Content-Based Image Retrieval from Weakly-Supervised Disentangled Representations

Anonymous Author(s)
Affiliation
Address
email

Abstract

In content-based image retrieval (CBIR) at category level, the relevance of a retrieved image is evaluated with respect to a category of interest. In the best-case scenario, there is an available labeled dataset that can be used to train models that create the necessary data representations for CBIR. However, this is not always the case. In this paper, we explore the use of disentangled representation learned via weak supervision by using data organized into groups with shared content information. We show that such models attain competitive retrieval performances compared to unsupervised models. Moreover, since disentangled representations separate the explanatory factors in datasets, the learned representations can be used also for CBIR with respect to other categories that might be of interest but for which there is little supervision available.

1 Introduction

In content-based image retrieval (CBIR) at category level the goal is to rank images of a database based on whether they share a relevant category with a query image. This technique has been used for person re-identification [1],product matching [2] and shape retrieval [3] among other applications.

The retrieval is usually done by creating/learning a low-level representation of images that can be used for efficient ranking, the representations can then be fine-tuned by using labeled data to improve the clustering of similar images together [4]. The ranking of images within the database is created by sorting their low-dimensional representation based on their distance to the representation of the image query. When provided with enough labels on the categories of relevance current models can achieve a high performance [4, 5].

However, data labelling can be an expensive task and the fine-tuning of the data representations restricts the retrieval of images to the available labels. For example, consider a database of object images in Figure 1. Images might be labeled according to a certain category of interest, such as the type of object. One can then do retrieval only based on a the category

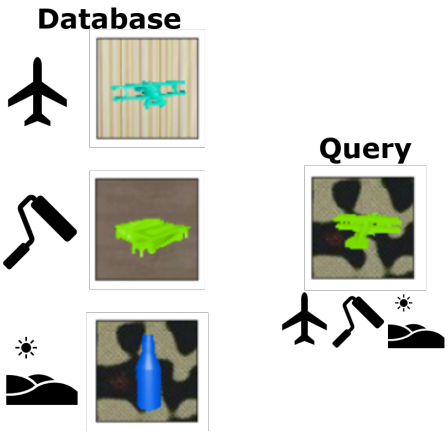


Figure 1: Content-Based Image Retrieval (CBIR) is the ordering of a database of images based on their relevance with respect to a query. However, the relevance of an image can depend not on a single criteria. For example, the relevance of an image can be in terms of the object depicted, in its color or even with respect to the background.

Submitted to NeurIPS 2021 Workshop on DGMs and Downstream Applications. Do not distribute.

of interest, however, retrieval based on other criteria such as the color of the object, the background, or the orientation might be not possible. There are different unsupervised methods for image retrieval [4], but they do not achieve a performance as high as that of supervised methods.

In general, weak supervision describes the information available from a generative process in which the relationship among datapoints can be known [6]. In this work we will focus on the weak supervision from match pairing in which the data can be organized into groups with a commonly shared generative factor. An example of such generative process in real life are the images from online retailers in which sellers upload images from the same product.

In this work, we explore the use of weak supervision for learning disentangled data representations, which separate the relevant factors of variation into separate dimensions of a vector representation [7], that can be used for CBIR with respect to the disentangled factors. We show that the weakly-supervised methods can help with the retrieval of semantic categories which are dependent on commonly shared generative factors.

2 Content-Based Image Retrieval

The goal of Content-Based Image Retrieval (CBIR) is to rank the images withing a database based on their similarity to an image query. This is usually done by creating a low dimensional representation of the images which allows efficient ranking. We will consider that there is an encoding function $h : X \to Z$ which maps images into a low dimensional representation space Z. Images for which their representation is closer to the query embedding are considered the most relevant and are ranked first. Initially, these representations were hand-engineered by practitioners [4], however with the advent of neural networks current approaches learn these representations automatically by using large amounts of data.

The ranking of the images in the database with respect to an image query is determined based on the distance between encoded data. Throughout this work we use as distance function $d : Z \times Z \to \mathbb{R}_{\geq 0}$ the cosine distance given by $d(z, z') = 1 - \frac{z^T z'}{\|z\|\|z'\|}$.

