDE των CNN χαρακτηριστικών όσο και στον VaDE των περιγραφών. Δοκιμάζονται αρχικά αρχιτεκτονικές με ίδιο αριθμό νευρώνων και στα 2 επίπεδα, με αυξήσεις ανά 500 νευρώνες όπως και

προηγουμένως.



Σχήμα 4.3: Ακρίβεια H για μεταβλητό αριθμό νευρώνων στα κρυφά επίπεδα

Παρατηρείται σχεδόν ίδια επίδοση σε κάθε συνδυασμό, που είναι περίπου 10% χειρότερη σε σχέση με τα πειράματα ενός κρυφού στρώματος. Έτσι δεν γίνεται επιπλέον πειράματα για 3 κρυφά επίπεδα ή για 2 κρυφά επίπεδα με διαφορετικό αριθμό νευρώνων σε κάθε επίπεδο.

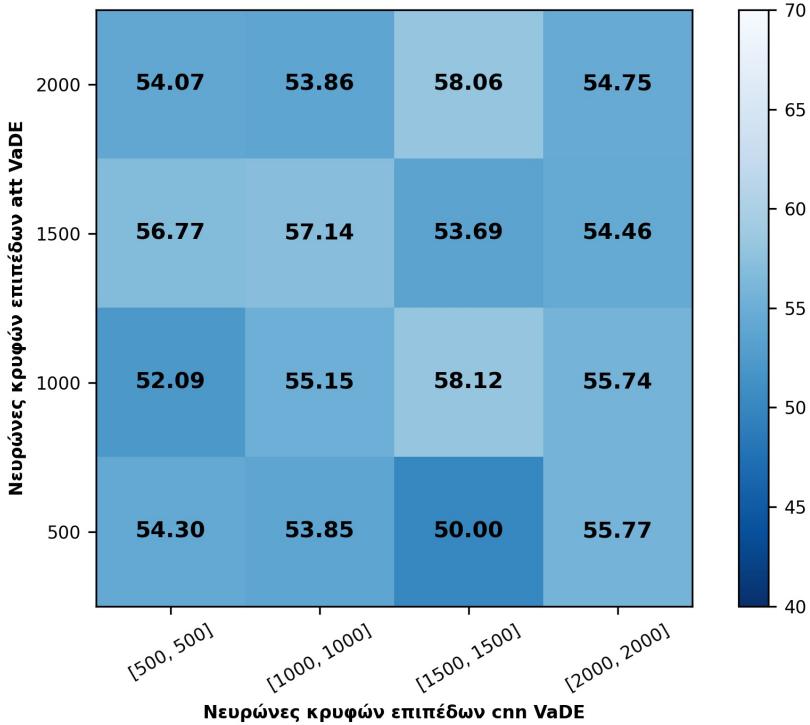
Αρχιτεκτονικές cnn με 2 επίπεδα και att με 1 επίπεδο

Ένα πείραμα που αξίζει να γίνει ωστόσο είναι να κρατήσουμε δύο κρυφά επίπεδα στον VaDE των CNN χαρακτηριστικών και ένα στον VaDE των περιγραφών. Αυτή η επιλογή αιτιολογείται από το μεγαλύτερο εύρος εισόδων, καθώς τα CNN χαρακτηριστικά διαφέρουν όλα μεταξύ τους και ανέρχονται σε χιλιάδες, ενώ οι κατηγορίες είναι μόλις δεκάδες ή εκατοντάδες.

Από τις εικόνες 4.2 - 4.4 είναι ξεκάθαρο η καλύτερη επίδοση επιτυγχάνεται για ένα κρυφό επίπεδο.

Συμπερασματικά η πλέον αποτελεσματική αρχιτεκτονική είναι η :

- $\text{hidden}(E_1) = [2000]$
- $\text{hidden}(D_2) = [2000]$
- $\text{hidden}(E_1) = [2000]$
- $\text{hidden}(D_2) = [2000]$



Σχήμα 4.4: Ακρίβεια H για μεταβλητό αριθμό νευρώνων στα κρυφά επίπεδα

4.2.3 Μοντέλο ταξινομητή

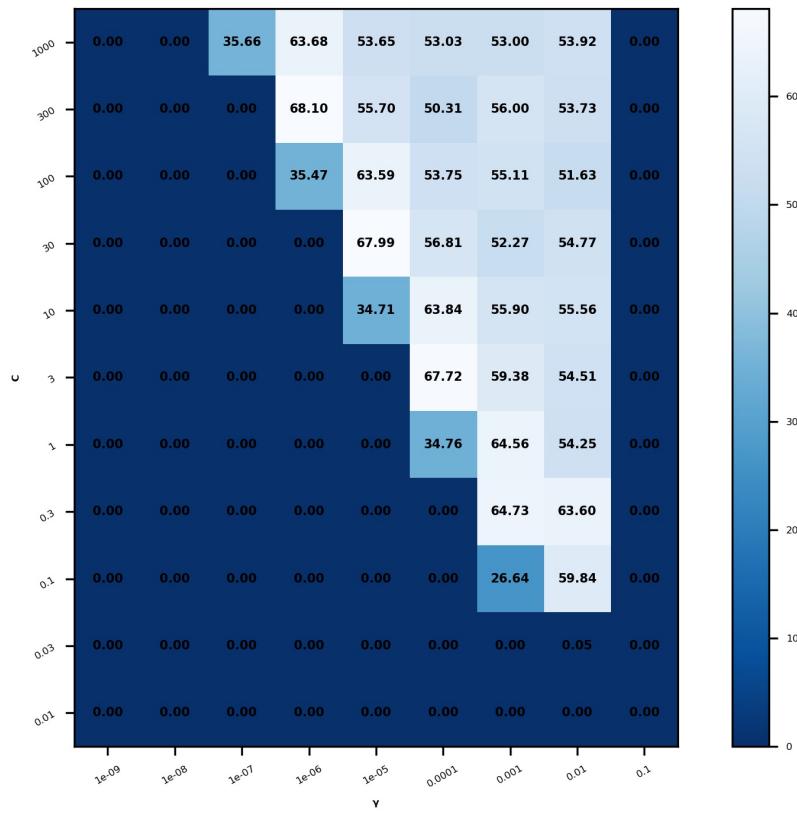
Φεύγοντας από το κομμάτι των VaDE, θα εξεταστεί αν η χρήση SVM οδηγεί σε καλύτερα αποτελέσματα σε σύγκριση με την Softmax. Για την υλοποίηση του SVM ταξινομητή χρησιμοποιείται η βιβλιοθήκη scikit-learn και έτσι οι παράμετροι του αλγορίθμου βελτιστοποίησης δεν λαμβάνονται υπόψιν. Ωστόσο υπάρχουν 2 σημαντικές παράμετροι του μοντέλου για τις οποίες γίνεται διεξοδική διερεύνηση: η C και η γ .

