

Embedding Arithmetic of Multimodal Queries for Image Retrieval

Guillaume Couairon
 Meta AI, Sorbonne Université
 gcouairon@fb.com

Matthijs Douze
 Meta AI
 matthijs@fb.com

Matthieu Cord
 Valeo, Sorbonne Université
 matthieu.cord@sorbonne-universite.fr

Holger Schwenk
 Meta AI
 schwenk@fb.com

Abstract

Latent text representations exhibit geometric regularities, such as the famous analogy: *queen is to king what woman is to man*. Such structured semantic relations were not demonstrated on image representations. Recent works aiming at bridging this semantic gap embed images and text into a multimodal space, enabling the transfer of text-defined transformations to the image modality.

We introduce the SIMAT dataset to evaluate the task of Image Retrieval with Multimodal queries. SIMAT contains 6k images and 18k textual transformation queries that aim at either replacing scene elements or changing pairwise relationships between scene elements. The goal is to retrieve an image consistent with the (source image, text transformation) query. We use an image/text matching oracle (OSCAR) to assess whether the image transformation is successful. The SIMAT dataset will be publicly available.

We use SIMAT to evaluate the geometric properties of multimodal embedding spaces trained with an image/text matching objective, like CLIP. We show that vanilla CLIP embeddings are not very well suited to transform images with delta vectors, but that a simple finetuning on the COCO dataset can bring dramatic improvements. We also study whether it is beneficial to leverage pretrained universal sentence encoders (FastText, LASER and LaBSE).

1. Introduction

Many works aim at learning a multimodal space in which images and text can be embedded and compared [4, 5, 27, 30]. This is especially useful for image/text retrieval, where given a sample from one modality, one has to find a corresponding sample of the other modality in a database. Such embeddings have other interesting applications. It has been shown in [11, 25] that textual input can be used to *replace* scene elements in images or change their properties with simple

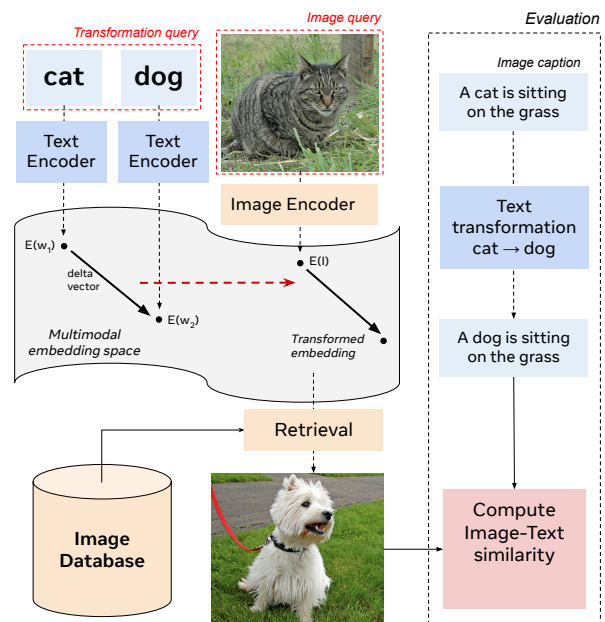


Figure 1. We aim at performing image retrieval, guided by an input image and a text query describing how the input image should be transformed. The transformation is mapped to a *delta vector*, added to the image embedding to produce a *transformed embedding*, for which a corresponding image is retrieved in a database. The evaluation module checks that the text-transformed caption is valid for the image result.

arithmetic operations in the latent space, combined with nearest neighbour search.

In this work, we study the following problem: given an image and a transformation query formulated in the text domain, the task is to find an image that satisfies the transformation query while being semantically as similar as possible to the input image.

For example, with the *cat*→*dog* transformation, an image showing a cat sitting in the grass should be transformed into

an image with a dog sitting in the grass (Figure 1).

Evaluation challenges. We seek to create a dataset to evaluate algorithms on this task. Such a dataset should contain feasible (image, text transformation) queries: the transformation *man*→*dog* can be applied to an image with “A man is running on the beach”, but not to “A man is speaking on the phone”. We create a corpus from Visual Genome images and annotations [13] and ensure that this requirement is met. Evaluating semantic text-to-image transfer is challenging: First, we need to ensure that the requested transfer was performed (the cat was replaced by a dog). Then, we need to verify that the modification is minimal: the dog should be sitting on grass and ideally, all other visual elements should not be changed. We use OSCAR [15] as an external oracle to assess whether these two conditions are met. OSCAR has been trained on captioned images with a binary cross-entropy loss to recognize whether or not a given text corresponds to an image.