3 Weak Supervision

Weakly-supervised disentanglement in the form of match pairing can be explained in terms of generative models [6]. Assume that there is a set of unobserved factors S which can explain the variability of data and is considered to partake in the data-generative process. Let $g : S \to X$ be a deterministic function that models the data generation from these underlying factors, notice that this function can be substituted with a probability distribution over the data space conditioned on the underlying factors. In particular, in this work we assume that the underlying unobserved factors can be divided into independent subsets $S = S_C \times S_P$. The set S_C represents the shared content in the images and the set S_P represents other variations in the data which we refer to as the perspective factors as in [8].

We assume that the perspective factors do not affect the discrimination of the underlying content. This means that a perfect content inference model $f : X \to S_C$ should be invariant to the visual changes in data due to perspective. That is, for two images generated from different perspective factors $s_p, s'_p \in S_P$ the inference model should provide the same results $f(g(s_C, s_P)) = f(g(s_C, s'_P))$. For example, changes in the images of an airplane with different lighting conditions, orientations and even changes in the color of the airplane do not change the inference of the content.

Let $\mathcal{X} = \{x_1^{(n)}, \ldots, x_K^{(n)}\}_{n=1}^N$ be a dataset of images consisting of N groups with a shared common content factor and K different perspectives. Each image can then be expressed as $x_k^{(n)} = g(s_C^{(n)}, s_P^{(k)})$ with the corresponding $n \in \{1, \ldots, N\}$ content $s_C^{(n)} \in S_C$ and $k \in \{1, \ldots, K\}$ perspective $s_P^{(k)} \in S_P$, see Figure 2.

We assume that the acquisition of weakly-supervised data can be a cheap procedure in some practical scenarios. Some examples can be video recordings where across several frames the content of the image is fixed, another example is an agent moving in space where across consecutive images the

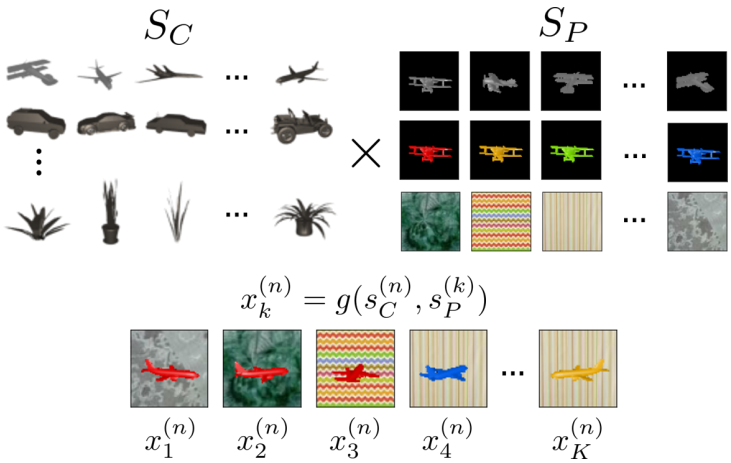


Figure 2: Data generation obtained from the combination of different content and perspective factors to create a weakly-supervised dataset. The content information of S_C consist of the information about the specific object instances depicted in the image. The different perspectives S_P can include changes in the orientation of the objects, their color, their background, etc.

content might be fixed. Even for product placement in classifieds platforms where users upload images from a product to be sold.

We would like to create data representations that exploit the structure provided by the data to create representations that separate each of the underlying factors. Then use those representations for CBIR to be able to retrieve data based on the different underlying factors at will by separating the data representations accordingly.

4 Data Representation Models

To show the use of weakly-supervised disentangled representations for CBIR we will use the Variational Autoencoder (VAE) [9] as a completely unsupervised baseline model. We later describe the basic ideas surrounding the weakly-supervised disentanglement models used and introduce a modification to the loss that could potentially be used to provide better clustering of data for CBIR.

