A Contrastive Learning Approach for Compositional Zero-Shot Learning

Muhammad Umer Anwaar (✉) *, Rayyan Ahmad Khan*, Zhihui Pan, Martin Kleinsteuber
Technische Universität München
Mercateo AG
Munich, Germany
umer.anwaar@tum.de

Figure 1: A potential scenario of a smart IR system, "enabling the customer" to express their mind better.

ABSTRACT

An object can be in several states. For different states (attributes) the object could look dramatically different. Thus, the smart information retrieval systems of the future need to learn good state-object representations. Such a system should not only be able to recognize state-object compositions unseen during training but also be able to retrieve images based on multi-modal (image-text) query. In the literature, these tasks are treated separately. In this work, we propose a unified model, ContraNet, which leverages the rich semantics of the state-object to learn multimodal representation in a contrastive manner. We adopt a deep metric learning approach and learn a multimodal representation by pulling similar images and texts closer to each other and pushing apart different ones. Our autoencoder based model learns the text-aware representation of image which is suitable for both tasks. The reconstruction losses provide additional regularization for learning of the representation. Our approach outperforms the state-of-the-art (SOTA) methods on widely-used benchmarks. Specifically, on the task of state-object composition, ContraNet achieves 8.7% and 8.1% performance gain on UT-Zappos and MIT-States on best HM metric, respectively. For the image retrieval task, ContraNet surpasses the SOTA performance by 4% on MIT-States and 5.3% on Fashion200k.

CCS CONCEPTS

• **Information systems** → *Information retrieval*; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Compositional Learning, Multi modal, Contrastive Learning, CZSL

ACM Reference Format:
Muhammad Umer Anwaar (✉) *, Rayyan Ahmad Khan*, Zhihui Pan, Martin Kleinsteuber. 2021. A Contrastive Learning Approach for Compositional Zero-Shot Learning. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462244.3479904

* Equal Contribution.

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479904

1 INTRODUCTION

Objects, in the real world, exist in certain state(s). Human cognition is well set up to process vague concepts and easily categorize them and assign attributes (states) to objects. For instance, if a user wishes to buy a dress, their mind turns the attention towards what color, size, style, price etc., they would prefer. Traditional information retrieval (e.g. e-commerce websites) offer their customers a unimodal query, i.e., either a text or an image. After the query, the user has to select many filters to "help the algorithm" narrow down the options for the user. Future information retrieval (IR) systems should be smart enough to "help the customer" in expressing the concept in their mind by allowing a multi-modal query (see Fig. 1).

In this work, we focus our attention to a major aspect of such smart systems: learning good state-object representations. Specifically, (1) we aim to learn such models which understand different states of an object and can recognise even unseen combinations of them. (2) The model should be able to retrieve images based on multi-modal (image-text) query, where the text describes the changes sought by the user in the query image.

In the literature, these two tasks have been treated separately. The first task is referred to as compositional zero shot learning (CZSL). In contrast to image classification, the goal of CZSL classification is to simultaneously identify the class of the object and the state in which the object appears. Sometimes the visual differences of the same object in two states can be huge and that is where traditional classification methods fail. Several CZSL methods have been proposed to address this challenge. Li et al. [16] subject the learning of object embeddings to symmetry constraints under different state transformations. Misra et al. [19] maximize similarity of learned joint state-object embeddings with image features. Nagarajan et al. [20] treat states as operators (transformation matrices) and apply them to objects to yield the joint pair embeddings and then maximize the similarity with image. The second task is image retrieval based on a multi-modal query. Vo et al. [26] proposed the Text Image Residual Gating (TIRG) method for composing the query image and text for image retrieval. Anwaar et al. [1] proposed that the target image representation is an element-wise rotation of the representation of the source image in a complex space. The information about the degree of rotation is specified by the query text features.

The above mentioned methods do not focus on solving the two tasks in a systematic way. Interestingly, the state-of-the-art (SOTA) methods for CZSL task employ such data splits where all the objects and states are seen by the model during training. They refer to it as zero-shot because the test set contains images with pairs (state-object combinations) which were not seen in training. This will prove to be quite limiting for the real-world image retrieval application. Since the model is expected to have seen not only all the objects but also all the attributes the user can come up with. Naturally, a user can use different words which carry the same semantic meaning. Thus, a good model must have the ability to generalize to both unseen objects and unseen attributes. On the other hand, the second class of methods do not learn any object or state classification. They are focused on learning the fusion of query image and text for directly improving the image retrieval. This approach results in poor performance on the first task (see Sec.4.2).

In this work, we propose a unified approach, ContraNet, which bridges the existing gap in these two tasks. ContraNet aims to predict a composition of multiple semantic concepts in images. We adopt a contrastive learning approach to learn embeddings which are visually grounded and semantically meaningful. In recent years, contrastive learning has shown impressive results on a variety of tasks [4, 10, 25, 32]. Our rationale behind adopting contrastive learning is that current SOTA methods (discussed above), overwhelmingly rely on labels for learning. They overlook the fact that the underlying data lives on a much complicated manifold than what sparse labels could capture. Therefore, purely supervised methods converge to rigid solutions. In other words, they lead to good task-specific solutions, rather than learning the multiple semantic concepts in the data. ContraNet utilizes pretrained image and text models and then learn their mapping onto the multimodal embedding space via contrastive loss. That is, by maximizing the similarity between actual image-text pairs against random pairs through a bidirectional contrastive loss between the text and image modalities. These embeddings from this multimodal embedding space are then utilized for the downstream tasks. We use cross-entropy losses for the task 1 and soft-triplet loss for the task 2.