Embedding arithmetic of multimodal queries. We choose to transform images by encoding the transformation query as a *delta vector* in the multimodal space, before adding it to an image embedding and retrieving the closest image in a database (see Figure 1). This operation solely relies on the image/text alignment without needing any transformation example. However, it requires a well structured multimodal space to be able to transfer text transformations to images. We know that word and sentence embeddings trained on vast amounts of data have been shown to possess geometric properties that can be useful for text transformation [17, 18]. Previous work [11] has hinted that such geometric properties could also be present in multimodal spaces, without quantitative evidence.

In this paper, we study the suitability of multimodal embedding spaces like CLIP [21] to perform image retrieval with image-text queries. The CLIP multimodal space was trained with a contrastive image/text matching loss [23] on 400M image-text pairs crawled from the internet, which makes it suitable for use in a zero-shot evaluation setting like ours. Given that word and sentence embeddings have been shown to display strong geometric properties [17, 18], we study whether it is necessary to use them as building blocks to multimodal embeddings better suited to image retrieval with multimodal queries. In particular we use LASER [2] and LaBSE [6], which have been pre-trained on large corpora of multilingual data.

The contributions of this work are:

- SIMAT, a dataset of 6 000 images and 18 000 transformation queries, to evaluate algorithms on the task of image retrieval with multimodal queries, which comes with an evaluation metric based on OSCAR.
- detailed experiments to measure which multimodal embeddings work best within the *delta vectors* framework to transform images.

2. Related Work

The geometric properties of word embeddings have been observed notably in [7, 18], with the famous example *king* is to *queen* what *man* is to *woman*. These properties have also been studied for sentence embeddings [17]. Recent work has demonstrated that state-of-the-art multimodal embeddings can be obtained by scaling up image-text alignment pretraining [11, 21]. It has been observed that such embeddings exhibit properties similar to word embedding analogies, allowing textual and image queries to be combined. A few examples are shown in [11], but without any quantified evaluation.

Image Retrieval with Multimodal Queries. In general, Image Retrieval with Multimodal Queries is a form of image retrieval where some textual inputs serve as instructions to modify an existing image through retrieval. The instruction is a simple word pair in our case, but it can also be a more complex instruction in natural language. Several datasets exist to tackle this problem: The CSS dataset [26] which is a synthetic dataset with simple colored geometrical objects based on CLEVR [12]. The Fashion200k dataset [9] provides around 200k images of fashion products, each annotated with a compact attribute description. Similarly, the Fashion-IQ dataset [8] was built to advance research on interactive fashion image retrieval. The MIT-States dataset [10], also commonly used, is a dataset of $\approx 60k$ images, each annotated with an object/noun label and a state/adjective label such as *new car* or *broken window*. Those datasets are designed to evaluate image retrieval with multimodal queries on narrow domains, which gives more control over what attributes can be changed and ensures that the transformation is always feasible. Another common characteristic is the focus on changing object properties rather than objects relationships. We focus on more realistic images, and study object transformations where an object should be replaced by another without changing the high-level subject-object interaction.

Methods for solving this image retrieval task [1, 24, 26] all focus on supervised learning: a fraction of the dataset is used for training and the remaining for testing. Instead, we want to measure if multimodal embeddings trained with an image/text matching objective can be used to solve this task without any transformation example.

Text-driven Image Editing. Some work focus on directly modifying the pixels of query images instead of performing the retrieval step [22]. [29] encode images as a graph of interacting objects which lets the user modify an image by editing its scene graph. GANs are also frequently used to modify images based on some natural language input [14, 19, 28]. Lastly, CLIP [21] can be used in combination with a StyleGAN generator to make semantic edits in images, as exemplified in [20].

3. The SIMAT database

3.1. Requirements

First, we need a list of images with some transformation queries (e.g. a man sleeping on the beach, with the query *man*→*woman*). We want simple images (so that the query is unambiguous) and relevant transformation queries. Second, we need a database of images that we will use for the retrieval step. And finally, we need a criterion to decide, based on the retrieved image, if the transformation is successful or not. It is the case if only the element designated by the transformation query has changed, while keeping the rest of visual elements as similar as possible.