4.1 Variational Autoencoders

A Variational Autoencoder (VAE) [9] models the data creation by introducing a set of unobserved latent variables Z that condition the data generative process. The VAE uses neural networks to estimate the parameters of an approximation to the true posterior distribution with an element of a parametric family of probability distributions $\mathbb{Q}_{Z|x}$. The loss function for training the VAE consists of the negative Evidence Lower Bound (ELBO) which for a datapoint $x \in X$ is given by:

$$\mathcal{L}(x) = -\mathbb{E}_{z\sim\mathbb{Q}_{Z|x}}\left[\log P_{X|z}(x)\right] + \beta KL(\mathbb{Q}_{Z|x}|\mathbb{P}_Z) \tag{1}$$

The first term of the loss is denominated the reconstruction loss and the second term corresponds to the KL divergence between the posterior and the prior. The β hyperparameter was introduced in [10] and is a weighting factor that balances the contribution of the KL divergence, when $\beta = 1$ we return to the original standard VAE [9].

In this work we consider the latent space as a D-dimensional Euclidean latent space $Z = \mathbb{R}^D$ where D is much lower than the data dimensionality. The posterior approximate is taken as a Gaussian distribution with a diagonal covariance matrix. The location parameter of the posterior and the individual standard deviations per dimension are estimated with neural networks i.e. $\mu : X \to \mathbb{R}^D$ and $\sigma : X \to \mathbb{R}^D_{\geq 0}$. Similarly, the distribution $\mathbb{P}_{X|z}$ is considered as Bernoulli, the resulting reconstruction loss corresponds to the binary cross entropy between the normalized pixels in the original images and the reconstructed ones. The prior $\mathbb{P}_Z$ is chosen as a standard normal Gaussian with mean in the origin and the identity matrix as the covariance. In this work we consider the

encoding function h used for CBIR (see Section 2) for all models to be the location parameter of the posterior i.e. $h(x) = \mu(x)$.

4.2 Weakly-Supervised Models for Disentangled Representation Learning

In disentangled representation learning the goal is to find an encoding function that creates low-dimensional representations in which each dimension captures the information about the factors of variation. Currently, there is no agreed upon definition of what are the characteristics that should define such low-dimensional representations.

In recent years several models for weakly-supervised disentangled representation learning have been proposed [11, 12, 13]. In this work we evaluate the representations obtained from training a Grouped Variational Autoencoder (GVAE) [11] and the Adaptive Grouped Variational Autoencoder (AdaGVAE) [13] in CBIR.

The GVAE works with data organized into groups with shared content and different perspectives. The latent space Z is then separated into two subspaces $Z = Z_C \times Z_P$ which should capture the corresponding content and perspective information. Consider two images $x = g(s_C, s_P)$ and $x' = g(s_C, s'_P)$ with shared content, the main idea is to average the distribution over the latent dimensions that should encode the same information. The distribution for the dimensions that encode the information about the perspective should then be different. For the image x, and similarly for x', the resulting posterior distribution becomes

$$\hat{\mathbb{Q}}_{Z_C|x} = a(\hat{\mathbb{Q}}_{Z_C|x'}, \hat{\mathbb{Q}}_{Z_C|x}) \quad \text{and} \quad \hat{\mathbb{Q}}_{Z_P|x} = \mathbb{Q}_{Z_P|x}. \tag{2}$$

where a is a function that averages two distributions. In the case of the GVAE and AdaGVAE the average of the posteriors corresponds to a normal distribution with location parameter given by $\frac{\mu(x)+\mu(x')}{2}$ and the standard deviation given by $\frac{\sigma(x)+\sigma(x')}{2}$. The loss function for the data pair is

$$\mathcal{L}_{WS}(x, x') = -\mathbb{E}_{z\sim\hat{\mathbb{Q}}_{Z|x}}\left[\log P_{X|z}(x)\right] - \mathbb{E}_{z\sim\hat{\mathbb{Q}}_{Z|x'}}\left[\log P_{X|z'}(x)\right] \tag{3}$$

$$+ \beta KL(\hat{\mathbb{Q}}_{Z|x}|\mathbb{P}_Z) + \beta KL(\hat{\mathbb{Q}}_{Z|x'}|\mathbb{P}_Z). \tag{4}$$

Notice that this loss function can be extended to any number of datapoints with shared content and for the GVAE it is not restricted to pairs.