Despite the simplicity of our model, ContraNet outperforms the SOTA methods on benchmark datasets, namely: MIT-States, UT-Zappos and Fashion200k. Our experimental evaluation shows that projecting the text and image features onto a common embedding space and learning the representations via contrastive loss significantly enhances the performance of ContraNet.

Our main contributions are summarized below:

- We propose ContraNet, a unified contrastive learning approach which not only can recognize unseen combinations of state-object pairs but is also able to retrieve images of never seen objects based on multi-modal query.
- ContraNet outperforms the SOTA methods by a substantial margin on state-object composition zero-shot learning tasks. i.e., 8.7% on UT-Zappos and 8.1% on MIT-States on the best HM metric.
- For the image retrieval task, ContraNet utilizes the common embedding space learnt via contrastive loss and surpass the SOTA performance by 4% on MIT-States and 5.3% on Fashion200k dataset on Recall@1 metric.

2 RELATED WORK

In computer vision, a bulk of research centered around object recognition and classification, treats attribute as a mid-level "feature" learned from the visual patterns. This has proved amazingly successful in various tasks, e.g., zero-shot recognition, image description [7, 15, 17] and visual question answering [14].

Recently, there has been an increasing interest in compositional learning of attributes (states) and objects. In compositional learning, learning correct attributes is given the same importance as object prediction. That is, the model needs to predict the state-object pair. Several approaches have been proposed to tackle this problem. LabelEmbed (LE) proposed by [19] uses GloVe vectors[22] for state and object. They employ a 3-layered MLP to transform the word

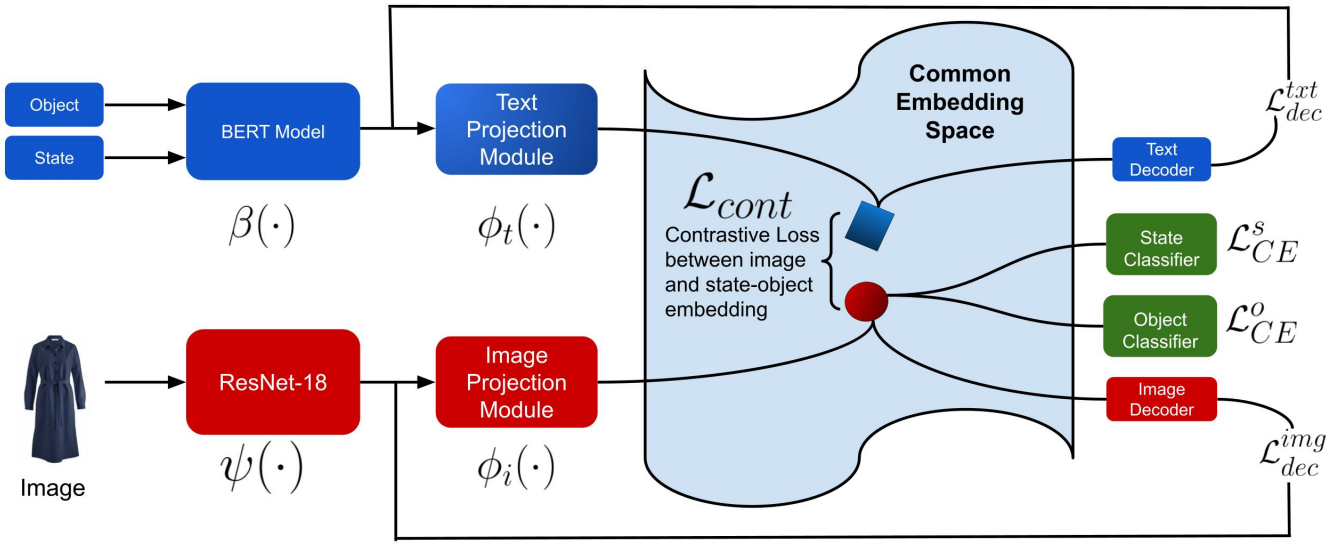

Figure 2: ContraNet Architecture: Learning the compositional labels via contrastive learning

embeddings into a transformation matrix. The prediction of the classifier is obtained by the product of transformation matrix with image features. AnalogousAttr [3] trains several linear classifiers for seen compositions and then leverages tensor completion techniques to do predictions for the unseen pairs. AoP [20] uses GloVe vectors for objects. But they consider states (attributes) as linear transformation matrices, which "operate" on the objects to yield pair embeddings. The pair embedding with the minimum distance to image embedding in the joint embedding space is the prediction of the model. Red Wine is another method proposed by [19]. It replaces the GloVe vectors in LE with the SVM weights. TAFE [29] employs word2vec vectors [18] of state-object composition pair to generate binary classifier for each composition. Task-driven Modular Networks (TMN) [23] configures a set of modules (fully connected layers operating in semantic concept space) through a gating function in a task-driven way. It generalizes to unseen compositions by re-weighting these primitive modules. SymNet [16] learns object embeddings showing symmetry under different state transformations. They emphasize that leaning in such a fashion yields better embeddings for the compositional learning task.