Previous work [25] has tried to solve these requirements with a dataset of $\approx 1,500$ images from Google Image Search queries, dubbed SIC112, with each image annotated with a actor-action-environment triplet, such as (*woman, walking, street*), among a set of 112 possible triplets. Transformation queries then consist in changing either the subject, action or environment. This set of images is also used as a database for retrieval, which has two advantages: (i) transformation queries are always possible by design of the dataset, and (ii) the quality of the retrieved image is measured by checking if its annotation triplet was indeed the one expected by the transformation query.

We scale this approach to a larger number of annotation triplets, that take the more general form of (subject, relationship, object). However, we observed that due to the larger triplet vocabulary, images can be accurately described by multiple such triplets, which skews the evaluation metric: an image would often be rejected for not being annotated with the expected triplet while still being visually correct. Therefore, we chose to use a different metric for evaluating the quality of transformed images: we evaluate whether the semantic transformation is successful by querying OSCAR [15]. OSCAR computes the probability $\mathbb{P}_O(I, T)$ that a caption T accurately described an image I , based on the concatenation of the text tokens in T and the object tags and features detected by faster R-CNN on image I (we provide the triplet to OSCAR in the form of a caption written in natural language). Note that this OSCAR-based evaluation method does not involve image annotations in the retrieval database and thus could potentially be applied to a much larger database of non-annotated images, which we leave for future work.

3.2. Construction

Similarly to [25], we create a list of images annotated with (subject, relationship, object) triplets, and perform the retrieval step inside the same list of images to ensure that transformation queries always have a valid solution in the dataset. We start from annotations from the Visual Genome dataset [13]. Each image in the dataset contains a list of such

triplets with subject and object bounding boxes, which we use to crop square regions of images that minimally contain the subject and object in the image. We then filter this list and compute possible transformations:

Subject/Relation filtering. Only keep triplets for which the subject is a human or animal, and the relation is a non-positional relationship in Visual Genome. The full lists are shown in Figure 2.

Object filtering. Only keep objects O for which there exists at least two triplets (S, R, O) and (S', R', O) with $R \neq R'$. This ensures that the selected objects have at least two different types of interaction in images. Then, only keep the 10 most frequent objects for a single (subject, relation) pair. This gives a list of 645 distinct triplets.

Building transformation queries. For each image I with associated triplet (S, R, O) , add in the list of transformation queries $(I, O \rightarrow O')$ if there is a triplet (S, R, O') in the database. Do the same for S and R . This ensures that transformation queries consists of pairs of *objects* that can have the same (subject, relation) pair, and symmetrically for *subjects* and *relations*.

Writing captions for OSCAR. For each of the 645 triplets, we manually wrote a caption in natural language, e.g. (*man, sitting on, chair*) \rightarrow *A man sitting on a chair*.

We now have a database of images and transformation queries, but we have noticed some noise in the annotation procedure: an image can have a triplet annotation which does not well describe the main action in it, because the cropping procedure included an object that is more important than the extracted triplet. Also, transformation queries sometimes consisted in synonyms. We solve this problem using OSCAR to filter transformation queries: Given an image I with query triplet t_1 and target triplet t_2 , we keep the corresponding transformation if $\mathbb{P}_O(I, t_1) > 0.9$ and $\mathbb{P}_O(I, t_2) < 0.1$. This ensures that not modifying the image is not a valid solution to the problem.

The distribution of images being quite skewed (see Figure 2), the transformation queries also have a bias towards the more frequent subjects, relations and objects. We alleviate this problem by using reweighting in the scoring metric (see below).

In summary, our SIMAT dataset (for Semantic Image Transformation) consists of:

- 5 989 images, each annotated with a subject-relation-object triplet.
- 17 996 transformation queries on those images, with queries asking to change the subject, the relation, or the object.
- A list of 645 distinct subject-relation-object triplets with corresponding captions, each triplet having at least 2 corresponding images.