Στο σχήμα 4.5 δοκιμάζονται τιμές για την παράμετρο C από 0.01 έως 1000 ενώ για την παράμετρο γ από 10^{-9} μέχρι 0.1. Η αύξηση των τιμών στον χώρο αναζήτησης είναι εκθετική, και όπως βλέπουμε υπάρχουν αρκετές δυάδες παραμέτρων που συναγωνίζονται για την καλύτερη επίδοση. Καθώς έχουν την ίδια επίδοση, θα εξεταστούν και οι μέσες ακρίβειες πάνω στις γνωστές και άγνωστες κλάσεις.

(C, γ)	S	U	H
(300, 1e-6)	67.6	68.6	68.1
(30, 1e-5)	67.4	68.1	67.7
(3, 1e-4)	67.6	68.4	68

Πίνακας 4.1: Ακρίβειες S , U και H για τις υποψίες υπερπαραμέτρους

Σύμφωνα με τον πίνακα 4.1 όλες οι τιμές δίνουν ίδια ακρίβεια στις γνωστές και άγνωστες κλάσεις, που είναι ιδανικό δεδομένης μια αρμονικής ακρίβειας H . Συνεπώς οποιαδήποτε τιμή από αυτές θα μπορούσε να κρατηθεί. Ωστόσο εφόσον ο SVM ταξινομητής δεν έχει πολύ καλύτερη επίδοση από τον Softmax, στα πειράματα της επόμενης ενότητας εξετάζουμε και τους 2.



Σχήμα 4.5: Ρύθμιση των υπερπαραμέτρων C και γ του ταξινομητή SVM

4.3 Διερεύνηση παραμέτρων εκπαίδευσης

Πέρα από τις υπερπαραμέτρους που ορίζουν την αρχιτεκτονική του μοντέλου, σημαντικό ρόλο στην επίδοση του παίζουν και αυτές που χρησιμοποιήθηκαν για την εκπαίδευση του. Πιο συγκεκριμένα πρέπει να εξεταστούν:

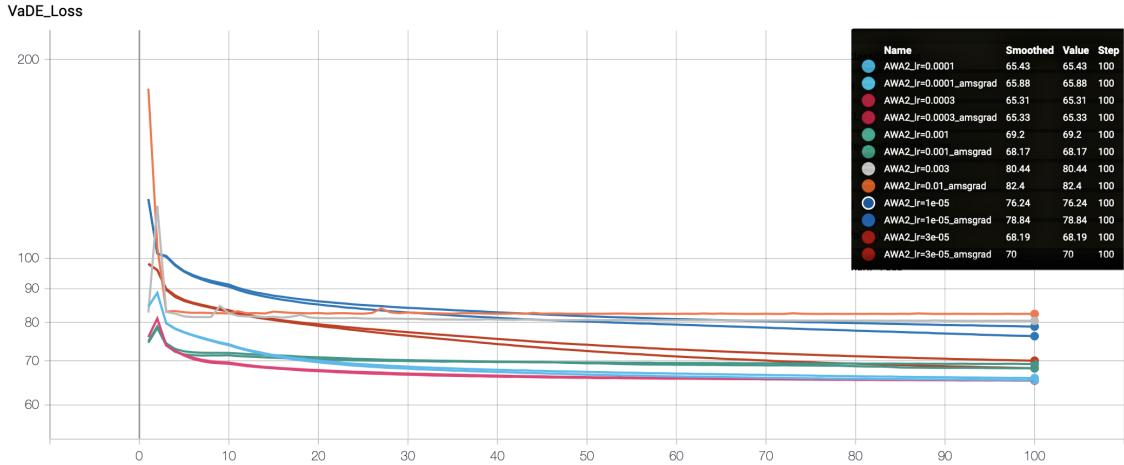
- οι παράμετροι του αλγορίθμου βελτιστοποίησης Adam που χρησιμοποιείται για την ελαχιστοποίηση του σφάλματος στους VaDE.
- η αναλογία δειγμάτων εκπαίδευσης από γνωστές και άγνωστες κλάσεις για την εκπαίδευση του ταξινομητή.

Όμοια με πριν, και ως συνέχεια των πειραμάτων της προηγούμενης ενότητας, χρησιμοποιούνται οι παράμετροι του μοντέλου CADA-VAE ως προεπιλεγμένες παράμετροι με εξαίρεση τις υπερπαραμέτρους που έχουν ήδη ρυθμιστεί.

4.3.1 Παράμετροι του αλγορίθμου Adam

Αν και σε αυτήν την υποενότητα θα μπορούσαν να εξεταστούν όλοι οι παράμετροι του Adam, καθώς και άλλοι γνωστοί αλγόριθμοι βελτιστοποίησης, εξετάζονται μόνο ο ρυθμός μάθησης και η χρήση ή όχι της παραλλαγής AMSGrad του αλγορίθμου.

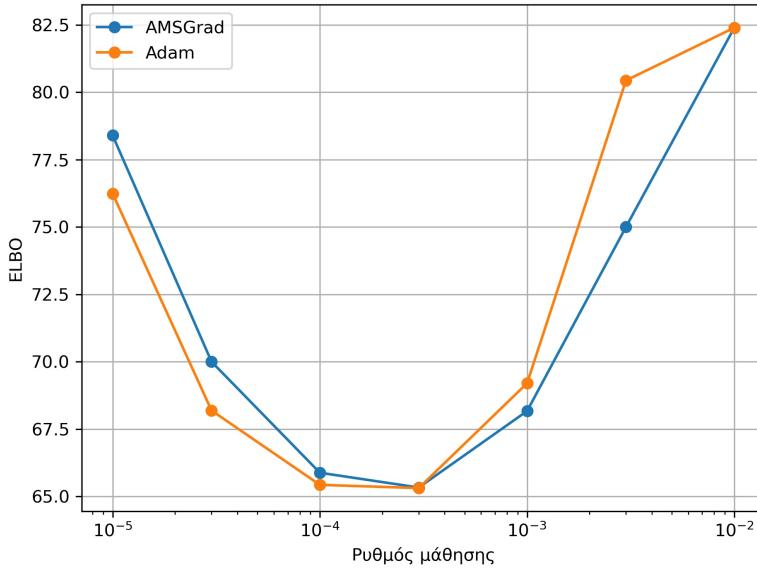
Πιο συγκεκριμένα εξετάζονται 9 ρυθμοί μάθησης από 10^{-5} έως 10^{-1} , και για κάθε ρυθμό μάθησης οι αλγόριθμοι AMSGrad και Adam.