The task of image retrieval based on multimodal query has also been explored in literature. Two state of the art methods are: TIRG [26] and ComposeAE [1]. In TIRG, the authors employ gated feature connection in order to keep the composed representation of query image and text in the same space as that of the target image. They also incorporate a residual connection which learns the similarity between concatenation of image-text features and the target image features. The authors of ComposeAE [1] argue that the source image and the target image lie in a common complex space. They are rotations of each other and the degree of rotation is encoded via query text features. Some other approaches which are also closely related to this task are interactive image retrieval task [8, 24] and attribute-based product retrieval task [33]. These approaches have their limitations such as that the query texts are limited to a fixed set of relative attributes [33], require multiple rounds of natural language queries as input [8, 24] or that query texts can be only one word i.e. an attribute [9]. Unlike our task, the focus of these methods is on modeling the interaction between user and the agent.

3 APPROACH

3.1 Problem Setting and Overview

Let $\mathcal{X}$ denote the set of images, $\mathcal{S}$ denote the set of states, $\mathcal{O}$ the set of objects and $\mathcal{T} = \mathcal{S} \times \mathcal{O}$ denote the set of composition labels. Each image x is associated with a compositional label $t = (s, o)$. We tackle the following two tasks:

(1) Prediction of composition label (state, object) for a given image. During testing, most of the composition labels are novel i.e. not seen during training. Hence, this task is also called Compositional Zero Shot Learning Task. The model, $f : \mathcal{X} \rightarrow \mathcal{T}^{test}$, is trained to maximize the number of correct predictions of composition labels.
(2) Image Retrieval based on a multi-modal (image-text) query. Specifically, the query text prompts some *state* modification in to the query image x. The task is to retrieve images with same object label o as in the query image but with the desired state label s. The model, $g : (\mathcal{X}, \mathcal{T}) \rightarrow \mathcal{X}^{target}$, is trained to maximize the similarity between composed representation of (image-text) query and the target image representation.

3.2 Task # 1: Learning the Compositional Zero Short Prediction

ContraNet is an autoencoder based approach to learn composition of multiple semantic concepts in images. In this section, we discuss the architecture of ContraNet and the loss functions involved in CZSL task. Fig. 2 presents an overview of ContraNet.

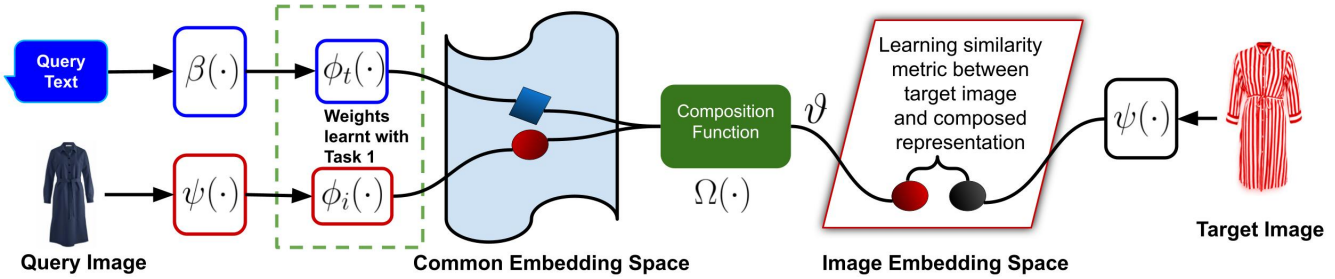

Figure 3: ContraNet Architecture: Image retrieval based on multi-modal query. The weights of the projection modules learned during task 1 are frozen.

In the figure, $\psi(\cdot)$ denotes the pre-trained image model (e.g. ResNet-18), which takes an image as input and returns image features in a d-dimensional space. Analogously, $\beta(\cdot)$ denotes the pre-trained text model (e.g. BERT), which takes an text as input and returns text features in a h-dimensional space. It is to be noted that $\beta(\cdot)$ takes as input both the state and object text and returns a single h-dimensional feature vector. Typically in image-text joint embeddings [26, 28], these features are combined using fully connected layers or gating mechanisms. In contrast to this, we project these features onto a common multi-modal embedding space via separate projection modules. These modules are denoted by $\phi_t : \mathbb{R}^h \mapsto \mathbb{R}^k$ and $\phi_i : \mathbb{R}^d \mapsto \mathbb{R}^k$ for text and image respectively.

In this common embedding space, we aim to learn such representations which better capture the underlying cross-modal dependencies. ContraNet utilizes the text information to learn multiple semantics present in the images. We adopt a deep metric learning approach (DML) to train ContraNetin a contrastive fashion. During training, we sample a batch of B input image-text pairs (x, t). We calculate the similarities between the representations of images and texts in the common embedding space and then calculate the contrastive loss, $\mathcal{L}_{cont}$, as follows:

$$\Delta_t = \Big\|_{j=1}^{B} \phi_t(\beta(t_j)), \tag{1}$$

$$\Delta_i = \Big\|_{j=1}^{B} \phi_i(\psi(x_j)), \tag{2}$$

$$\mathcal{E} = \Delta_t * \Delta_i^T, \tag{3}$$

$$\mathcal{L}_{cont} = \frac{1}{2B}\sum_{j=1}^{B} -\log\left\{\frac{\exp\{\mathcal{E}_{jj}\}}{\sum_{p=1}^{B}\exp\{\mathcal{E}_{jp}\}}\right\} - \log\left\{\frac{\exp\{\mathcal{E}_{jj}\}}{\sum_{p=1}^{B}\exp\{\mathcal{E}_{pj}\}}\right\}, \tag{4}$$

where $\|$ denotes the concatenation operation of representation for each sample of the batch, T denotes transpose of a matrix, $*$ denotes matrix multiplication, exp denotes exponential function and $\mathcal{E}$ denotes the matrix of similarities of all image-text pairs in a batch. The first term in the loss function corresponds to text-to-image contrastive loss and the second term corresponds to image-to-text contrastive loss.