To allow hyperparameter selection, we make a 50-50

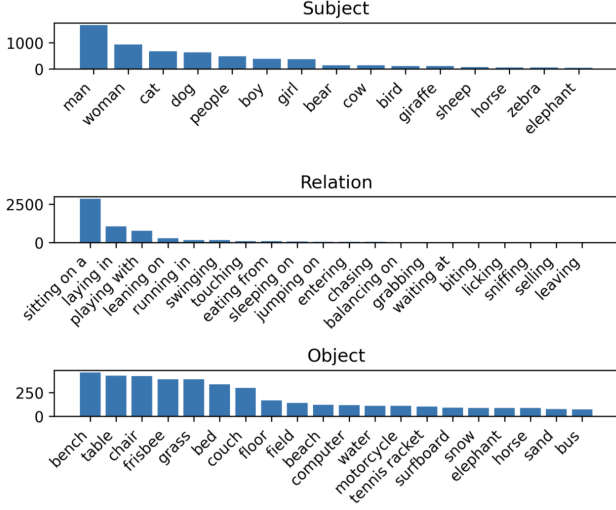


Figure 2. Statistics for SIMAT database. All subjects and relationships are represented, but only 25 objects out of 131 are listed here.

dev/set split on the list of images, and split the transformation queries accordingly.

3.3. Evaluation Metric

Let $(I_i, w_1 \rightarrow w_2, T_i)$ be a sample in our dataset where $w_1 \rightarrow w_2$ is the transformation query and T_i is the caption associated to the target triplet of this sample. For this sample, we consider that a retrieved image J_i corresponds to a successful transformation if OSCAR outputs a probability $\mathbb{P}_O(J_i, T_i) > 0.5$. The final score is simply a weighted accuracy over all dataset samples:

$$S = \sum_{(I_i, T_i, \mu_i) \in S} \mu_i \mathbb{P}_O(J_i, T_i) > 0.5 \quad (1)$$

where the coefficients μ_i are the contributions of each sample to the total score. We adopt an inverse square root reweighting to downsample the most frequent transformations.

4. Methods

Starting from semantic transformations in text, we show how text transformations can be transferred to images via multimodal embeddings. We then present our procedure for fine-tuning multimodal embeddings.

4.1. Text delta vectors for semantic transformations

Semantic properties in sentences can be modified by word replacement: in the sentence “A man walking on the beach”, the semantic property *subject gender* can be changed by replacing the word *man* with the word *woman*. In a latent space, where direct word replacement is not possible, we can apply

semantic transformations by doing arithmetic operations. By encoding sentences as the sum of their word embeddings, applying a transformation $w_1 \rightarrow w_2$ on a sentence embedding $E(s)$ amounts to adding the vector $E(w_2) - E(w_1)$, which we call a *delta vector*. In principle, the textual form of the transformed sentence can be found by retrieving the sentence embedding closest to $E(s) + E(w_2) - E(w_1)$ in a database.

However, there is some ambiguity in the process since bag of words representations do not take into account the order of words. That is why we consider more complex non-linear sentence embeddings which have been shown to display similar properties as above [17], in addition to better reflecting the meaning of sentences [2].

We study four sentence embeddings: *CLIP*, obtained by a contrastive loss on a large set of image/text pairs; *FastText*, obtained with a weighted sum of FastText word embeddings [3]; *LaBSE*, which are trained by matching parallel sentences in different languages with a contrastive loss [6]; and the *LASER* embeddings [2] which are trained with a multilingual translation task.

4.2. From text delta vectors to images

Semantic transformations, seen as *delta vectors* as defined above, can be added to image embeddings in multimodal spaces. We transform images with text delta vectors in the following manner (Fig 1): given an image encoder E_{img} and a text encoder E_{txt} that embed both modalities into a shared latent space, we retrieve in the SIMAT database the image that has the highest cosine similarity with the latent vector

$$x = E_{img}(I) + \lambda \cdot (E_{txt}(w_2) - E_{txt}(w_1)) \quad (2)$$

The *scaling factor* λ is a hyper-parameter that can be adjusted to increase the strength of the transformation. The natural choice is $\lambda = 1$ but it has been noted that a higher value can help to better enforce the transformation [11]. The image embeddings are quite sparse due to the relatively small size of the image database, so we found it helpful to enforce the rule that the retrieved image should be different from the input image.

4.3. Finetuning multimodal embeddings

We consider multiple choices for the image and text encoders: Our default setup is to use the CLIP embeddings for both modalities (63M parameters for the text encoder, 87M for the image encoder), and we experiment with using two ImageNet-pretrained ResNets (ResNet50 and ResNet152, respectively 23M and 63M parameters) as image encoders, and FastText, LASER and LaBSE as text encoders. We can evaluate the vanilla CLIP embeddings without retraining; however, other encoding choices are not directly compatible and we have to fine-tune the encoders to be able to encode image and text into a shared latent space. We use a very

simple fine-tuning scheme on COCO [16] where we train linear adaptation heads after the frozen encoders (Fig 3) for 30 epochs with a learning rate of 1e-3 and a batch size of 4096. Fine-tuning a model takes approx. 3 hours on 8 Tesla V100 GPUs.

