Content-Based Image Retrieval from Weakly-Supervised Disentangled Representations

Luis A. Pérez Rey [ab], **Mike Holenderski** [a], **Dmitri Jarnikov** [ab]
[a] Eindhoven University of Technology, Eindhoven, The Netherlands
[b] Prosus, Amsterdam, The Netherlands
{l.a.perez.rey, m.holenderski, d.s.jarnikov}@tue.nl

Abstract

In content-based image retrieval (CBIR), a database of images is ordered based on the similarity to a query image. Similarity criterion is usually determined with respect to a shared category e.g. whether the database images contain an object of the same type as depicted in the query. Depending on the situation, multiple similarity criteria can be relevant such as the type of object, its color, or the depicted background. Ideally, a dataset labeled with all possible criteria information is available for training a model for computing the similarity. Typically, this is not the case. In this paper, we explore the use of disentangled representations for CBIR with respect to multiple criteria. To alleviate the need for labels, the models used to create the representations are learned via weak supervision by using data organized into groups with shared information. We show that such models can attain better retrieval performances compared to unsupervised baselines.

1 Introduction

In content-based image retrieval (CBIR) at category level the goal is to rank images of a database based on their similarity to a query image, the similarity criterion is based on whether the images share a pre-defined category of interest. For example, whether a certain type of object is depicted in the images. This technique has been used for person re-identification [1], product matching [2], shape retrieval [3] among other applications.

In some applications, it can be useful to select, from a list of multiple options, the similarity criterion to be used with a query image [4]. For a given query image, the database ordering is conditioned on the selected criterion. For example, in online retail, the retrieval of relevant products could be based on the type of product, but also on the color, material, brand, etc. See Figure 1.

Typically, the ranking of a database of images is obtained by comparing the distances between low-dimensional representations of the database images with respect to a low-dimensional representation of the query image. The smaller the distance, the more similar an image of the database is considered to be. These low-dimensional representations are obtained from the visual features extracted by a model, commonly, neural networks [5, 6].

The models used to create the data representations are fine-tuned to improve the clustering of similar images using data labeled with information with the corresponding criterion [5]. However, labeling data for multiple criteria can be very expensive. Moreover, when changing the similarity criterion used for CBIR, the representations should also change accordingly.

Disentangled representation learning is a promising approach for this task since it has the goal of creating data representations which separate the factors of variation within data into different dimensions of a vector representation [7]. By selecting the appropriate dimensions that encode the information corresponding to a particular criterion, the retrieval can be tailored to different needs.

NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications

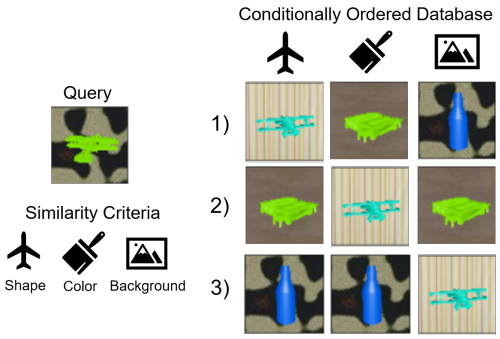


Figure 1: Content-Based Image Retrieval is the ordering of a database of images with respect to a query image and a similarity criterion.The retrieval criterion could be conditioned on different categories. For example, in terms of the object depicted, the color of the object or the background.

Initial disentangled representation learning approaches were unsupervised, however, it has been proven that completely unsupervised disentangled representation learning is not possible [8]. Recently, the use of weak supervision has been proposed as a way to attain disentangled representations using inductive priors which exploit known information about the generative process [9].

There are different forms of weak supervision. In particular, match-pairing describes weak supervision in which data is organized into groups with commonly shared information. This kind of supervision is easy to acquire in many applications. Examples include: images in online classifieds platforms organized into groups with the same product depicted, video recordings with shared content in consecutive frames or successive observations from agents interacting with an environment.