As for the decoder part of ContraNet, we learn two separate decoders from the representations embedded in the multi-modal embedding space. i.e., image decoder and text decoder denoted by d_{img} and d_{txt} respectively. The reason for using the decoders and reconstruction losses is two-fold: first, it acts as regularizer on learning of the embeddings and secondly, it forces the model to retain relevant text and image information in the common embedding space. Empirically, we have seen that these losses reduce the variation in the performance and aid in preventing overfitting.

Thus, we also add two reconstruction losses, $\mathcal{L}_{dec}^{img}$ and $\mathcal{L}_{dec}^{txt}$, in our training objective, each corresponding to d_{img} and d_{txt} respectively. They are given by:

$$\mathcal{L}_{dec}^{img} = \frac{1}{B}\sum_{i=j}^{B}\left\|\psi(x_j) - \hat{z}_i\right\|_2^2, \tag{5}$$

$$\mathcal{L}_{dec}^{txt} = \frac{1}{B}\sum_{i=j}^{B}\left\|\beta(t_j) - \hat{q}_i\right\|_2^2, \tag{6}$$

where $\hat{z}_i = d_{img}(\cdot)$ and $\hat{q}_i = d_{txt}(\cdot)$.

For this task, during inference, we have to predict the composition label (state, object) for a given image. Thus, ContraNet trains two separate classifiers for state and object classification. These classifiers take as input the image representation projected onto the common embedding space. They are implemented as two fully-connected (FC) layers, followed by a softmax layer. They are trained with cross-entropy losses, denoted as $\mathcal{L}_{CE}^{S}$ and $\mathcal{L}_{CE}^{O}$, for state and object classification.

The total loss is computed by the weighted sum of above mentioned losses. It is given by:

$$\mathcal{L} = \mathcal{L}_{cont} + \lambda_i\,\mathcal{L}_{dec}^{img} + \lambda_t\,\mathcal{L}_{dec}^{txt} + \lambda_o\,\mathcal{L}_{CE}^{O} + \lambda_s\,\mathcal{L}_{CE}^{S}, \tag{7}$$

3.3 Task # 2: Image Retrieval Based on a Multi-Modal Query

As discussed in Sec. 3.1, the goal in this task is to retrieve images which have same object label as the query image but possess the state as desired by query text. For instance, we have input image of an "unripe tomato" and the text prompts to retrieve images with the "ripe" state (see Fig. 5). Fig. 3 presents the modified ContraNet architecture for this task. Since the goal is to retrieve images, we hypothesise that it is better to compose the modalities in the multi-modal query and map them to the image embedding space. This ensures that the image and compositional features are "aligned". In this way, during inference, we can simply compare the composed features with the test images in the image embedding space.

Method	MIT-States						UT-Zappos					
	Attribute	Object	Seen	Unseen	HM	AUC	Attribute	Object	Seen	Unseen	HM	AUC
AoP	21.1	23.6	14.3	17.4	9.9	1.6	38.9	**69.9**	59.8	54.2	40.8	25.9
Red Wine	22.7	25.1	20.7	17.9	11.6	2.4	40.6	69.1	53.6	52.1	41.3	26.1
LabelEmbed	23.5	26.3	15.0	20.1	10.7	2.0	41.2	69.3	53.0	61.9	41.0	25.7
ComposeAE	23.8	26.4	21.4	22.6	14.9	2.7	41.9	68.8	57.2	58.9	44.2	29.2
TMN	23.3	26.5	20.2	20.1	13.0	2.9	40.8	69.2	58.7	60.0	45.0	29.3
SymNet	24.3	**27.3**	24.2	25.2	16.1	3.0	41.3	68.6	49.8	57.4	40.4	23.4
ContraNet	**28.9**	26.7	**28.1**	**27.4**	**17.4**	**4.7**	**52.7**	68.1	**60.7**	**62.5**	**48.9**	**34.7**
- without $\mathcal{L}_{cont}$	22.9	27.1	14.8	18.6	9.7	1.9	42.4	69.2	54.9	52.6	42.9	27.4
- without $\mathcal{L}_{dec}$	28.2	26.5	27.5	26.8	17.2	3.9	51.4	66.7	58.4	60.8	47.2	33.1

Table 1: Performance comparison on the Generalized CZSL split. The best performance is in bold and the second best is underlined.

For this task, we learn a composition function, $\Omega : \mathbb{R}^{2k} \mapsto \mathbb{R}^d$, for composing the embeddings of query image and text coming from the common embedding space. This function maps the embeddings directly in to the image embedding space. Lets denote the output of $\Omega(\cdot)$ by ϑ. In our experiments, we work with three different variants of $\Omega(\cdot)$, namely: MLP, residual gating and image rotation based on text (see Sec. 4.4 for details.)