When using the ResNet-based encoders, our initial study showed that only training a linear layer is not sufficient to get a reasonable performance on image-text retrieval, because the backbone network is only trained on image classification. Therefore, we freeze only the first three blocks of the ResNet models and add a simple 4-layer MLP architecture on top of the pooled features. We use an image-text InfoNCE [23] contrastive loss (which was used for training CLIP):

$$\begin{aligned}\mathcal{C}(I, T) &= \frac{1}{n} \sum_{i=1}^n \left(\frac{\exp(I_i \cdot T_i / \tau)}{\sum_{j=1}^n \exp(I_i \cdot T_j / \tau)} \right) \\ \mathcal{L} &= \frac{1}{2} \mathcal{C}(I, T) + \frac{1}{2} \mathcal{C}(T, I)\end{aligned}\quad (3)$$

where I and T are normalized image and text embeddings, τ a temperature parameter which is learnable in CLIP. However, we choose to keep it fixed to study its impact on the transformation score.

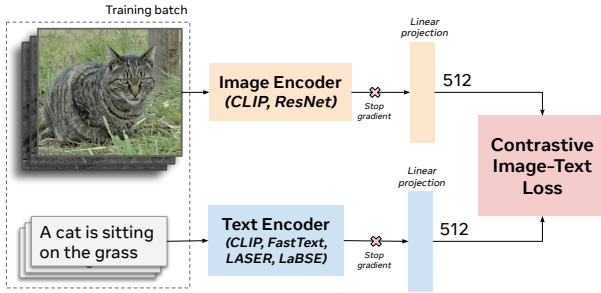


Figure 3. Layer adaptation learning on COCO. The image and text embeddings are projected to a shared multimodal space of dimension 512.

5. Experiments on SIMAT database

In this section, we analyze the ability of various multimodal embeddings to transfer text transformations to images via *delta vectors*.

5.1. Vanilla CLIP embeddings

We first study the performance of the vanilla CLIP embeddings for transferring text transformations to images with delta vectors. To put our results in perspective, we also evaluate the following baselines:

Text to Image: We directly provide the target captions to the CLIP text encoder and retrieve the image closest to that embedding. This is the standard image retrieval task, which is easier because the target subject-relation-object features

Method	SIMAT score	
	$n = 1$	$n = 5$
Delta Vectors ($\lambda = 1$)	15.9	39.2
Delta Vectors ($\lambda = 3$)	35.4	67.6
Image to Text to Image	39.7	71.0
Text to Image	65.9	95.6

Table 1. SIMAT score for delta vectors in the original CLIP multimodal space. The default score considers the nearest neighbour in the retrieval step ($n = 1$). We also report the SIMAT score for the the best image using $n = 5$ nearest neighbours.

are given as input. Hence it can be considered as an upper bound of our SIMAT score.

Image to Text to Image: We first find among the SIMAT captions, the text embedding that is closest to the query image. We then add the text delta vector corresponding to the transformation query and finally retrieve the closest image in the SIMAT database. This means that we do not transform the input image directly but we transform a textual representation of the input image.

Results are shown in Table 1. The delta vector method works for 15.9% of the transformation queries. A higher value of λ gives much better results (35.4%) which are nonetheless below the Image to Text to Image baseline (39%), and very far from the Text to Image upper bound (65.9%). It means that with our benchmark, transforming images works better by using text representations of images rather than the image embeddings themselves. However, in a real-world scenario, we don't want to get explicit context of images by converting them to text (which requires a form of image captioning); we want to use the image embeddings as implicit context.

5.2. Fine-tuning CLIP on COCO

In this section, we consider CLIP as image and text encoder, but we additionally train adaptation layers on COCO with different values for the temperature parameter τ . Figure 5 shows the SIMAT score as a function of the scaling factor λ , on the SIMAT dev set. The same curve for the vanilla CLIP embeddings is shown in black. We can see that all curves have an optimal value for λ , which depends on τ . This optimal value $\lambda^*(\tau)$ decreases as τ increases from 0.01 to 1, and the global optimum is reached for $\tau = 0.1$ and $\lambda = 1$. For these values, the SIMAT score is 48.2 which is a 33% absolute improvement over the zero-shot score.

