Multi-modal retrieval with smooth weighting of negatives

Aleksandr Timofeev, Julia Majkowska, Pawel Mlyniec, Sofia Blinova
EE-608 Project Report

***Abstract*—The multi-modal retrieval problem is a challenging task that aims to align semantically two distinct modalities such as images and textual descriptions. The last breakthroughs in this domain are based on contrastive loss functions, especially the maximal triplet loss. However, current studies lacks exploration of other contrastive losses. In our work, we bridge this gap showing that the maximal triplet loss can be replaced with a smooth objective that is more straightforward for optimization. We demonstrate that our method improves Recall@1 on the MS-COCO dataset with our model architecture of choice by 3.8% even with hyperparameters tuned for the baseline loss functions. In addition, we study the effect of other contrastive losses and conduct sample analysis which confirms meaningfulness of our model.**

I. Introduction

The multi-modal retrieval problem is a challenging task that aims to align semantically two distinct modalities such as images and textual descriptions. It allows retrieving the content of one modality given another. For example, one can find the best text description to a given image and vice versa. Though there are different candidate modalities, e.g. video and audio [1], image and video [2], text and audio [3], we focus on image and text modalities in this work. Multi-modal retrieval has many applications in multi-modal databases, e.g. search engines [4], [5], online content recommendations [6], [7], or remote sensing [8]. Additionally, generating multi-modal content embedding can be used as a pivot for other problems such as machine translation [9] or be a good one-shot model to solve tasks in one domain complementing it with information from another modality [10]. Therefore, creating strong models in this field may have tangible impact on research in other fields of study.

This problem is known for decades, however only recently it was tackled using neural networks [11]. The main idea is to build a common semantic embedding space for both modalities in order to discover the correspondences between them. The most popular approaches based on neural networks can be split into off-the-shelf and fine-tuned models [12]. The former methods utilize pre-trained models and improve their representations quality either by feature enhancement or by extracting representations for image patches. The original model parameters are not changed. In contrast, the latter methods fine-tune models towards the retrieval task and address the issue of domain shifts. These approaches may be supervised or unsupervised. In our work, we stick to the last family of methods in the supervised setting. Specifically, we design our models to find a common embedding space for a given set of images and their textual descriptions.

Our method is based on the contrastive learning similar to [13]. To design a meaningful common embedding space, this approach generates embeddings of both modalities and distribute them such that similar ones or positive examples would lay in the neighborhood of each other while distinct ones or negative examples would be far, i.e. it shapes semantic clusters.

One of the main components for this method is a loss function which is paid the least attention in research literature. The state-of-the-art results are shown with the maximal triplet loss, introduced in [13], which considers only *hard* negatives for training. However, other options are not studied well. Moreover, it is not the best choice in the family of contrastive losses as shown in [14] for images. The usage of the maximal triplet loss raises several theoretical and practical problems. Firstly, taking maximum squeezes information available for our model discarding all other samples. Secondly, if the maximum is an outlier, this may mislead the training process. Thirdly, this strategy guarantess a zero error for the train set but generalization is questionable. Finally, taking the maximum by batches induces randomization which makes it all work (despite the listed above problems) but theoretically the batch maximum is a biased estimator of the original loss. In our work, we show that it is possible to avoid this non-smooth objective and improve the results by using the SimCLR loss function [14] which can be seen as smooth weighting of *all* negatives.

The choice of a model architecture has a significant impact on the results of retrieval. There is a great variety of options which can be divided into methods with separate and joint processing of both modalities. In the case of separate training, it is common to choose famous models in their domain and use them to encode input modalities. For example, one can apply a GRU-based model for texts and ResNet for images [13], DenseNet and BERT [15], or Hierarchical LSTM and R-CNN [16]. In joint training, one can leverage visual-language BERT [17], ImageBERT [18], or Unicoder-VL [19]. Since joint training usually involves huge models we prefer the separate training in our work. Concretely, we choose ResNet32 [20] to generate image embeddings and DistillBERT [21] for text embeddings. We consider this choice as an optimial trade-off between fast/light and powerful models.