We employ triplet loss with soft margin, $\mathcal{L}_{trip}$, for learning this composition function. The loss aims to maximize the similarity between the composed features ϑ and the target image features $\psi(y)$ extracted from the image model. It also pushes the composed representation ϑ away from non-similar images in $\mathbb{R}^d$. Let $\kappa(\cdot,\cdot)$ denote the similarity kernel, which we implement as a dot product between its inputs. The loss is given by:

$$\mathcal{L}_{trip} = \frac{1}{MB}\sum_{j=1}^{B}\sum_{m=1}^{M}\log\Big\{1+\exp\{\kappa(\vartheta_j,\psi(\tilde{y}_{j,m})) \; - \; \kappa(\vartheta_j,\psi(y_j))\}\Big\}, \tag{8}$$

where M denotes the number of triplets for each sample j and $\psi(\tilde{y}_j)$ denote the randomly selected negative image from the batch. This is equivalent to the the soft triplet based loss used in [11, 27].

4 EXPERIMENTS

4.1 Experimental Setup

Datasets: In our experiments, we use three benchmark datasets, namely: MIT-States[12], UT Zappos [31] and Fashion200k[9]. MIT-States consists of 53753 diverse real-world images where each image is described by an object-attribute composition label, i.e. an attribute (state) and a noun (object), e.g. "ripe tomato". There are 245 objects, 115 attributes and 1962 possible pairs. UT-Zappos is a dataset of only shoes with fine-grained annotations. A composition label consists of shoe type-material pair. There are 12 shoe types (objects), 16 different materials (states) and 116 possible pairs.

Two different settings of these benchmark datasets are proposed in the literature for evaluating the models. (1) Generalized CZSL (GCZSL) split [23]. Following Chao et al. [2], they argue that performance of the model should also be evaluated on seen pairs. In this setting, MIT-States utilizes 1262 pairs (the seen pairs) for training, whereas the test set has 400 seen and 400 unseen pairs. This split also contains a validation set consisting of 300 seen and 300 unseen pairs. The UT-Zappos dataset makes use of 83 pairs (the seen pairs) for training, whereas the test set has 18 seen and 18 unseen pairs. The validation set has of 15 seen and 15 unseen pairs. (2) CZSL split [19, 20], which ensures that there is no overlap between pairs in the train and test set. For MIT-states, the train set still consists of 1262 pairs/34562 images and the test set now has 700 pairs/19191 images. For UT-Zappos, the train set still has 83 pairs/24898 images and the test set now contains 33 pairs/4228 images.

The third dataset Fashion200k [9] is only used for the second task. It consists of ∼200k images of different fashion categories. Each image has a human annotated caption, e.g. "beige bolero jacket dress". In order to ensure fair comparison, for the second task, we also follow the same train-test split as proposed by our competitors (TIRG [26] and ComposeAE [1]) for both Fashion200k and MIT-States. Now the split for MIT-States is that 196 objects out of total 245 are reserved for training and the rest 49 objects are used during test. This split ensures that there is no overlap between training and testing queries in terms of objects.

Implementation Details: Following [16, 23, 29], we compare the results of ContraNet with several baselines as well as previous state-of-the-art (SOTA) methods. Unless otherwise specified, the default image feature extractor for all methods is ResNet-18. We also extract 512-dimensional image features using ResNet18 pre-trained on ImageNet [5]. ContraNet employs pretrained BERT [6] for encoding texts. Concretely, we employ BERT-as-service [30] and use Uncased BERT-Base++ which outputs a 1024-dimensional feature vector. To ensure fair comparison, we employ the same strategy for hyperparameter tuning as SymNet [16]. We use cross-validation to determine the hyper-parameters, e.g., learning rate, weights, epochs. We employ Adam [13] optimizer. For both datasets, $M = 3$ in $\mathcal{L}_{trip}$ and the weights of the losses are: $\lambda_s = \lambda_o = 1$ and $\lambda_i = \lambda_t = 0.1$. We repeat each experiment 10 times and report the average performance of the models.

Metrics: For the task of CZSL, we follow [16, 20] and report Top-1, 2, 3 accuracies on the unseen test set as evaluation metrics. For the task of GCZSL, we follow the evaluation protocol from TMN [23] and use the same metrics as proposed by them. Namely: best accuracy on only images of seen/unseen compositions (*best seen/best unseen*), best harmonic mean (*best HM*) and Area Under the Curve (AUC) for seen and unseen accuracies by varying the bias values. For the task of image retrieval, following the evaluation protocol