We therefore conclude that the temperature parameter τ has a great importance for transferring text delta vectors to images, and that the fine-tuned embeddings work best with delta vectors for $\tau = 0.1$ and $\lambda = 1$.





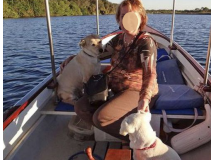





Image Query					
Transformation Query	Woman → Man	Leaning on → Jumping over	Toilet → Suitcase	Kite → Rail	Boat → Bed
Target Caption	A man balancing on a surfboard.	A horse jumping over a fence.	A cat sitting on a suitcase .	A man leaning on a rail .	A woman sitting in a bed .
Retrieved Image					
Success (OSCAR)	YES	YES	YES	NO	NO

Figure 4. Transformation examples from the CLIP model finetuned on COCO with temperature $\tau = 0.1$. Rows 1-3 show examples of successful subject, relation and object transformations. Row 4 shows an example of an unsuccessful object transformation: the retrieved image contains a bench instead of a rail, but we can note some visual similarity with a rail. Row 5 shows a frequent mode of failure: the object is the correct one but the relation has been modified. We assume that our algorithm prioritized keeping the dog in the image.

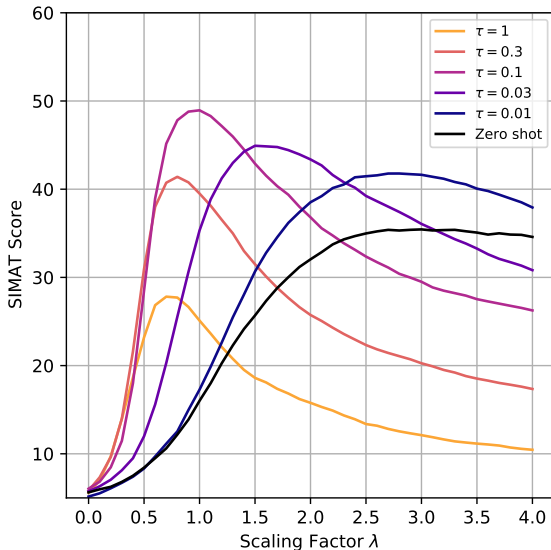


Figure 5. SIMAT score as a function of the scaling factor λ (on development set). The overall best score is obtained for $\tau = 0.1$ and a scaling factor of exactly 1.

Here, we make the important observation that the empirical optimal value for λ is exactly the theoretical value of 1 that should be used to transform bag of word embeddings.

Given that $\lambda = 1$ was suboptimal for vanilla CLIP embeddings, we make the hypothesis that multimodal embeddings that are optimal for $\lambda \neq 1$ can be projected to embeddings better suited for delta vectors (hence having better geometric regularities) that maximize transformation accuracy for $\lambda = 1$.

Transformation examples on SIMAT obtained with this model are presented in Figure 4.

Note that the best image retrieval and text retrieval evaluations on COCO are obtained for $\tau = 0.01$, which hints towards the fact that smaller temperatures are better for image-text retrieval and higher temperatures ($\tau \sim 0.1$) are more compatible with the delta vector framework. In the rest of the paper, we use a fixed temperature of $\tau = 0.1$.

5.3. Using pretrained text encoders

We show in Figure 6, that a value of $\tau = 0.1$ which is optimal for CLIP, is also near-optimal for all other considered text embeddings, FastText, LASER and LaBSE. It seems to be a value that works well for delta vectors. In Table 2, we analyse our different choices for the image and text encoders. The *Retrieval upper bound* metric corresponds to the Text to Image baseline of section 5.1. The *Text delta vector* metric is an evaluation of how well the text-defined delta vector can accurately transform the caption of the input image (and not the image itself). We also compute the standard image/text retrieval metrics (Image R@1 and Text R@1) on the COCO

test set.

We can see that the key contributing factor in the different SIMAT scores is which sentence encoder has been used. If we fix the sentence encoder, the image encoder has an important influence on the image-text retrieval metrics but very little impact on the SIMAT score. Therefore we conclude that improving multimodal embeddings at the task of image/text retrieval will not necessarily improve their geometric properties (in the context of delta vectors).