The GVAE predefined the separation of the latent dimensions into the two subspaces Z_C and Z_P, however this restricts the disentanglement to only separate the content from the perspective. The AdaGVAE proposes a modification to the GVAE that allows the separation of multiple factors by identifying the latent dimensions that should capture the factor variations among pairs of data. The only requirement for this model is that only a few factors are changing among the pairs, the least factors are changing between images, the better.

4.3 Normalized Temperature/Prototypes for CBIR

Disentangled representation models are not optimized for CBIR in which ideally, the distance between encoded datapoints is important to determine the similarity among images. The closer similar images are mapped together, and the farther dissimilar images are mapped, the better the ranking. Because of this, we propose a modification of both the GVAE and the Ada-GVAE by adding the prototypical loss as a contrastive loss that encourages clustered data [14].

The prototypical network [14] was introduced as a metric learning approach for few-shot classification where for each class a prototypical embedding is estimated. The prototype to which an embedded datapoint is closest determines its class. In our case we propose to estimate for each grouped data a prototype, for a given group of images $\{x_1^n, \ldots, x_K^n\}$ the corresponding prototype in the representation space $z^{(n)} \in Z$ is given by

$$z^{(n)} = \frac{1}{K}\sum_{k=1}^{K} h(x_k^{(n)}) \tag{5}$$

For a given datapoint x the probability that the content s_C in the image corresponds to that of group n is given by:

$$P_{S_C|x}(n)\frac{\exp\left(-d(h(x),z^{(n)})/\tau\right)}{\sum_{n'=1}^{N}\exp\left(-d(h(x),z^{(n')})/\tau\right)} \tag{6}$$

where $\tau \in \mathbb{R}_{>0}$ is the temperature. The smaller the temperature the tighter the clustering. Throughout this work we have used $\tau = 1$.

By adding the log-likelihood of this probability as an extra loss term we encourage that the embeddings with shared content to be mapped close to their prototype which should result in more compact data representations.

For a given datapoint belonging to the n-th group we maximize the log-likelihood in Equation 6 resulting in the normalized temperature loss term

$$\mathcal{L}_{NT}(x^{(n)}) = d(h(x^{(n)}), z^{(n)}/\tau) - \log\sum_{n'=1}^{N}\exp\left(-d(h(x^{(n)}), z^{(n')})/\tau\right). \tag{7}$$

For a dataset $\{x_1^{(1)},\ldots,x_K^{(1)}\}_{n=1}^{N}$ the loss is given by

$$\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_{WS}\left(x_1^{(n)},\ldots,x_K^{(n)}\right) + \alpha\frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{NT}(x_k^{(n)}) \tag{8}$$

with α a weighting factor that balances the contribution of the prototypical loss pull and the VAE objective. We refer to GVAENT and AdaGVAENT as the modified models that incorporate the normalized temperature loss.