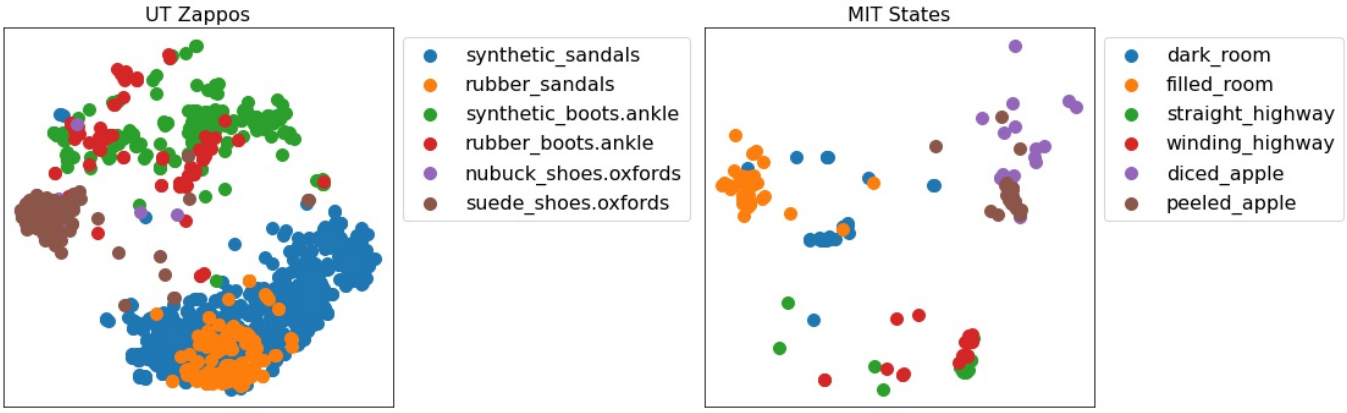

Figure 4: Visualization (using t-SNE) of test image instances projected unto the learned common embedding space

from TIRG [26], we use recall at rank k ($R@k$), as our evaluation metric. $R@k$ estimates the proportion of queries where the target (ground truth) image is within the top k retrieved images.

4.2 Discussion of Results for the GCZSL and CZSL Tasks

Table 1 summarizes the results of the performance comparison on the GCZSL task. In the following, we discuss these results to gain some important insights into the problem.

First, we note that our proposed method ContraNet consistently outperforms the SOTA methods by a significant margin on both benchmark datasets. Specifically, on UT-Zappos in terms of AUC metric, the performance improvement over the second best method (TMN) is 18.4% and 32.9% over the third best method (RedWine). Similarly, for best HM metric on UT-Zappos, ContraNet outperforms the second best method (TMN) by 8.67% and by 15.53% over the third best method (RedWine). On MIT-States dataset, ContraNet still outperforms the competitive methods but the margins are less than those on UT-Zappos dataset. For instance, in terms of AUC and best HM metric, the performance improvement over the second best method (SymNET) is 3.6% and 8.1% respectively.

Second, we observe a very interesting pattern for the object classification accuracy metric. On UT-Zappos dataset, the top-2 performing methods (AoP and LabelEmbed+) are fairly simple models in comparison to more recent models like SymNET and TMN. For the MIT-States, this pattern is reversed and AoP and LabelEmbed+ are performing worse than advance models. Here, SymNET and ContraNet achieves best performance with insignificant difference of 0.7%. We hypothesise that simpler model are better able to capture the object class for UT-Zappos. We test this conjecture by removing $\phi(\cdot)$'s and $\mathcal{L}_{cont}$ from ContraNet. The results support this conjecture as the performance is improved by 1.1 and 0.4 for ContraNet without $\mathcal{L}_{cont}$ on UT-Zappos and MIT-States respectively. Detailed investigation of this behavior is out of scope for this work.

Third, we note that all the methods perform better on UT-Zappos as compared to MIT-States. This is due to differences in the underlying distributions of the two datasets, which makes one dataset more "difficult" to learn. Another reason is that MIT-states dataset was automatically labeled, which leaves the room for more incorrect labels. Our quantitative analysis (Sec. 4.5) provides a better view of this issue. Briefly, even though the algorithm retrieves semantically similar images but they will not be considered correct due to noisy labelling. For instance, for the second query in Fig. 5, we can see that the second and fifth image are semantically similar. But according to automated labels, only the first and third images are "correct images". This issue has also been discussed in depth by Nawaz et al.[21].

Method	MIT-States			UT-Zappos		
	Top-1	Top-2	Top-3	Top-1	Top-2	Top-3
AnalogousAttr	1.4	-	-	18.3	-	-
LabelEmbed	13.4	17.6	22.4	25.8	39.8	52.4
Red Wine	13.1	21.2	27.6	40.3	52.8	67.1
AoP	14.2	19.6	25.1	46.2	56.6	69.2
TAFE-Net	16.4	26.4	33.0	33.2	45.8	57.3
SymNet	19.9	28.2	33.8	52.1	67.8	76.0
ContraNet	**22.1**	**33.7**	**38.2**	**54.6**	**73.1**	**80.4**
- without $\mathcal{L}_{cont}$	14.7	18.8	24.6	39.9	50.2	62.4
- without $\mathcal{L}_{dec}$	<u>20.4</u>	<u>31.2</u>	<u>35.9</u>	<u>53.3</u>	<u>69.7</u>	<u>78.6</u>

Table 2: Performance comparison on the CZSL split. The best performance is in bold and the second best is underlined.

Table 2 presents the results of CZSL task. We observe that ContraNet significantly outperforms all the competitive methods. The second best method is also ContraNet without $\mathcal{L}_{dec}$. In this variant, we remove both the image and text decoder of our model. Without these "regularizers" on the common embedding space, there is slight drop in performance but still $\mathcal{L}_{cont}$ and classification losses are able to maintain a decent performance.

Finally, we observe from Tables 1 and 2, that the variant of ContraNet without $\mathcal{L}_{cont}$ results in a huge degradation in the performance. This reinforces our claim that contrastive learning of image and text embeddings in a common space helps in learning better representations. These representations are then particularly useful for the compositional learning tasks.

4.3 Visualization in Latent Space

We plot the representations of images projected unto the common embedding space to visualize how well they are separated with respect to their labels. Fig.4 presents these visualizations using t-SNE for the UT-Zappos and MIT-States datasets. It shows that compositions with similar state-object are closer to each other than other compositions.