Also, quite unexpectedly, the SIMAT score does not seem correlated to the Text delta vector score, which measures how well delta vectors can transform text embeddings: the fine-tuned CLIP text embeddings have a text transformation accuracy of 82.4% whereas the fine-tuned FastText embeddings reach 94.4%. Yet they have very similar SIMAT scores (48.2% vs 47.5%). It seems to show that within our constraints, a slightly lower performance on text delta vector (which indicates an embedding space with less geometric structure on the text side) is not the current limitation.

5.4. Sentence-based delta vectors

In our default method for using text-based delta vectors, we used single words as input to the text encoder. This is particularly well suited for the FastText embeddings which are based on word embeddings, but not so much for the LASER and LaBSE sentence encoders which are built to encode sentences and not single words. This could explain the performance gap between FastText and LASER/LaBSE. To test this hypothesis, we changed our definition of delta vectors so that it is computed by encoding sentences rather than single words. We define the *sentence average delta*

Image Encoder	Sentence Encoder	MS Coco		Text delta vectors	SIMAT score	Retrieval upper bound
		Text R@1	Image R@1			
RN50	CLIP	25.4	22.3	88.0	44.5	76.2
RN152		27.6	23.5	87.2	46.0	77.7
CLIP		45.2	34.8	82.4	48.2	75.4
RN50	FastText	17.6	15.2	95.3	44.6	65.6
RN152		19.1	16.3	95.5	46.7	68.0
CLIP		28.2	21.9	94.4	47.5	70.6
RN50	LaBSE	18.8	16.8	91.0	38.8	66.9
RN152		20.4	17.9	90.7	39.9	69.0
CLIP		31.4	24.9	92.9	41.9	69.9
RN50	LASER	17.0	15.4	92.1	37.0	67.0
RN152		19.0	16.9	92.6	36.0	67.6
CLIP		29.6	22.8	92.8	37.7	67.6

Table 2. Comparison of different image and sentence encoders for the evaluation of delta vectors ($\tau = 0.1$).

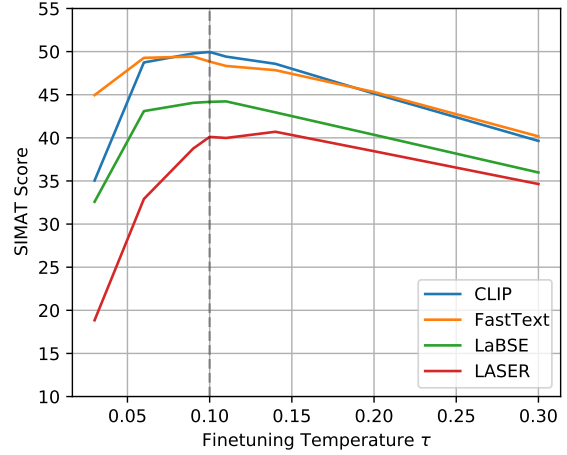


Figure 6. SIMAT score ($\lambda=1$) as a function of training temperature, for several text encoders. For all, the maximum SIMAT score is always obtained for $\tau \approx 0.1$.

Sentence Encoder	Single word	Sentence Average		
		$\lambda = 1$	$\lambda = 1.2$	$\lambda = 1.5$
CLIP	48.2	46.7	51.5	53.5
FastText	47.5	44.6	46.5	45.8
LASER	37.7	43.8	45.0	44.2
LaBSE	41.9	44.6	46.5	45.5

Table 3. Comparison of two methods to calculate delta vectors: *Single word* and *Sentence average*. With the latter, all the encoders have very similar SIMAT scores.

vector of transformation $w_1 \rightarrow w_2$ as the average of delta vectors $E(s_2) - E(s_1)$ where s_1 and s_2 go over all pairs of SIMAT captions such that s_2 is the result of the text transformation $w_1 \rightarrow w_2$ applied to s_1 .

We show the results in Table 3. With this new method, the performance gap between the different text encoders is much smaller, the SIMAT score being higher for LASER and LaBSE, and smaller for FastText. We observed that we can use a higher scaling factor to boost the SIMAT score, up to $\lambda = 1.5$ for CLIP. We suspect this is due to the fact that the second method produces more reliable delta vectors with a smaller norm.

Note that the role of this experiment is to shed light on the reasons behind the performance spread with respect to the text encoders. The captions of SIMAT should be reserved for evaluation only and not used within the algorithm. A better algorithm may use the COCO captions to create better sentence-based delta vectors, but we leave this for future work.