In this work, we explore the use of existing weak supervision methods that use matched-pairs for learning disentangled data representations which separate the relevant factors of variation into separate dimensions of a vector representation [8]. We propose a modification to these methods by introducing a prototypical loss to encourage better clustering of data. We evaluate whether the obtained representations can be used for CBIR conditioned on different similarity criteria. We provide empirical results showing that weakly-supervised disentangled representations can attain better performance compared to the unsupervised baselines.

2 Conditional Content-Based Image Retrieval

The goal of Content-Based Image Retrieval (CBIR) is to rank the images within a database based on their similarity to an image query. The similarity whether they share a previously specified category depicted in an image query [5]. This is usually done by creating low-dimensional representations for the images. The similarities for these representations can be efficiently computed, both memory and computation-wise, in contrast to direct image-to-image comparisons. We will assume that the data representation is obtained through an encoding function $h : X \to Z$ which maps images X into a low dimensional vector representation space Z.

For a given query, the images in a database are ordered based on the distance between their low-dimensional data representations to the query embedding. The closer the low-dimensional representations of the database images are to the query, the more relevant they are considered .

In the case of conditional CBIR from multiple similarity criteria consider that there is a set of M similarity criteria $\{Y_m\}_{m=1}^{M} Y$ where each criterion Y_m is a set of possible discrete categories. An image $x \in X$ is described in terms of each of this categories $\{y_x^{(m)}\}_{m=1}^{M}$ where $y_x^{(m)} \in Y_m$. An image x is similar to x' according to the m-th criterion if $y_x = y_{x'}^{(m)}$ and dissimilar otherwise.

Given a query image, the ranking of a database is conditioned on a selected similarity criterion Y_m, distances in the low-dimensional representations of the database and queries should reflect the selected criterion. Approaches [10, 4] where the data representations are masked to capture the information about the different similarity criterion in different parts of the representation have been proposed. However, they require labels for all the similarity criteria for each image.

The separation of information about the different similarity criteria can be attained through disentangled representation learning [7] where the goal is to create data representations that separate the factors of variation within data into different dimensions of a vector representation. The assumption is that the factors of variation capture the information about the similarity criteria. One can condition the retrieval of images by selecting different dimensions of the data representation. A recent method for conditional CBIR has proposed the use of disentangled data representations [11]. However, this method is completely supervised. In this work, we explore the use of weak supervision to create disentangled data representations for conditional CBIR with fewer labels.

3 Weak Supervision for Conditional Content Based Image Retrieval

Weak supervision in disentangled representation learning describes the information available from a certain generative process used to create data representations that separate the factors of variation of data. Different types of weak supervision have been characterized in [9] namely: restricted labeling, match pairing and rank pairing. In this work we will explore the use of weak supervision in the form of match pairing to create disentangled representations useful for conditional CBIR.

Weakly-supervised disentanglement in the form of match pairing can be explained in terms of generative process that creates data that can be organized into groups with shared information [9]. 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. 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 information among a group of data, which we refer to as the 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 [12].

For example, consider the set S_C consisting of all the information that describes the identity of a particular object instance and S_P as the set that represents changes in the orientation of the object, its color and its background. The n-th group of data is created by first sampling a code element corresponding to the information of a particular object from S_C, then sampling K combinations of different perspectives from S_P to generate the data via g. See Figure 2.

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 different perspectives. For two images generated from different perspective factors $s_p, s'_p \in S_P$ the inference model should provide the same results $f(g(s_C, s_P)) = f(g(s_C, s'_P))$.

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 $x_k^{(n)} = g(s_C^{(n)}, s_P^{(n,k)})$ with the corresponding $n \in \{1, \ldots, N\}$ content factor $s_C^{(n)} \in S_C$ and $k \in \{1, \ldots, K\}$ perspective factor $s_P^{(k)} \in S_P$.

We would like to create models 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 different similarity criteria associated to the different underlying factors by separating the data representations accordingly.