Σχήμα 4.6: Σφάλμα εκπαίδευσης VaDE των πρώτων 100 εποχών για διάφορες τιμές παραμέτρων του Adam

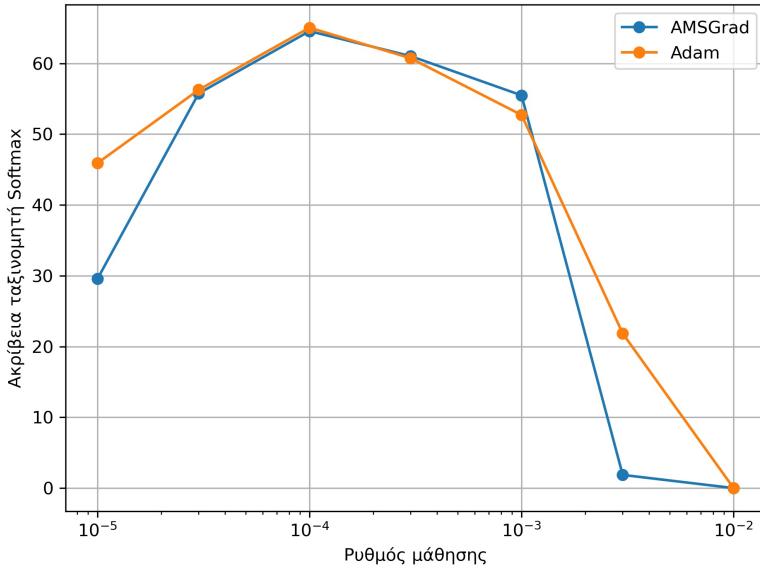
Το σχήμα 4.6 δείχνει την μείωση του σφάλματος εκπαίδευσης των VaDE για 100 εποχές για κάθε έναν από τους 14 συνδυασμούς που αναφέρθηκαν. Βλέπουμε πως οι ρυθμοί μάθησης που επιτυγχάνουν την καλύτερη βελτιστοποίηση είναι οι lr=0.0001 και lr=0.0003 ανεξάρτητα από την χρήση ή όχι του αλγορίθμου AMSGrad.

Για να γίνει πιο σαφή σύγκριση μεταξύ των 4 αυτών τιμών, παρακάτω δίνονται τα διαγράμματα της τελικής τιμής ELBO (δηλαδή μετά από 100 εποχές εκπαίδευσης), της ακρίβειας ενός ταξινομητή Softmax και ενός ταξινομητή SVM. Για την Softmax και τον SVM χρησιμοποιούμε τις τιμές που βρέθηκαν στις προηγούμενες ενότητες και για την δημιουργία του συνόλου εκπαίδευσης του ταξινομητή 200 δείγματα ανά γνωστή κλάση και 400 ανά άγνωστη.



Σχήμα 4.7: Σφάλμα εκπαίδευσης μετά από 100 εποχές για διάφορες τιμές παραμέτρων του Adam

Βλέπουμε πως έχουμε ελαφρώς καλύτερα αποτελέσματα για lr=0.0001 χωρίς την χρήση AMSGrad, οπότε από εδώ και πέρα θα χρησιμοποιείται αυτός ο αλγόριθμος βελτιστοποίησης



Σχήμα 4.8: Ακρίβεια ταξινομητή Softmax έπειτα από εκπαίδευση του χώρου κοινής ενσωμάτωσης με τον Adam

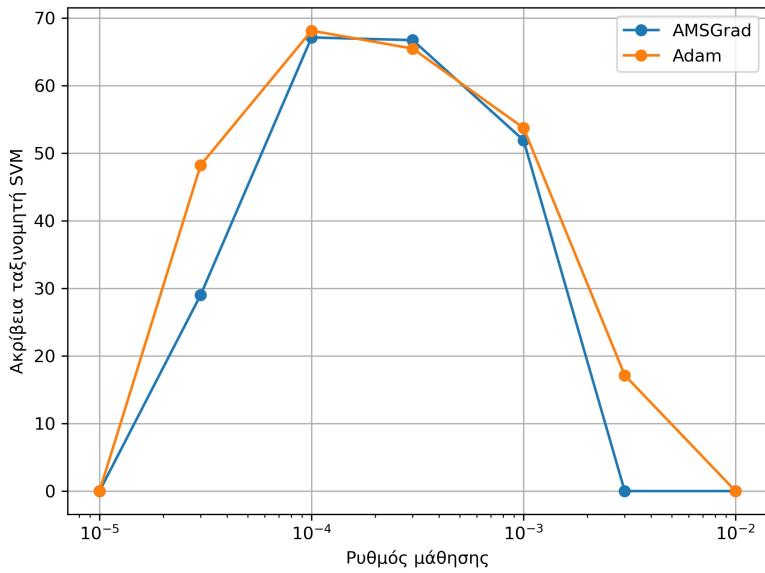
των υπερπαραμέτρων.

4.3.2 Αναλογία γνωστών και άγνωστων κλάσεων στην εκπαίδευση του ταξινομητή

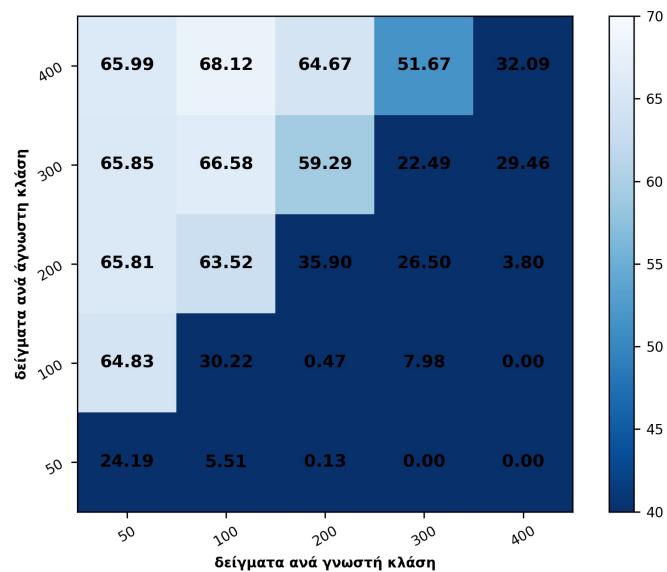
Τέλος θα εξετάσουμε τον αριθμό των δειγμάτων από γνωστές και άγνωστες κλάσεις κατά την διάρκεια της εκπαίδευσης του ταξινομητή. Δοκιμάζονται όλοι οι δυνατοί συνδυασμοί μεταξύ 50, 100, 200, 300 και 400 δειγμάτων, καθώς με αυτές τις τιμές καλύπτονται αναλογίες μέχρι και 8:1, προσφέροντας μεγαλύτερο εύρος του πλήρους χώρου αναζήτησης σε σχέση με το [27].