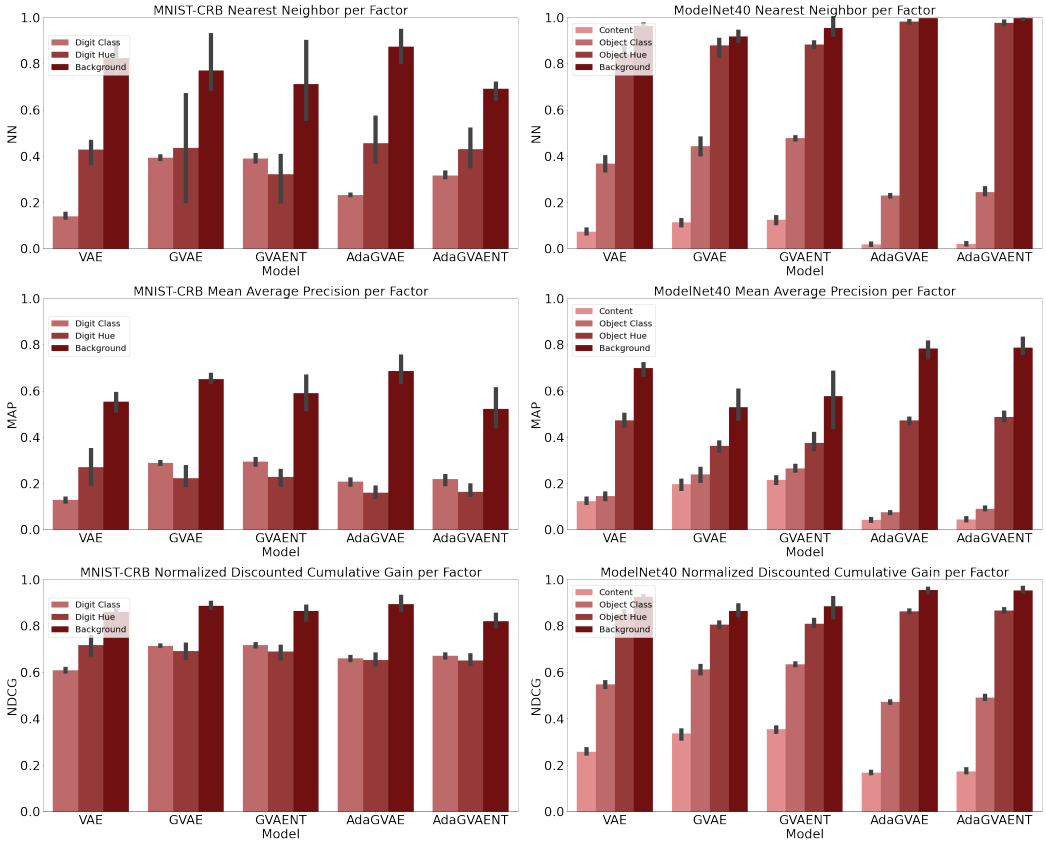
5 Experiments

In order to evaluate the CBIR performance of the models we have created two synthetic datasets that are organized into N groups with shared content and $K = 6$ perspective variations as described in Section 4.2. The generative process used to create each dataset was to first retrieve a unique content factor and then apply K transformations by uniformly sampling different perspective factors to modify the data accordingly. Each perspective variable corresponds to a certain change in the appearance of the particular content depicted as it is been shown in Figure 2.

For each of the datasets we evaluate the models retrieval capabilities to retrieve images with respect to different criteria. We evaluate the capabilities of the models for CBIR of semantic categories such as digit and object class in the MNIST-CRB and ModelNet40 dataset respectively. We also evaluate the CBIR capabilities of the learned representations to retrieve images based on different perspectives and such as color of the digit or object, the background and in the case of the ModelNet40 datasets the capability to retrieve images from with the same content.

We assume that for a trained data representation model an oracle is available which can identify the latent dimensions which best capture the relevant information that represents this criteria. This is achieved by measuring the mutual information between the encoded data and the true labels, the latent dimensions with higher mutual information than the mean are selected to retrieve the images with the information of interest. Notice however that if a truly disentangled representation is available very few labeled data is required to identify the relevant latent dimensions for each criteria. However, we propose an exploration of such separation of the latent dimensions with less data as future work.

MNIST-CRB This dataset consists of the MNIST digit dataset [15] where the content variable per group corresponds to an instance of a digit, there are 60,000 different digit instances in the training dataset and 10,000 in the test. The images are transformed such that the different perspectives correspond to changes in the hue of the digit (10 hues), its orientation (10 orientations) and the background (10 different backgrounds) in which it is embedded. The backgrounds are chosen from the visual domain decathlon's describable textures [16]. See Figure 4a. Images are encoded in a $D = 10$ dimensional latent space. In the case of the GVAE the representation space is divided into $Z_P = \mathbb{R}^3$ and $Z_C = \mathbb{R}^10$.