4.4 Discussion of Results for the Image Retrieval Based on Multi-modal Query Task

Tables 3 and 4 present the results of this task for MIT-States and Fashion200k respectively. Depending on the composition function Ω, we get three variants of ContraNet. Ω_{res} means that we compose the modalities via the gated and residual connections introduced by TIRG. Ω_{rot} denotes that composition is modelled as rotation of image embeddings from common embedding space. This idea was proposed by ComposeAE[1]. However, unlike them, we do not employ any rotational symmetry loss or complex rotation. The third variant, Ω_{mlp}, simply means that we fuse the two modalities via MLP (two-fully connected layers with non-linear activations).

First, we note that the variant ContraNet-Ω_{res} achieves impressive gains over all the competitive methods. Specifically, on MIT-Sates in terms of R@10 metric, the performance improvement over the second best method (ComposeAE) is 7.3% and 19.3% over the third best method (TIRG). Similar performance trend can be seen for fashion200k dataset. This supports our claim that the semantic meaning hidden in the labels helps in learning better composition.

Second, we observe that several methods (like AoP, SymNet, FiLM) could not outperform simple baseline like concatenation of raw image and text features. While TIRG, ComposeAE and ContraNet consistently outperforms this baseline by a significant margin on both benchmark datasets.

Third, we note that all the variants of ContraNet outperform or achieve comparable performance to ComposeAE and TIRG. Although ContraNet-Ω_{res} comes out to be the best variant, but ContraNet-Ω_{mlp} and ContraNet-Ω_{rot} are also able to perform the task reasonably well. We hypothesize that the reason is that the projection modules to the common embedding space learned in the task 1 have learned the underlying shared information between two modalities quite well. Thus, a simple composition function like MLP is also able to achieve competitive results. In order to confirm this intuition, we drop the projection modules to the common embedding space from our best performing variant, i.e., ContraNet-Ω_{res} without Common Embedding Space. Consequently, we observe an enormous drop in performance on both datasets.

4.5 Qualitative Results

Fig.5 presents some qualitative retrieval results for MIT-States dataset. For the first query, we see that three correct "barren mountain" images are retrieved, shown with green boundary. We can observe that other retrieved images share the same semantics and are visually similar to the target images. In second query, we note that same objects in different states can look drastically different. This highlights the importance of incorporating the text information in the composed representation. Some qualitative retrieval results for Fashion200k dataset are presented in Fig. 6. In these results, we observe that the model is able to capture the style and color information quite well. In the first row, we see similar black floral blouses. Similarly, in the second query, the model successfully images from the same product category, i.e. skirts. Moreover, the retrieved images seem to follow the desired modifications expressed in the query text remarkably well. It is pertinent to highlight that the captions under the images are the ground truth. They are not available to the model as input during training or inference.

Method	R@1	R@5	R@10
Raw Image features only	3.3	12.8	20.9
Raw Text features only	7.4	21.5	32.7
Concatenation [Image,Text]	11.8	30.8	42.1
Show and Tell	11.9	31.0	42.0
AoP	8.8	27.3	39.1
Relationship	12.3	31.9	42.9
FiLM	10.1	27.7	38.3
SymNet	11.2	29.5	41.4
TIRG	12.2	31.9	43.1
ComposeAE	<u>13.9</u>	<u>35.3</u>	<u>47.9</u>
ContraNet-Ω_{res}	**14.5**	**40.7**	**51.4**
ContraNet-Ω_{rot}	13.9	39.1	50.8
ContraNet-Ω_{mlp}	13.7	36.9	48.8
ContraNet-Ω_{res} without Common Embedding Space	12.0	31.2	42.9

Table 3: Model performance comparison on MIT-States. The best number is in bold and the second best is underlined.

Method	R@1	R@10	R@50
Raw Image features only	3.5	22.7	43.7
Raw Text features only	1.0	12.3	21.8
Concatenation [Image,Text]	11.9	39.7	62.6
Show and Tell	12.3	40.2	61.8
Param Hashing	12.2	40.0	61.7
Relationship	13.0	40.5	62.4
FiLM	12.9	39.5	61.9
SymNet	11.7	38.6	60.4
TIRG	14.1	42.5	63.8
ComposeAE	<u>22.8</u>	<u>55.3</u>	<u>73.4</u>
ContraNet-Ω_{res}	**24.0**	**58.4**	**79.2**
ContraNet-Ω_{rot}	18.5	54.8	76.3
ContraNet-Ω_{mlp}	22.9	56.7	77.5
ContraNet-Ω_{res} without Common Embedding Space	17.8	50.6	71.1

Table 4: Model performance comparison on Fashion200k. The best number is in bold and the second best is underlined.

5 CONCLUSION

In this work, we propose ContraNet, a novel approach to learn text-aware image representations in a common embedding space. The core idea is to ensure, in a contrastive manner, that the (state, object, image) triples are distinguishable across the training batches.