Από τα σχήματα 4.10 και 4.11 παρατηρείται πως η αναλογία γνωστών και άγνωστων κλάσεων δεν επηρεάζει τον ταξινομητή SVM ωστόσο επηρεάζει αρκετά τον ταξινομητή Softmax. Πιο συγκεκριμένα παρατηρούμε ικανοποιητική απόδοση για αριθμό δειγμάτων άγνωστων κλάσεων τουλάχιστον διπλάσιο από τον αριθμό δειγμάτων των γνωστών κλάσεων, ενώ πολύ χαμηλή απόδοση σε αντίθετη περίπτωση.

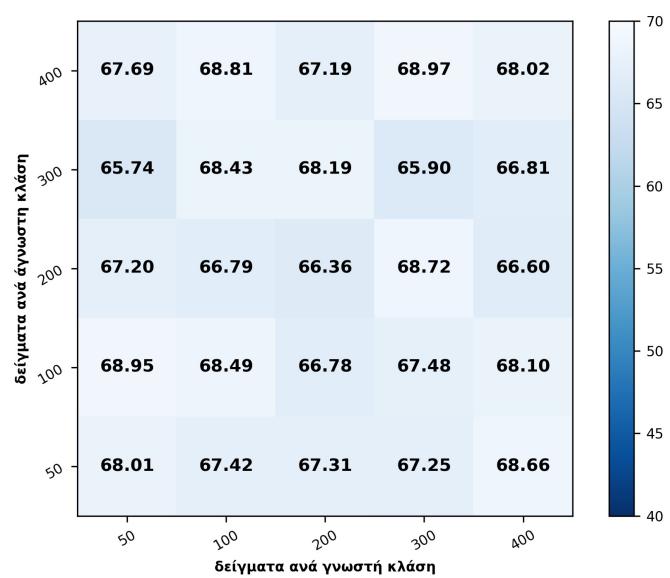
Η καλύτερη αναλογία για τον ταξινομητή Softmax είναι 100 δείγματα ανά γνωστή κλάση και 400 ανά άγνωστη, που δίνει και μια από τις καλύτερες επιδόσεις του SVM. Συνεπώς στα πειράματα τις επόμενες ενότητας επιλέγουμε αυτές τις τιμές.



Σχήμα 4.9: Ακρίβεια ταξινομητή SVM έπειτα από εκπαίδευση του χώρου κοινής ενσωμάτωσης με τον Adam



Σχήμα 4.10: Ακρίβεια ταξινομητή Softmax για διάφορες τιμές δειγμάτων ανά γνωστή και άγνωστη κλάση



Σχήμα 4.11: Ακρίβεια ταξινομητή SVM για διάφορες τιμές δειγμάτων ανά γνωστή και άγνωστη κλάση

4.4 Αποτελέσματα

Μετά τα πειράματα των προηγούμενων δύο ενοτήτων έχουμε καταλήξεις στις υπερπαραμέτρους:

- $d = 200$
- $(E_1, D_1) = ([2000], [2000])$
- $(E_2, D_2) = ([2000], [2000])$
- Ταξινομητής Softmax
- Χρήση του αλγόριθμου βελτιστοποίησης Adam με lr = 0.0001 για 100 εποχές
- 100 δείγματα ανά γνωστή κλάση και 400 ανά άγνωστη

Χρησιμοποιείται ταξινομητής Softmax αντί του SVM, γιατί είναι πιο γρήγορος σε σύνολα δεδομένων με περισσότερες κατηγορίες όπως το CUB και το SUN, ενώ τελικά για κατάλληλο αριθμό γνωστών και άγνωστων κλάσεων έχει την ίδια επίδοση με τον SVM.

Με βάση το παραπάνω μοντέλο CSVaDE θα εξεταστεί η επίδοση της προτεινόμενης μεθοδολογίας σε όλα τα σετ δεδομένων. Για την εκπαίδευση χρησιμοποιείται το 80% του συνόλου εκπαίδευσης και επικύρωσης, ενώ για την αξιολόγηση των γνωστών κλάσεων το υπόλοιπο 20%. Για τις άγνωστες κλάσεις χρησιμοποιείται το σύνολο ελεγχου. Ο διαμερισμός είναι ίδιος με το [27].

4.4.1 Κατηγοριοποίησης εικόνων χωρίς καθόλου παραδείγματα

Το πρώτο σενάριο για το οποίο γίνεται έλεγχος του προτεινόμενου μοντέλου είναι αυτός της κατηγοριοποίησης χωρίς καθόλου παραδείγματα (GZSL). Εξετάζουμε την απόδοση των μοντέλων στα σύνολα δεδομένων AWA2, CUB, SUN και για top-1 και top-3 ακρίβεια.

	top-1			top-3		
	S	U	H	S	U	H
AWA2	76.26	42.93	54.93	89.18	86.40	87.77
CUB	29.55	47.57	36.46	57.56	71.65	63.84
SUN	6.82	45.21	11.85	21.67	69.86	33.08

Πίνακας 4.2: Αποτελέσματα GZSL σε όλα τα σύνολα δεδομένων

4.4.2 Κατηγοριοποίησης εικόνων με λίγα παραδείγματα

Το επόμενο και τελευταίο σενάριο είναι αυτός της κατηγοριοποίησης με λίγα παραδείγματα. Τυπικά στην βιβλιογραφία εξετάζονται 1-20 επιπλέον παραδείγματα ανά άγνωστη κατηγορία. Συνεπώς σε αυτό το πείραμα, κατά την διάρκεια της εκπαίδευσης, το μοντέλο εκπαιδεύεται με 1, 2, 5 και 10 δείγματα από κάθε άγνωστη κατηγορία. Η ακρίβεια είναι top-1. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα.