4 Data Representation Models

In this section we first describe the unsupervised [13] Variational Autoencoder (VAE) as an unsupervised method for disentangled representation learning. We later describe the basic ideas surrounding the used weakly-supervised disentanglement models based on the VAE and introduce a modification to the loss that we have proposed to include to provide better representations for CBIR.

4.1 Variational Autoencoders

A Variational Autoencoder (VAE) [13] 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

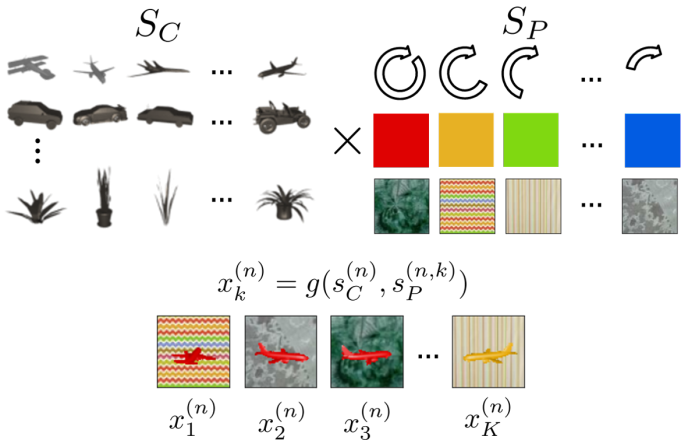


Figure 2: Data generation obtained from the combination of different content and perspective factors to create a weakly-supervised dataset. The content information of S_C consist of the information about the specific object instances depicted in the image. The different perspectives S_P can include changes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

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:

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

where $P_{X|z}$ is the probability density function of the distribution $\mathbb{P}_{X|z}$ and $\mathbb{P}_Z$ is the prior distribution over the latent variables.

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 [14] and is a weighting factor that balances the contribution of the KL divergence. When $\beta = 1$ the model returns to the original standard VAE training scheme [13].

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 as 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 [15, 16, 17]. In this work we evaluate the representations obtained from training a Grouped Variational Autoencoder (GVAE) [15] and the Adaptive Grouped Variational Autoencoder (AdaGVAE) [17] for conditional CBIR.

The GVAE works with data organized into groups with shared content and different perspectives. The latent space Z is separated into two subspaces $Z = Z_C \times Z_P$ which should capture the corresponding content and perspective factors $S = S_C \times S_P$. Consider two images $x = g(s_C, s_P)$ and $x' = g(s_C, s'_P)$ with shared content s_C. The main idea to disentangle the content from the perspective information is to average the distribution over the latent dimensions that should encode

the content information. The distribution for the dimensions that encode the information about the perspective should be different. The resulting posterior distribution $\hat{Q}$ for the content latent subspace depends on both images x and x' while for the perspective latent subspace the posterior depends only on the corresponding image as in

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

where a is a function that averages the input 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 standard deviation given by $\frac{\sigma(x)+\sigma(x')}{2}$. The loss function for the data pair is

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

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

In the GVAE this loss can be extended to any number of datapoints with shared content for the AdaGVAE it is 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 that encourages clustered data [18].

The prototypical network [18] 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 group of data with shared content 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

We evaluate the performance of the models for CBIR with different criteria of similarity on two synthetic datasets obtained through a generative process as described in Section 4.2. For a given image query, we evaluate the sorting of an image database with respect to different similarity criterion including content and perspective factors.

Two synthetic datasets were used: MNIST-CRB consisting of images of digit numbers and ModelNet40 consisting of rendered images of 3D models from 40 categories. Each dataset is organized into N groups with shared content and a fixed number of perspective variations $K = 6$ as described in Section 4.2, the groups are separated into training and test set. Each of the N groups in the dataset are generated by first selecting an image with unique fixed content, and then creating K different images from applying transformations corresponding to different perspectives. In the case of ModelNet40 we first select an object and sample a random rendered image from the available orientations. These perspectives are obtained by applying to the original image a set transformations associated to different combinations of perspective factors, for example, changes in the hue of the object, its orientation and the background see Figure 2.