(a) Scores for MNIST-CRB dataset (b) Scores for ModelNet40 dataset

Figure 3: Nearest Neighbor (NN), Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) scores measured for retrieval of the semantic class of interest, the hue of the object instances and the background using VAE, GVAE, GVAENT, AdaGVAE and AdaGVAENT models.

(a) MNIST-CRB examples (b) ModelNet40 examples

Figure 4: Example of the datasets used. In each of the datasets a batch is formed by images with a fixed content and varying perspective.

ModelNet40 The ModelNet40 [17, 18] dataset consists of 3D models from 40 different categories aligned to a standard orientation. The fixed content in this dataset corresponds to a particular instance of a 3D model there are 9843 training 3D models and 2468 for the test dataset. Images from twelve distinct orientations are rendered. Each image is then transformed to change the hue of the object (6 hues), its orientation (12 orientations) and the background (6 backgrounds) using the visual domain decathlon's describable textures [16]. See Figure 4b. Images are encoded in a $D = 100$ dimensional latent space. In the case of the GVAE the representation space is divided into $Z_P = \mathbb{R}^3$ and $Z_C = \mathbb{R}^97$.

6 Evaluation

For each dataset, we divide the test dataset of images generated from unseen content factors. We create an image database by randomly selecting a single image per content variable in the test set and a query set by selecting three images per content variable. Each image in both the database and the query are encoded into the low dimensional space and a distance matrix is computed by comparing the cosine distance between images . For each query a ranked list of the images in the database is created by sorting the database image representations based on their cosine distance to the query. Depending on the retrieval factor to be taken into account the true positive images in the ranked list are identified.

We evaluate the performance of each model for CBIR by measuring the Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG) and Nearest Neighbor (NN) precision which are typical metrics to evaluate information retrieval [19]. For each metric the bar height represents the mean and the whiskers the 95% confidence interval across three repetitions. We assess the performance of each model to retrieve new unseen instances of the data based on the characteristics of interest i.e. digit number, hue of digit, background and orientation for the MNIST-CRB dataset and object category, object hue, background and orientation for ModelNet40 dataset.

7 Results

The resulting MAP and NDCG values provide an overview of the performance across the complete database of images and the values measured are shown in Figure 3 for the retrieval with respect to different criteria. The results show that the performance of the retrieval for number class in MNIST-CRB is higher for weakly supervised models compared to the VAE. This shows that including such information into the data modelling can help in the identification of semantic classes of interest. Both GVAE and GVAENT achieve the highest values compared to their Adaptive counterparts. On the other hand, these models obtain the same performance as a VAE for other visual factors such as the color of the digit or the background which are the most visually salient features in terms of number of pixels in the image.

On the other hand for the ModelNet40 dataset also the performance of the weakly-supervised methods to retrieve the object class is higher for both GVAE and GVAENT models. The adaptive methods (AdaGVAE and AdaGVAENT) do not provide an increase in the retrieval performance for the category semantic factors nor the color hue. However, they show an increase in the performance for the identification of the background depicted in images.

The NN results show the retrieval capabilities of the models to retrieve the most relevant image in a database, the results are very similar to the aggregated MAP and NDCG. The only difference lies in the higher precision of models to retrieve the most relevant content compared to the MAP values across the complete database.

The normalized temperature loss for clustering the encodings for CBIR does not provide a noticeable improvement in the retrieval with respect to each of the factors. However, it might be interesting to explore the use of different temperature values to identify whether there is a strong dependency to this hyperparameter.

8 Conclusions

The use of weak-supervision seems to provide a benefit in CBIR of semantic categories which depend on the content information provided by the weak supervision such as object type and class digit compared to completely unsupervised methods. The obtained data representations can be used to retrieve other visually salient features similarly to what unsupervised models such as VAEs can achieve.

We have proposed the use of a contrastive loss to encourage better clusters for retrieval, however we have not observed empirical evidence that this can provide a significant increase in the performance for retrieval. It might be then interesting to test the dependency of the model's performance on the choice of hyperparamters and the use of different contrastive losses to achieve better clustering in semi-supervised settings.