	1-shot	2-shot	5-shot	10-shot
AWA2	60.79	67.06	74.53	75.07
CUB	36.39	37.14	38.40	38.88
SUN	13.37	13.22	14.32	16.76

Πίνακας 4.3: Αποτελέσματα GFSL σε όλα τα σύνολα δεδομένων

4.5 Σχολιασμός αποτελεσμάτων

Παρατηρούμε πως στο σύνολο δεδομένων AWA2 έχουμε την καλύτερη επίδοση, συγκρίσιμη με τις πλέον σύγχρονες και επιτυχημένες μεθόδους του Πίνακα 2.2. Στα σύνολα CUB και SUN έχουμε αρκετά χαμηλές επιδόσεις, πιθανότατα επειδή χρησιμοποιήθηκαν οι ίδιες υπερπαράμετροι που ρυθμίστηκαν στο AWA2. Αυτή η εικασία ενισχύεται από το γεγονός πως η ακρίβεια στις άγνωστες κλάσεις είναι πιο υψηλή από κάθε σύγχρονη μέθοδο, αλλά η ακρίβεια στις γνωστές κλάσεις πολύ χαμηλή. Συνεπώς χρειάζεται εκ νέου ρύθμιση των παραμέτρων πάνω σε αυτά τα σύνολα.

Επιπλέον παρατηρούμε πως ακόμα και στο σύνολο δεδομένων AWA2 που έχουμε πολύ καλή επίδοση, υπάρχει μια αξιοσημείωτη διαφορά ακρίβειας μεταξύ γνωστών και άγνωστων κλάσεων, καθώς και μια διαφορά της τάξης του 13% σε σχέση με το σύνολο επικύρωσης. Αυτό σημαίνει πως χρειάζεται καλύτερη ρύθμιση των υπερπαραμέτρων, ίσως χρησιμοποιώντας διαφορετικούς διαχωρισμούς του συνόλου δεδομένων κατά την επικύρωση, πολλαπλά πειράματα για κάθε συνδυασμό υπερπαραμέτρων ή ακόμα και συνδυασμούς υπερπαραμέτρων στα πειράματα αντί της σειριακής τους εξέτασης.

Στα πειράματα για κατηγοριοποίηση εικόνων με λίγα παραδείγματα, στο σύνολο AWA2 παρατηρούμε σημαντική αύξηση της ακρίβειας, χρησιμοποιώντας μόνο μερικά επιπλέον παραδείγματα. Αυτό δηλώνει πως η προτεινόμενη μέθοδος κλιμακώνει καλά και στο σενάριο του GFSL, παράγοντας αποτελέσματα συγκρίσιμα με το CADA-VAE.

Βέβαια για την περαιτέρω βελτίωση της προτεινόμενης μεθοδολογίας χρειάζεται ένας ποιοτικότερος σχολιασμός των αποτελεσμάτων καθώς και μια επιθεώρηση του χώρου κοινής ενσωμάτωσης.

Αρχικά παρατηρούμε πως η top-3 ακρίβεια βελτιώνει σημαντικά τα αποτελέσματα, ειδικά στις άγνωστες κλάσεις. Αυτό σημαίνει πως η προτεινόμενη μεθοδολογία, ακόμα και στις περιπτώσεις που δεν επιλέγει την σωστή κατηγορία, την θεωρεί αρκετά πιθανή ώστε να την έχει στις 3 πιο πιθανές επιλογές.

Κάτι ακόμα που θα πρέπει να σημειωθεί είναι πως τα μοντέλο "μπερδεύει" παρόμοιες κατηγορίες μεταξύ τους. Αυτό φαίνεται πιο εύκολα βλέποντας τον πίνακα σύγχυσης των άγνωστων κλάσεων του σχήματος 4.12. Για παράδειγμα το 40% περίπου από τις καμηλοπαρδάλεις ταξινομείται σωστά ως καμηλοπάρδαλη ενώ ένα άλλο 35% ταξινομείται ως άλογο, που έχει κοινά χαρακτηριστικά με την καμηλοπάρδαλη. Ένα 5% ταξινομείται ως αγριόγατες και το υπόλοιπο 20% πάει σε κατηγορίες από γνωστές κλάσεις. Ακόμα και τα δελφίνια, που ταξινομούνται 100% λάθος, ταξινομούνται κυρίως ως μπλε φάλαινες. Αξίζει να σημειωθεί πως τα δελφίνια θεωρούνται κατηγορία φαλαινών και μοιάζουν αρκετά εμφανισιακά με μπλε φάλαινες, με εξαίρεση το μέγεθος φυσικά.

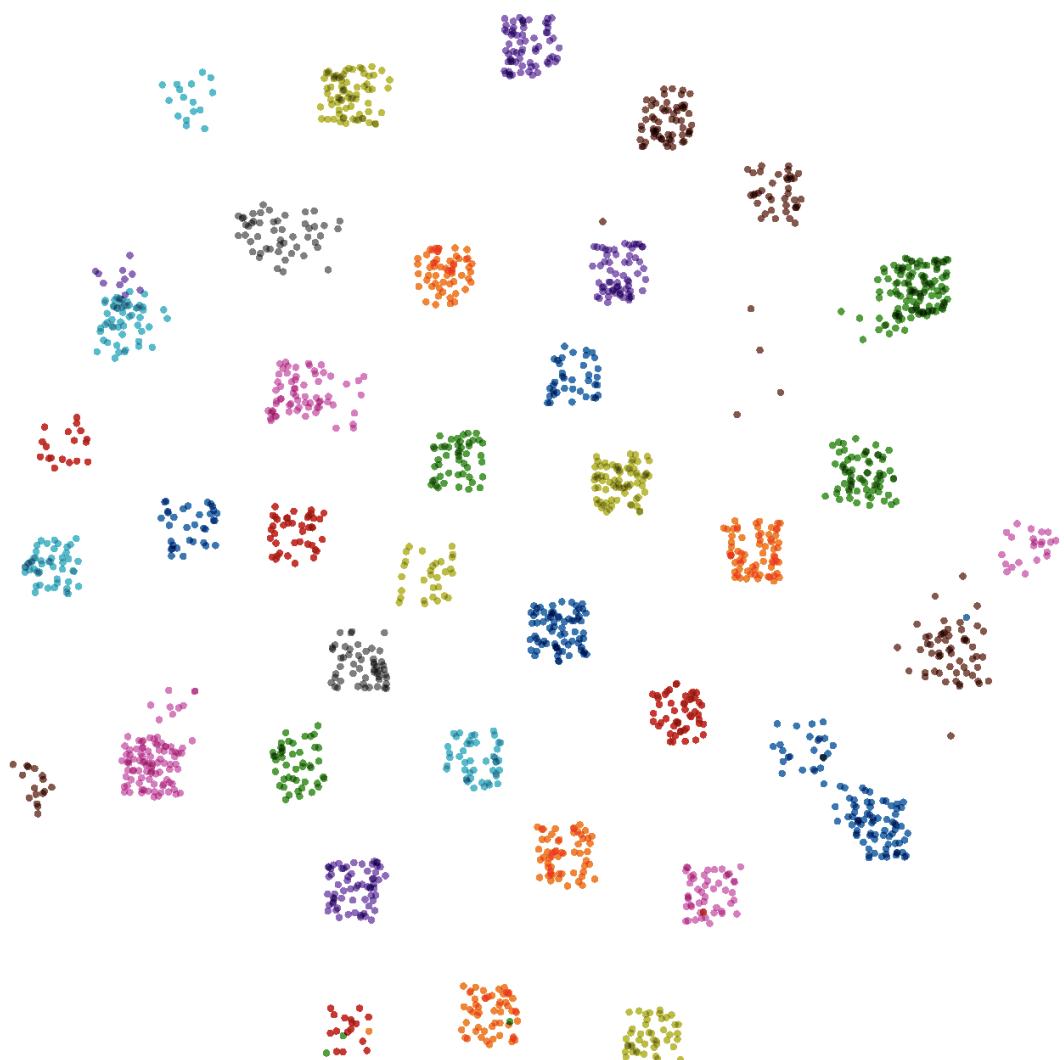
Τέλος δίνονται εικόνες από τον χώρο κοινής ενσωμάτωσης, με την χρήση του αλγόριθμου t-SNE για την μείωση των διαστάσεων από 200 σε 2. Παρατηρούμε πως τα δεδομένα του