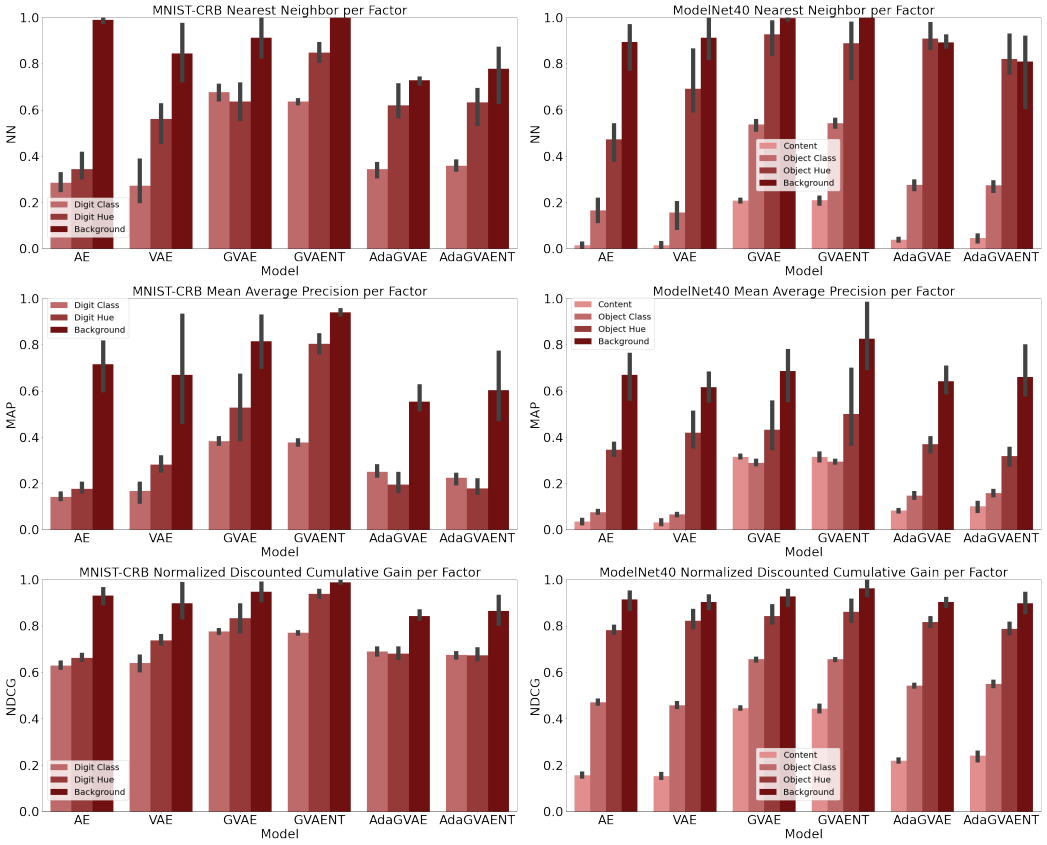
MNIST-CRB This dataset consists of the MNIST digit dataset [19] where the content variable per group corresponds to an instance of a digit, there are 60,000 different digit instances in the training set and 10,000 in the test set. The images are transformed to create different perspectives which correspond to changes in the hue of the digit (10 hues), its orientation (10 orientations) and the background (10 different backgrounds). The backgrounds are chosen from the visual domain decathlon's describable textures [20]. 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 $Z_C = \mathbb{R}^7$.

ModelNet40 The ModelNet40 [21, 22] dataset consists of 3D models from 40 different categories aligned to a standard orientation. The fixed content in each group corresponds to an 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 transformed to create different perspectives corresponding to the change of hue in the object (6 hues), its orientation (12 orientations) and the background (6 backgrounds) using the visual domain decathlon's describable textures [20]. See Figure 4b. 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 $Z_C = \mathbb{R}^7$.