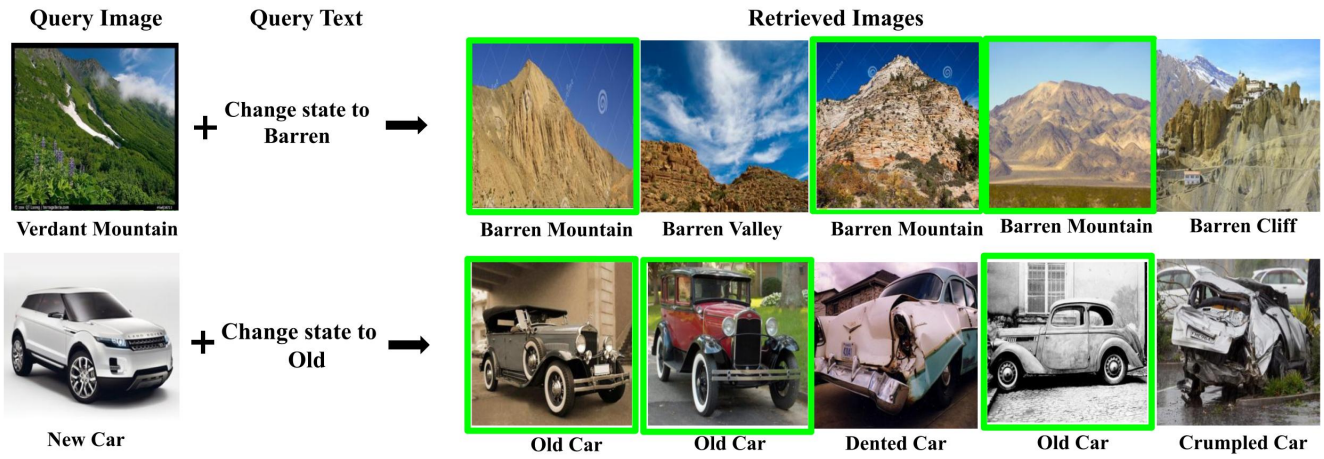

Figure 5: Qualitative results for image retrieval task: MIT-States

Figure 6: Qualitative results for image retrieval task: Fashion200k

This prompts ContraNet to learn efficient embeddings by simultaneously leveraging the information in multiple modalities. The contrastive loss is complemented by image and text reconstruction losses that not only regularize the learnt embedding but also attempt to preserve the information in the individual modalities. The resultant architecture achieves SOTA performance on generalized CZSL as well as multi-modal query-based image retrieval task, as demonstrated by the extensive comparison with many competitive baselines on three popular benchmark datasets.

ACKNOWLEDGMENTS

This work has been supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy through the *WoWNet* project IUK-1902-003// IUK625/002.

REFERENCES

[1] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. 2021. Compositional Learning of Image-Text Query for Image Retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1140–1149.

[2] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. 2017. An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild. arXiv:1605.04253 [cs.CV]

[3] Chao-Yeh Chen and Kristen Grauman. 2014. Inferring Analogous Attributes. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*. IEEE Computer Society, USA, 200–207. https://doi.org/10.1109/CVPR.2014.33

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709 [cs.LG]

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[7] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 1778–1785. https://doi.org/10.1109/CVPR.2009.5206772

[8] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*. 678–688.

[9] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE International Conference on Computer Vision*. 1463–1471.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. arXiv:1911.05722 [cs.CV]

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).

[12] Phillip Isola, Joseph J Lim, and Edward H Adelson. 2015. Discovering states and transformations in image collections. In *CVPR*.

[13] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR* (2015).

[14] Jayanth Koushik, Hiroaki Hayashi, and Devendra Singh Sachan. 2017. Compositional Reasoning for Visual Question Answering. In *Proceedings of the 34 th International Conference on Machine Learning, 2017*.

[15] Neeraj Kumar, Peter Belhumeur, and Shree Nayar. 2008. FaceTracer: A Search Engine for Large Collections of Images with Faces. In *Computer Vision – ECCV 2008*. Springer Berlin Heidelberg, Berlin, Heidelberg, 340–353.

[16] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. 2020. Symmetry and Group in Attribute-Object Compositions. In *CVPR*.

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 3730–3738. https://doi.org/10.1109/ICCV.2015.425

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*.

[19] Ishan Misra, Abhinav Gupta, and Martial Hebert. 2017. From red wine to red tomato: Composition with context. In *CVPR*.

[20] Tushar Nagarajan and Kristen Grauman. 2018. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*.

[21] Shah Nawaz, Kamran Janjua, Ignazio Gallo, Arif Mahmood, Alessandro Calefati, and Faisal Shafait. 2019. Do Cross Modal Systems Leverage Semantic Relationships? arXiv:1909.01976 [cs.CV]

[22] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[23] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. 2019. Task-driven modular networks for zero-shot compositional learning. In *ICCV*.

[24] Fuwen Tan, Paola Cascante-Bonilla, Xiaoxiao Guo, Hui Wu, Song Feng, and Vicente Ordonez. 2019. Drill-down: Interactive Retrieval of Complex Scenes using Natural Language Queries. In *Advances in Neural Information Processing Systems*. 2647–2657.

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748 [cs.LG]

[26] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*.

[27] Nam N Vo and James Hays. 2016. Localizing and orienting street views using overhead imagery. In *European Conference on Computer Vision*. Springer, 494–509.

[28] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE CVPR*. 5005–5013.

[29] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E. Gonzalez. 2019. TAFE-Net: Task-Aware Feature Embeddings for Low Shot Learning. arXiv:1904.05967 [cs.CV]

[30] Han Xiao. 2018. bert-as-service. https://github.com/hanxiao/bert-as-service.

[31] Aron Yu and Kristen Grauman. 2017. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *CVPR*.

[32] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2020. Contrastive Learning of Medical Visual Representations from Paired Images and Text. arXiv:2010.00747 [cs.CV]

[33] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. 2017. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1520–1528.