	horse	blue whale	sheep	seal	bat	giraffe	rat	bobcat	walrus	dolphin
horse	7.11	0.00	69.18	0.00	0.00	0.36	0.06	0.12	0.00	0.00
blue whale	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sheep	3.17	0.00	74.58	0.00	0.00	0.07	1.34	0.07	0.21	0.00
seal	0.30	4.86	0.00	8.81	0.30	0.10	8.70	0.30	31.78	0.51
bat	2.35	4.70	2.61	0.52	2.35	0.26	47.00	0.78	0.00	0.00
giraffe	34.78	0.00	1.83	0.00	0.00	38.44	0.00	5.24	0.00	0.00
rat	0.00	0.00	0.00	0.00	0.00	0.00	20.97	0.00	0.00	0.00
bobcat	0.00	0.00	0.48	0.00	0.16	0.00	3.02	71.11	0.00	0.00
walrus	1.40	9.77	1.86	6.05	0.00	0.00	5.12	0.00	35.81	0.93
dolphin	0.00	93.34	0.00	3.38	0.00	0.00	0.11	0.00	2.64	0.11

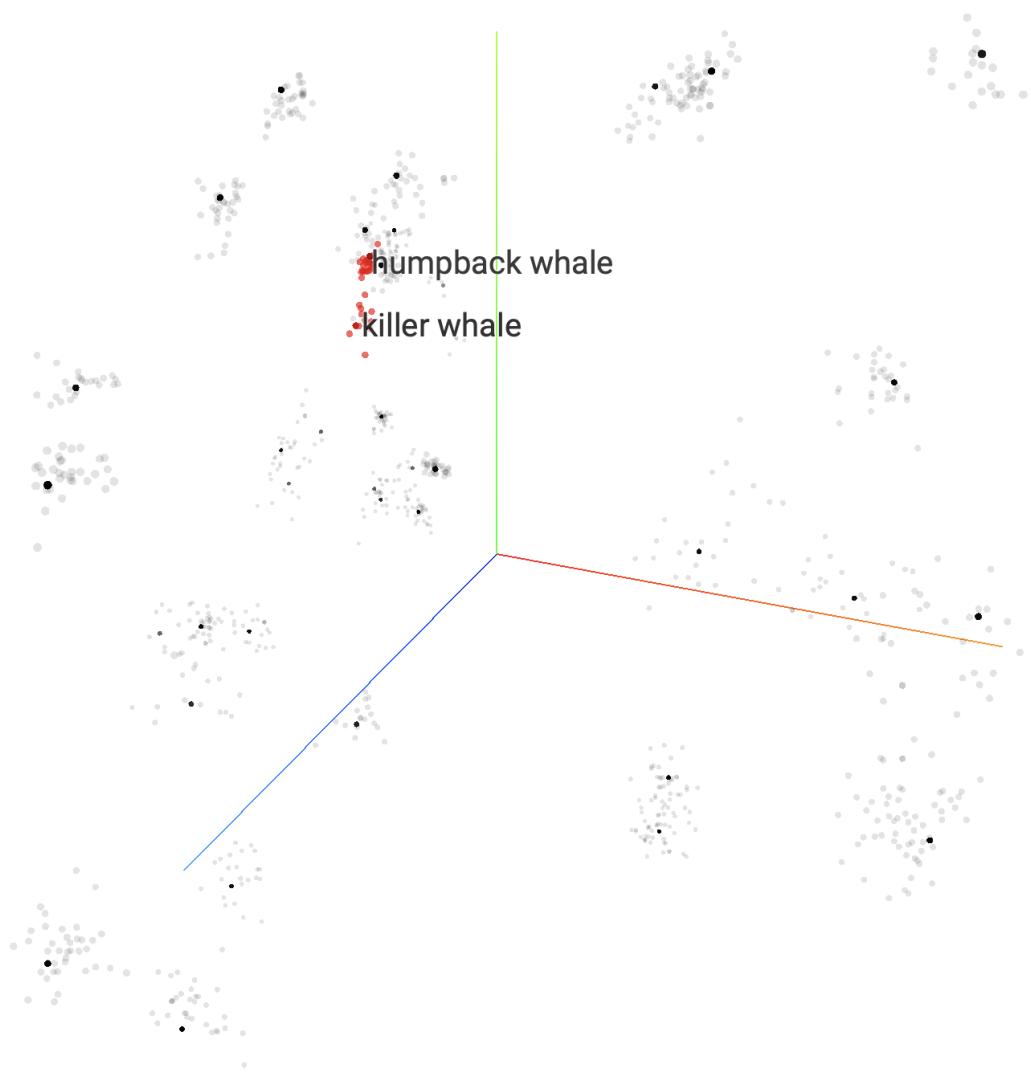
Σχήμα 4.12: Πίνακας σύγχυσης για τις άγνωστες κλάσεις

συνόλου εκπαίδευσης ομαδοποιούνται πολύ καλά σε συστάδες και είναι καλά διαχωρίσιμα. Επιπλέον με την χρήση ενός γραμμικού μετασχηματισμού όπως ο PCA φαίνεται πως οι σημασιολογικά κοντινές κλάσεις είναι πιο κοντά μεταξύ τους, και πως τα σημεία του χώρου που αντιστοιχούν σε CNN χαρακτηριστικά περιτριγυρίζουν τα σημεία που αντιστοιχούν σε περιγραφές, όπως ακριβώς περιγράφηκε θεωρητικά στο Κεφάλαιο 3.