6 Evaluation

As a comparison, we have used a regular Autoencoder (AE) and the Variational Autoencoder (VAE) [13] as completely unsupervised baselines. We leave a more thorough comparison with semi-supervised models as part of our future work. For each measure of similarity, the performance of each model is evaluated by measuring the Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG) and Nearest Neighbor (NN) precision, which are typical metrics for information retrieval [23]. The MAP and NDCG provide a notion of how well retrieval is achieved over a complete database while the NN give insights about the most similar images to a given query. To measure these quantities, the test dataset is divided into a database and query set which are formed by randomly sampling without replacement one and three images from each fixed content group respectively.

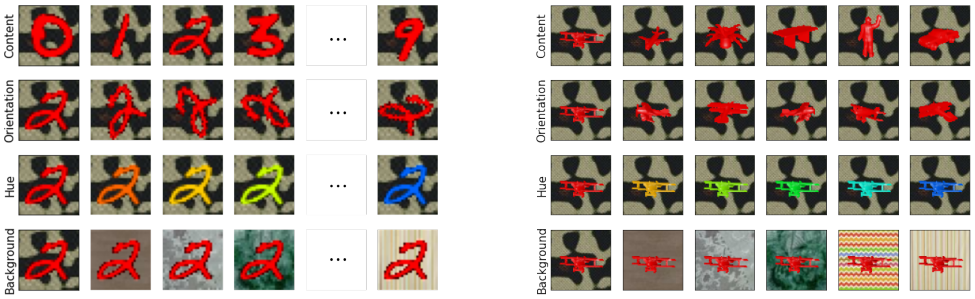
To evaluate the metrics, the query and databaset set images are encoded into their low-dimensional representation. For each query image, a ranked list of database images is created based on their



(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, the hue of the object instance or the background using VAE, GVAE, GVAENT, AdaGVAE and AdaGVAENT models.



(a) MNIST-CRB examples

(b) ModelNet40 examples

Figure 4: Examples of the datasets used. Each row shows different examples from an explanative factor: content, orientation, hue and background.

distance in the low-dimensional representation space. Depending on the measure of similarity being evaluated, the true positive images in the ranked list are identified and are used to compute all the metrics. See Appendix B for a more detailed description of the computation of the metrics.

Throughout this work we use the Euclidean distance between the vector representations in Z. The distance function $d : Z \times Z \to \mathbb{R}_{\geq 0}$ is given by

$$d(z, z') = \|z - z'\|_2^2 = (z - z')^T (z - z') \tag{9}$$

We evaluate the CBIR retrieval with respect to different similarity criteria such as digit class, digit hue and background for MNIST-CRB dataset and object class, object hue, background and content (3D model instance) for the ModelNet dataset.

For a trained data representation model we identify the latent dimensions which best capture the relevant information that represents these criteria by measuring the mutual information between the encoded data and the true similarity criteria labels. The latent dimensions with higher mutual information than the mean are selected to retrieve the images and calculate the distances with the information of interest. This in principle requires all training data to be labeled. Notice however that if a truly disentangled representation is available very little labeled data is required to identify the relevant latent dimensions for each criterion. However, we propose an exploration of such separation of the latent dimensions with little labeled data as future work.

7 Results

7.1 Qualitative Results

We present 2-dimensional projections of the embeddings obtained with Principal Component Analysis (PCA) for qualitative assessment, see Figure 5. We only show the results for MNIST-CRB, similar results were obtained for ModelNet40.

Notice that for both baselines, the AE and VAE, the projections show a similar clustering behaviour for the background and color of the digit. However there is no particular structure for the class information.

In the case of the GVAE and GVAENT, we show the projection of the content Z_C and perspective Z_P latent spaces separately. In both cases the perspective latent space Z_P appears to identify two main directions for the information about the background and color of the digits as can be seen in the corresponding projections where the digit color changes in a diagonal direction while the background changes in a perpendicular direction. On the other hand, the content latent space appears to create regions for the different types of digits. However the separation between the different digit types into regions is not that clear.

Finally, in the case of the AdaGVAE and the AdaGVAENT the projections show a clear separation of the color of the digits and the background. However, there appears to be no clear structure for the object types.