To sum up our contribution, we propose a method for a multi-modal retrieval based on ResNet32 and DistilBERT as well as the smooth SimCLR objective which eschew problems raised in the maximal triplet loss. We also study and provide results for another contrastive loss such as BarlowTwins [22] which works similarly. Based on the ranking task between two sets, we demonstrate that our method outperforms the maximal triplet loss on the MS-COCO dataset [23]. In the end, we exhibit the behaviour of our models on a sample analysis where we obtain meaningful results.

II. Related Work

There are many advances in the field of multimodal retrieval. One popular approach for this problem is joint training separate embedding models for each modality minimizing the similarity between the matched pairs of different modalities from the data set [16], [24], [25], [19].

In several works, images are preprocessesed into feature embeddings and a joint transformer based model [26] is trained. The model is then used to predict if the image feature embedding sequence is a match for a text label [18], [19], [27].

Apart from testing different architectures, several advances have been achieved in loss functions [13]. While most research is focused on dealing with text in English, some efforts have been made to generalize the multimodal embedding approach for multi-lingual multi-modal retrieval [28], [29], [17]. Other approaches add an additional goal of optimizing the embedding model to be used not only for cross-modal similarity queries but also for image-image and text-text similarity problems [30], [31].

In our work, we plan to use the principles of contrastive learning for the multi-modal embedding task. These principles are predominantly applied to self-supervised learning problems in Computer Vision [32], [33], [34]. There is some prior research on using this training strategy in a supervised setting for the multi-modal embedding problem [35], however we are not aware of any work for using the SimCLR loss [14] or the Barlow Twins loss [22] to train a model for multi-modal retrieval.

III. Preliminaries

In this section, we introduce several loss functions as well as the notation which are domain-specific but the reader should be aware of them for the better comprehension.

A. Sum Triplet

The sum triplet loss is one of the first choices for training of visual-text embeddings. It is used for example in [36], [37], [38] and shares similarities with the hinge loss:

$$\ell_{SH}(i,c) = \sum_{c'} \left[\alpha - s(i,c') + s(i,c)\right]_+ + \sum_{i'} \left[\alpha - s(i',c) + s(i,c)\right]_+ ,$$

where (c,i) is a positive caption-image pair of embeddings, i', c' are negative examples, α is a margin parameter, $s(\cdot,\cdot)$ is a similarity function, $[\cdot]_+ \equiv \max(\cdot, 0)$. This loss consists of two symmetric terms. The first one sums over all negative captions c', given query i. On contrary, the second is taken over all negative images i', given caption c. If i and c are laid closer to each other in the common embedding space than to any negative, by a margin α, the loss is zero. In order to make it loss feasible for training neural networks, it is applied on mini-batches of randomly sampled negatives. The runtime complexity of computing this loss approximation is quadratic in the number of image-caption pairs in a mini-batch.

B. Max Triplet

Originally, the max triplet loss introduced in [13]. It is designed, in contrast to the sum triplet loss, to mine the hardest negatives:

$$\ell_{MH}(i,c) = \max_{c'} \left[\alpha - s(i,c') + s(i,c)\right]_+ + \max_{i'} \left[\alpha - s(i',c) + s(i,c)\right]_+ , \quad (1)$$

where the notation corresponds one used for the sum triplet loss. The difference is only in the max operation instead of summation. This loss function is again applied with mini-batches. Empirically, the max triplet loss demonstrates better performance for multi-modal retrieval than its counterpart with summation.