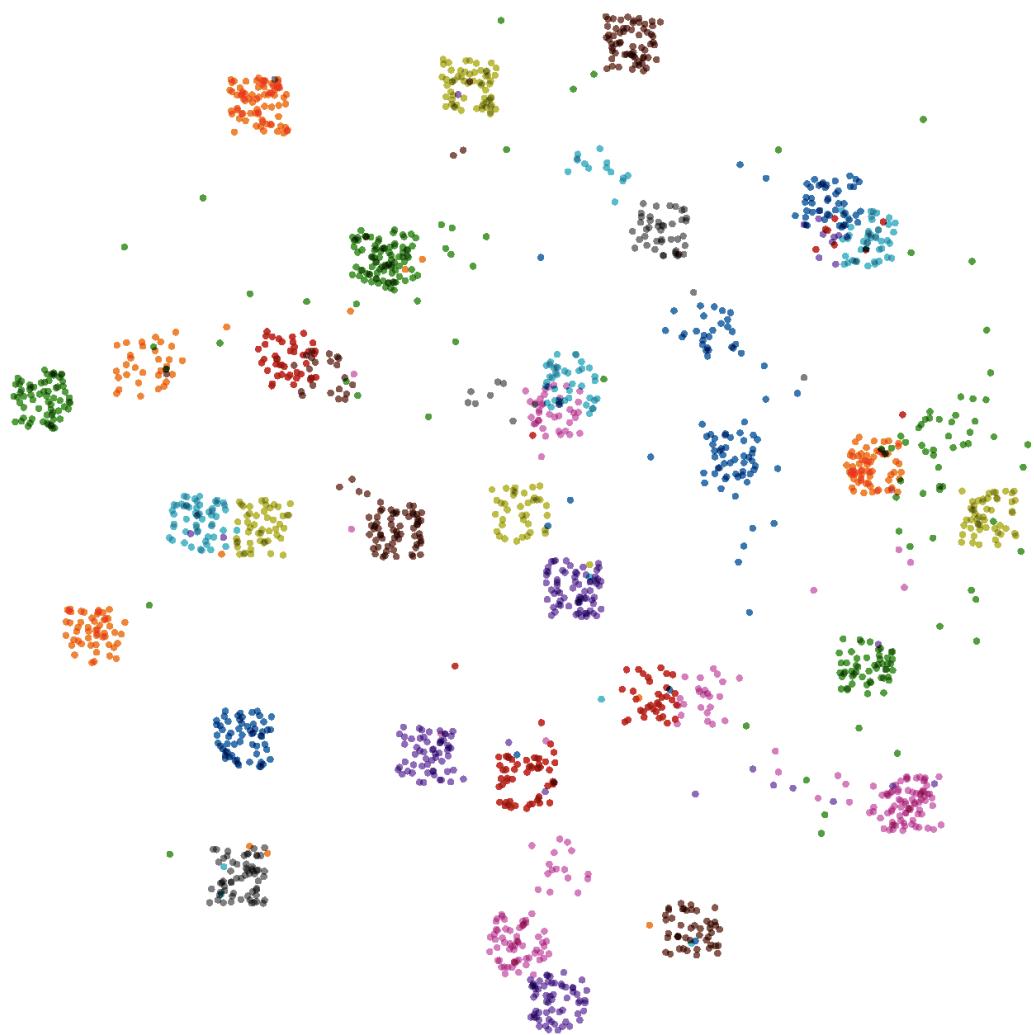
Τα δείγματα γνωστών κλάσεων που χρησιμοποιούνται για έλεγχο φαίνονται στο Σχήμα 4.15 και είναι επίσης αρκετά καλά διαχωρίσιμα παρά την χαμηλή διαστασημότητα σε σχέση με την αρχική. Τέλος όσο αφορά τις άγνωστες κλάσεις, παρατηρείται ομαδοποίηση, ωστόσο οι τα CNN χαρακτηριστικά ομαδοποιούνται μεταξύ τους και οι περιγραφές ομαδοποιούνται μεταξύ τους. Αυτό είναι ισχυρή ένδειξη, πως για την βελτίωση του μοντέλου πρέπει να βρεθεί ένας μηχανισμός καλύτερου συγχρονισμού των δύο VaDE.



Σχήμα 4.13: Μείωση διαστάσεων με τον αλγόριθμο t-SNE του χώρου κοινής ενσωμάτωσης με δείγματα γνωστών κλάσεων κατά την διάρκεια της εκπαίδευσης



Σχήμα 4.14: Μείωση διαστάσεων με τον αλγόριθμο PCA του χώρου κοινής ενσωμάτωσης με δείγματα γνωστών κλάσεων κατά την διάρκεια της εκπαίδευσης



Σχήμα 4.15: Μείωση διαστάσεων με τον αλγόριθμο t-SNE του χώρου κοινής ενσωμάτωσης με δείγματα γνωστών κλάσεων κατά την διάρκεια του ελέγχου



Σχήμα 4.16: Μείωση διαστάσεων με τον αλγόριθμο t-SNE του χώρου κοινής ενσωμάτωσης με δείγματα άγνωστων κλάσεων κατά την διάρκεια του ελέγχου

5. Συμπεράσματα και μελλοντικές προ-εκτάσεις

Στην παρούσα διπλωματική εργασία εξετάστηκε το πρόβλημα της κατηγοριοποίησης με λίγα ή καθόλου παραδείγματα. Αρχικά έγινε μια σύντομη αλλά περιεκτική βιβλιογραφική επισκόπηση των τελευταίων 10 χρόνων στον τομέα. Έπειτα προτάθηκε μια καινούρια μεθοδολογία αντιμετώπισης του προβλήματος, το CSVaDE, που συνδυάζει την έξυπνη ιδέα του συγχρονισμού δύο αυτοκωδικοποιητών αλλά κρατάει την συνάρτηση σφάλματος απλή χωρίς να προσθέτει αλληλοαντικρουσόμενους όρους. Αυτή η μεθοδολογία αναλύθηκε, τόσα θεωρητικά όσο και πειραματικά, ώστε να σχηματίζει ένα χώρο κοινής ενσωμάτωσης όπου τα δείγματα εικόνων θα ομαδοποιούνται με βάση την κατηγορία τους και οι περιγραφές θα βρίσκονται στο κέντρο της κάθε ομάδας.