7.2 Quantitative Results

The measured MAP, NDCG and NN for both datasets are plotted in Figure 3. 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 different similarity criteria i.e. by digit number, hue of digit, background and orientation for the MNIST-CRB dataset and by object content, category, object hue and background for ModelNet40 dataset.

Weak Supervision for Content-Related Similarity Criteria The results show that the performance of weakly-supervised models is higher for similarity criteria associated to the content information such as the digit class, object class and the content information itself. This indicates that by grouping data with the same content, one can obtain insights about semantic classes associated to the identities of items depicted. Both, the GVAE and the GVAENT achieve the highest metric values. Note that the use of the normalized temperature does not provide any gain in these similarity criteria.

Weak Supervision in Perspective-Related Similarity Criteria For the unsupervised perspective factors such as digit hue and background, the AE and VAE baselines attain a high performance. These methods tend to favor capturing information about factors that contribute to the reconstruction loss. This means that factors with large pixel features in the images are favoured to be captured. The adaptive methods AdaGVAE and AdaGVAENT perform similarly or slightly worse compared to the baselines. However, the GVAE and the GVAENT have the highest performance. By using the normalized temperature loss in the GVAENT also a slight increase can be observed in the performance for example for the digit hue compared to the GVAE.

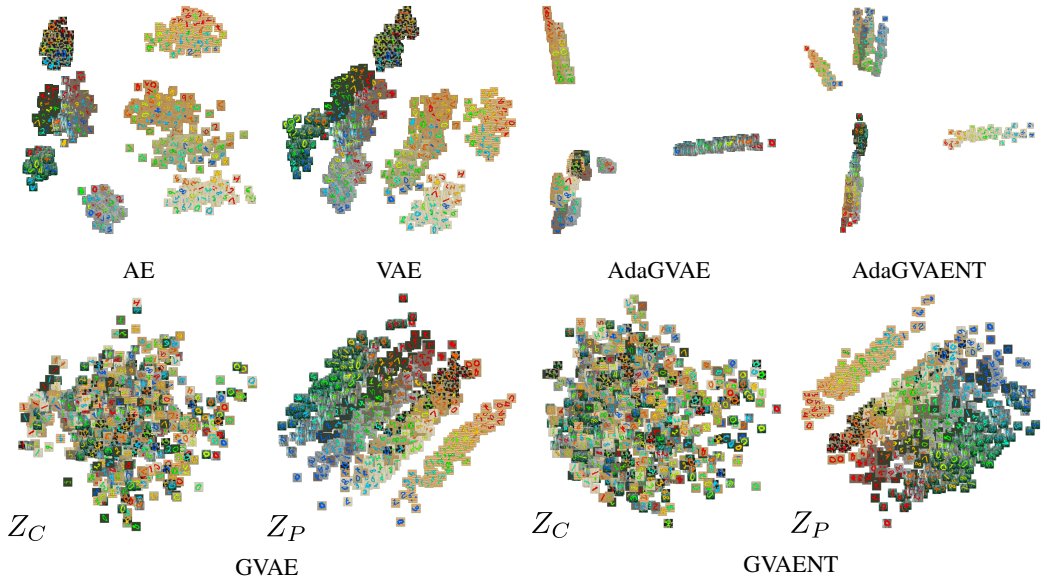


Figure 5: Two-dimensional projections of test image embeddings for MNIST-CRB dataset using Principal Component Analysis (PCA). For each method 1000 embedding images from the test dataset are plotted. Each projected embedding is shown with a miniature image. For the GVAE and GVAENT methods the projections of the content Z_C and perspective Z_P embeddings are shown.

8 Conclusions

Weak-supervision for learning disentangled representations can benefit CBIR compared to completely unsupervised methods, especially w.r.t. similarity criteria that depends on the fixed content information which is contained in the the weak supervision. The obtained data representations can be used to retrieve other visually salient features similarly to what unsupervised models such as AEs and VAEs can achieve.