A Hyperparameters & Neural Network Architectures

A.1 Hyperparameters

Each model was trained for 100 epochs with the Adam optimizer with learning rate 0.001, the hyperparameters for the losses of Equation 8 and Equation 4.2 were fixed to $\alpha = 100$ and $\beta = 1$ throughout all experiments.

The hardware used across all experiments was a DGX station with 4 NVIDIA GPUs V100 and 32GB. Only one GPU was used per experiment.

A.2 Architectures

We use the same architecture as in [13] for the ModelNet40 dataset and a modified version for MNIST-CRB. The details of the architectures are presented in Table1 and Table2 respectively.

Table 1: Encoder and decoder architectures used for ModelNet40 dataset.

Encoder	
Input	size (64,64, number channels)
Conv	filters 32, kernel 4, stride 2, ReLU
Conv	filters 32, kernel 4, stride 2, ReLU
Conv	filters 64, kernel 4, stride 2, ReLU
Conv	filters 64, kernel 4, stride 2, ReLU
Dense	units 256, ReLU
Dense(x2)	Latent dim for mean and standard deviation
Decoder	
Input	Nnumber of latent dimensions
Dense	units 256, ReLU
Dense	units 4*4*64, ReLU
Reshape	(4,4,64)
ConvT	filters 64, kernel 4, stride 2, ReLU
ConvT	filters 32, kernel 4, stride 2, ReLU
ConvT	filters 32, kernel 4, stride 2, ReLU
ConvT	filters (number channels), kernel 4, stride 2, Sigmoid

Table 2: Encoder and decoder architectures used for ModelNet40 dataset.

Encoder	
Input	size (28 , 28, number channels)
Conv	filters 32, kernel 3, stride 2, ReLU
Conv	filters 64, kernel 3, stride 2, ReLU
Dense	units 256, ReLU
Dense	units 128, ReLU
Dense(x2)	Latent dim for mean and standard deviation
Decoder	
Input	Number of latent dimensions
Dense	units 128, ReLU
Dense	units 256, ReLU
Dense	units 7*7*64, ReLU
Reshape	(7,7,64)
ConvT	filters 64, kernel 3, stride 2, ReLU
ConvT	filters (number channels), kernel 3, stride 2, Sigmoid

References

[1] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-identification: A Benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124. IEEE, 12 2015.

[2] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem(1):1096–1104, 2016.

[3] Juefei Yuan, Hameed Abdul-Rashid, Bo Li, and Yijuan Lu. Sketch/Image-Based 3D Scene Retrieval: Benchmark, Algorithm, Evaluation. *Proceedings - 2nd International Conference on Multimedia Information Processing and Retrieval, MIPR 2019*, pages 264–269, 2019.

[4] Wei Chen, Yu Liu, Weiping Wang, Erwin Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S. Lew. Deep Image Retrieval: A Survey. pages 1–20, 1 2021.

[5] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823. IEEE, 6 2015.

[6] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly Supervised Disentanglement with Guarantees. *International Conference on Learning Representations*, (Lld):1–8, 10 2020.

[7] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 09–15 Jun 2019.

[8] Nicki Skafte Detlefsen and Søren Hauberg. Explicit Disentanglement of Appearance and Perspective in Generative Models. In *Advances in Neural Information Processing Systems*, number NeurIPS, pages 1018–1028, 2019.

[9] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[10] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning Basic Visual Concepts With a Constrained Variational Framework. *International Conference on Learning Representations*, (July):1–13, 2017.

[11] Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. *IJCAI International Joint Conference on Artificial Intelligence*, 2019-Augus(i):2506–2513, 2019.

[12] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 2095–2102, 2018.

[13] Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6348–6359. PMLR, 13–18 Jul 2020.

[14] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 2017-Decem:4078–4088, 2017.

[15] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[16] Tomas Jakab Hakan Bilen, Sylvestre Rebuffi. Visual domain decathlon, 2017.

[17] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A Deep Representation for Volumetric Shapes. 2014.

[18] N. Sedaghat and T. Brox. Unsupervised generation of a viewpoint annotated car dataset from videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[19] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval*, volume 38. Cambridge University Press, Cambridge, 2008.