Διεξάχθηκαν πειράματα που δείχνουν ότι η μέθοδος έχει πολύ υποσχόμενα αποτελέσματα, αφού στο σύνολο δεδομένων AWA2 επιτεύχθηκε επίδοση συγκρίσιμη με τις πλέον σύγχρονες μεθόδους. Επιπλέον, μεταφέρθηκαν δείγματα από το σύνολο ελέγχου στο σύνολο εκπαίδευσης και το CSVaDE απέδωσε πολύ ικανοποιητικά στο γενικότερο πρόβλημα του GFSL. Όσο αφορά τις αστοχίες του μοντέλου, οι λανθασμένες προβλέψεις είναι πολύ κοντά σημασιολογικά ή οπτικά στις πραγματικές κατηγορίες των εικόνων (για παράδειγμα τα δελφίνια μπερδεύονται με τις φάλαινες).

Μια από τις πιο προφανείς μελλοντικές κατευθύνσεις για την βελτίωση του μοντέλου είναι η καλύτερη ρύθμιση των παραμέτρων, είτε μέσω πολλαπλών πειραμάτων για κάθε τιμή υπερπαραμέτρων, είτε για εξέταση όλων τα υπερπαραμέτρων συγχρόνως (και όχι σειριακά), είτε μέσω επικύρωσης *k* τημημάτων. Μια άλλη είναι η χρήση των ποσοτήτων CA και DA, που εισήγαγε το CADA-VAE, στην συνάρτηση σφάλματος.

Ωστόσο, από μια πιο ποιοτική άποψη, πρέπει να επιτευχθεί καλύτερος συγχρονισμός των αυτοκωδικοποιητών στον χώρο κοινής ενσωμάτωσης. Όπως φάνηκε από την επιθεώρηση του χώρου κοινής ενσωμάτωσης της ενότητας 4.5, τα δείγματα εικόνων και περιγραφών των γνωστών κλάσεων κωδικοποιούνται συγχρονισμένα στον χώρο κοινής ενσωμάτωσης, ωστόσο των άγνωστων κλάσεων, αν και οργανώνονται σε συστάδες, δεν είναι συγχρονισμένα. Αυτό έχει ως αποτέλεσμα της επιλογής παραπλήσιων κατηγοριών από τον ταξινομητή.

Βιβλιογραφία

- [1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, “Label-embedding for image classification”, *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1425–1438, 2015.
- [2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, “Evaluation of output embeddings for fine-grained image classification”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2927–2936.
- [3] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein gan”, *arXiv preprint arXiv:1701.07875*, 2017.
- [4] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, “Generating sentences from a continuous space”, *arXiv preprint arXiv:1511.06349*, 2015.
- [5] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, “Synthesized classifiers for zero-shot learning”, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5327–5336.
- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database”, in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.
- [7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, “Describing objects by their attributes”, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 1778–1785.
- [8] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, “Devise: A deep visual-semantic embedding model”, in *Advances in neural information processing systems*, 2013, pp. 2121–2129.
- [9] C. R. Givens, R. M. Shortt, *et al.*, “A class of Wasserstein metrics for probability distributions.”, *The Michigan Mathematical Journal*, vol. 31, no. 2, pp. 231–240, 1984.
- [10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets”, in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [11] A. Gretton, K. Borgwardt, M. J. Rasch, B. Scholkopf, and A. J. Smola, “A kernel method for the two-sample problem”, *arXiv preprint arXiv:0805.2368*, 2008.

- [12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition”, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [13] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, “Variational deep embedding: An unsupervised and generative approach to clustering”, *arXiv preprint arXiv:1611.05148*, 2016.
- [14] D. P. Kingma and M. Welling, “Auto-encoding variational bayes”, *arXiv preprint arXiv:1312.6114*, 2013.
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [16] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai, “Generalized zero-shot learning via synthesized examples”, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4281–4289.
- [17] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer”, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 951–958.
- [18] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks”, in *Advances in neural information processing systems*, 2017, pp. 700–708.
- [19] D. G. Lowe, “Object recognition from local scale-invariant features”, in *Proceedings of the seventh IEEE international conference on computer vision*, Ieee, vol. 2, 1999, pp. 1150–1157.
- [20] ——, “Distinctive image features from scale-invariant keypoints”, *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space”, *arXiv preprint arXiv:1301.3781*, 2013.
- [22] A. Mishra, S. Krishna Reddy, A. Mittal, and H. A. Murthy, “A generative model for zero shot learning using conditional variational autoencoders”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2188–2196.
- [23] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, “Zero-shot learning by convex combination of semantic embeddings”, *arXiv preprint arXiv:1312.5650*, 2013.
- [24] G. Patterson and J. Hays, “Sun attribute database: Discovering, annotating, and recognizing scene attributes”, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 2751–2758.
- [25] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global Vectors for Word Representation.”, in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

- [26] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, “Contractive auto-encoders: Explicit invariance during feature extraction”, 2011.
- [27] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, “Generalized zero-and few-shot learning via aligned variational autoencoders”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8247–8255.
- [28] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, “Style transfer from non-parallel text by cross-alignment”, in *Advances in neural information processing systems*, 2017, pp. 6830–6841.
- [29] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, “Zero-shot learning through cross-modal transfer”, in *Advances in neural information processing systems*, 2013, pp. 935–943.
- [30] K. Sohn, H. Lee, and X. Yan, “Learning structured output representation using deep conditional generative models”, in *Advances in neural information processing systems*, 2015, pp. 3483–3491.
- [31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions”, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [32] Y.-H. H. Tsai, L.-K. Huang, and R. Salakhutdinov, “Learning robust visual-semantic embeddings”, in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017, pp. 3591–3600.
- [33] V. K. Verma and P. Rai, “A simple exponential family framework for zero-shot learning”, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2017, pp. 792–808.
- [34] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset”, 2011.
- [35] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, “Latent embeddings for zero-shot classification”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 69–77.
- [36] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, “Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly”, *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.
- [37] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, “Feature generating networks for zero-shot learning”, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5542–5551.