We have proposed the use of a contrastive loss to encourage retrieval, we have observed that for perspective factors the use of this loss provides better clustering, however there is no evidence of gain for the retrieval with respect to content-based similarity criteria. It might be of interest 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 500 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 [17] for the ModelNet40 dataset and a modified version for MNIST-CRB. The details of the architectures are presented in Table1 and Table2 respectively.

B Metric Description

In this work we evaluate the Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG) and the Nearest Neighbor (NN) metrics for different similarity criteria. Consider that

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

Encoder		Decoder	
Input	size (64,64, number channels)	Input	Number of latent dimensions
Conv	filters 32, kernel 4, stride 2, ReLU	Dense	units 256, ReLU
Conv	filters 32, kernel 4, stride 2, ReLU	Dense	units 4*4*64, ReLU
Conv	filters 64, kernel 4, stride 2, ReLU	Reshape	(4,4,64)
Conv	filters 64, kernel 4, stride 2, ReLU	ConvT	filters 64, kernel 4, stride 2, ReLU
Dense	units 256, ReLU	ConvT	filters 32, kernel 4, stride 2, ReLU
Dense(x2)	units latent dim value (mean, std)	ConvT	filters 32, kernel 4, stride 2, ReLU
		ConvT	filters 3, kernel 4, stride 2, Sigmoid

Table 2: Encoder and decoder architectures used for MNIST-CRB dataset.

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

the selected similarity criterion to be evaluated can be described in terms of a set of possible discrete categories Y.

For a query image x_q and a test database of N images $\mathcal{X}$. The database is encoded and ordered based on the distances to the query embedding i.e. $d(h(x_q), h(x'))$ with $x' \in \mathcal{X}$. An image in the ordered database is considered a true positive if it shares the same category as the query. Let $\{y_{n'}\}_{n'=1}^{N}$ be the set of ordered true categories associated to the ordered images of the database and y_q the category of the query where $y_q, y_{n'} \in Y$ for all $n' \in \{1, \ldots, N\}$. The precision for the n-th retrieved image with respect to query image x_q is measured as

$$P_q(n) = \frac{\sum_{n'=1}^{n} \delta_{y_q, y_{n'}}}{n}. \tag{10}$$

Where $\delta_{y_q, y_{n'}}$ is the Kronecker delta between y_q and $y_{n'}$ and represents the relevance of the retrieved n' image. The total of relevant images for query q is given by $R = \sum_{n'=1}^{N} \delta_{y_q, y_{n'}}$. The NN metric for query image x_q corresponds to the precision of the first retrieved element, i.e. $\mathrm{NN}_q = P_q(1)$.

The MAP score for query x_q is given by the average precision of the relevant retrieved elements and calculated as

$$\mathrm{MAP}_q = \frac{\sum_{n'=1}^{N} \delta_{y_q, y_{n'}} P_q(n')}{R}. \tag{11}$$

The NDCG is a metric gives weights the retrieved images based on their position in the ordered database. For a query image x_q it is calculated as

$$\mathrm{NDCG}_q = \frac{\sum_{n'=1}^{N} \frac{2^{\delta_{y_q, y_{n'}}} - 1}{\log_2(n'+1)}}{\sum_{n'=1}^{R} \frac{1}{\log_2(n'+1)}}. \tag{12}$$

Notice that the NDCG metric has values in the interval $[0, 1]$ where 1 corresponds to perfect retrieval. All metrics are averaged across all possible queries to produce the final score.

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] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:1781–1789, 2017.

[5] 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.

[6] 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.

[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. (1993):1–30, 6 2012.

[8] 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.

[9] 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.

[10] Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. Bayesian representation learning with oracle constraints. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, pages 1–16, 2016.

[11] Yuxin Hou, Eleonora Vig, Michael Donoser, and Loris Bazzani. Learning Attribute-driven Disentangled Representations for Interactive Fashion Retrieval. *International Conference on Computer Vision*, 2021.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

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

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

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

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

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