6. Discussion

In this paper, we introduce SIMAT, a novel dataset to study the task of text-driven image transformation. It is much larger in size and variety of transformations than existing approaches like SIC112. Due to this larger size, we argue that evaluation cannot be performed solely by using the caption of the retrieved image, and we propose to use OSCAR to assess whether an input image has been successfully transformed.

We use SIMAT to study the geometric properties of multimodal embedding spaces trained with an image-text alignment objective. We use a simple linear approach (*delta vectors*) for transferring text-defined transformations to images in multimodal spaces, which should work well for well-structured spaces. This provides a novel way to study multimodal embedding spaces compared to standard image/text retrieval metrics in the literature.

After having evaluated vanilla CLIP multimodal embeddings, we have studied embeddings obtained by training with an image/text alignment on COCO, that use pretrained text encoders (FastText, LASER, LaBSE) and pretrained image encoders (CLIP, Resnet50, Resnet152). We emphasize below our findings:

- Vanilla CLIP embeddings, although very powerful for image/text retrieval, are not very well suited for delta-vector based transformation (Tab. 1).
- Finetuning CLIP on COCO brings substantial improvements for delta-vector based transformations and the best performance is obtained for $\tau = 0.1$ and $\lambda = 1$ (Fig. 5). We also observe that $(\tau = 0.1, \lambda = 1)$ is the best functioning point for all considered pretrained text encoders (FastText, LASER, LaBSE, Fig. 6). Since $\lambda = 1$ is also the optimal theoretical value with the delta vector framework, we conclude that finetuning at $\tau = 0.1$ helps to improve the geometric properties of the multimodal embedding space.
- We did not find any evidence that using geometric properties of pretrained sentence embeddings was helpful. While we expected multimodal embedding spaces built on top of these well-behaved text spaces to display better linear properties, experiments have showed the opposite: (a) higher accuracy for text transformation is not correlated to better image transformation (Tab. 2); (b) Using LASER and LabSE was actually harmful (Fig. 6) but we show in Tab. 3 that this is almost entirely due to the fact that we only have access to single words to compute the text delta vector.

For future work, we would like to extend SIMAT to richer semantic transformations: we expect that for transformations like *young*→*old* or *dirty*→*clean*, higher-level semantic knowledge embedded in language models will be critical to

do meaningful transformations.

References

- [1] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1140–1149, 2021. 2
- [2] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. 2, 4
- [3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. 4
- [4] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3984–3993, 2018. 1
- [5] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 1
- [6] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, 2020. 2, 4
- [7] Louis Fournier, Emmanuel Dupoux, and Ewan Dunbar. Analogies minus analogy test: measuring regularities in word embeddings, 2020. 2
- [8] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. *arXiv preprint arXiv:1905.12794*, 2019. 2
- [9] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1463–1471, 2017. 2
- [10] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015. 2
- [11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 1, 2, 4
- [12] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 2

- [13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. [2](#), [3](#)
- [14] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020. [2](#)
- [15] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. [2](#), [3](#)
- [16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. [5](#)
- [17] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018. [2](#), [4](#)
- [18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. [2](#)
- [19] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *arXiv preprint arXiv:1810.11919*, 2018. [2](#)
- [20] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. [2](#)
- [21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. [2](#)
- [22] Jing Shi, Ning Xu, Trung Bui, Franck Dérnoncourt, Zheng Wen, and Chenliang Xu. A benchmark and baseline for language-driven image editing. In *Proceedings of the Asian Conference on Computer Vision*, 2020. [2](#)
- [23] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1857–1865, 2016. [2](#), [5](#)
- [24] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019. [2](#)
- [25] Nam Vo, Lu Jiang, and James Hays. Let’s transfer transformations of shared semantic representations. *arXiv preprint arXiv:1903.00793*, 2019. [1](#), [3](#)
- [26] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6439–6448, 2019. [2](#)
- [27] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016. [1](#)
- [28] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse image generation and manipulation. *arXiv preprint arXiv:2012.03308*, 2020. [2](#)
- [29] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Jianming Zhang, Ning Xu, and Jiebo Luo. Semantic layout manipulation with high-resolution sparse attention. *arXiv preprint arXiv:2012.07288*, 2020. [2](#)
- [30] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020. [1](#)