C. SimCLR

We introduce the SimCLR loss function here in the form as it was proposed in [14]. However, this form is not appropriate for our task and we show necessary modifications in Section IV-A. The SimCLR loss is a contrastive loss and originally designed to train neural networks to find semantically similar images using self-supervised learning in a way that is robust to different augmentations. This is also a good pre-training technique to improve results for downstream tasks. The SimCLR loss can be viewed as a smooth relaxation of the maximal triplet loss (in self-supervised learning computer vision tasks the symmetrization is unnecessary):

$$\ell_{SimCLR}(i,i') = -\log \frac{\exp\left(s(i,i')/\tau\right)}{\sum_{i' \in I} \mathbb{1}[i' \neq i] \exp\left(s\left(i,i'\right)/\tau\right)}, \quad (2)$$

where i and i' are embeddings of a positive pair obtained from one image by different augmentations, $s(a,b) = \frac{a^T b}{\|a\|\|b\|}$, τ is a temperature parameter, I is a set of images from a mini-batch. It is worth noting that $|I| = 2N$, where N is the batch size, i.e. we sum over all negative images of both augmentations. Again, negative samples are sampled randomly. In [14], it is shown that the SimCLR loss outperforms the max triplet loss for computer vision tasks.

D. BarlowTwins

Now, we introduce another contrastive loss function which has been shown to perform even better than SimCLR in computer vision tasks. By analogy to the SimCLR loss function, Barlow Twins [22] creates a pair of images for every original image. The pair is created by applying two randomly sampled augmentations. Similar pairs of images are created for every sample in the mini-batch. The model extracts the normalized (over the batch dimension) representations of the two corresponding distorted versions. The aim is then to make the cross-correlation matrix of them as close as possible to the identity matrix. Specifically, the loss function is

$$\ell_{\mathcal{BT}} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2, \quad (3)$$

where λ is a positive coefficient, $\mathcal{C}$ is a cross-correlation matrix of the embedding from the mini-batch. The advantages of Barlow Twins is that this decorrelation (the second term) removes redundant information about samples in the output units. Unlike other recent self-supervised methods, Barlow Twins does not require large batches.

IV. Method

In this section, we describe our approach to the multi-modal retrieval task.

A. Loss

To train our models for the multi-modal retrieval, we apply new contrastive loss functions, specifically, SimCLR mentioned in Section III-C and BarlowTwins from Section III-D. However, they have to be modificated for the multi-modal retrieval task. The BarlowTwins loss is already symmetric with respect to both dimensions of the similarity matrix. The only change that it needs is to replace one of the image embeddings in Equation (3) with caption embeddings. Yet, as it is seen in Equation (2), the SimCLR loss has to be symmetrized over the second dimension

of the similarity matrix, i.e. over both modalities of negatives. We do it in the following way:

$$\ell_{SimCLR}(i,c) = -\log \frac{\exp\left(s(i,c)/\tau\right)}{\sum_{c' \in C} \mathbb{1}[c' \neq c] \exp\left(s(i,c')/\tau\right)} - \log \frac{\exp\left(s(i,c)/\tau\right)}{\sum_{i' \in I} \mathbb{1}[i' \neq i] \exp\left(s(i',c)/\tau\right)}, \quad (4)$$

where $s(i,c) = \frac{c^T i}{\|c\|\|i\|}$, τ is a temperature parameter, I and C are sets of images and texts from a mini-batch. Now, Equation (4) can truly be called the smooth relaxation of Equation (1).

The fact that the maximal triplet loss is not smooth entails several drawbacks. Firstly, taking maximum squeezes information available for our model discarding all other samples. Secondly, if the maximum is an outlier, this may mislead the training process. Thirdly, this strategy guarantess a zero error for the train set but the generalization is questionable. Finally, taking the maximum by batches induces randomization which makes it all work (despite the listed above problems) but theoretically the batch maximum is a biased estimator of the original loss.

Our smooth relaxation (Equation (4)) automatically resolves many problems of the maximal triplet loss because we switch from the maximal to the average minimization, which is well-studied and provides guarantees by the empirical risk minimization framework [39]. But unlike the sum triplet loss, this loss function allows assigning different weights to negatives accounting for their relative hardness. This may be viewed as soft mining of hard negatives. In addition, [14] demonstrate that the SimCLR loss is empirically better than both triplet losses.

B. Model architecture

Figure 1 shows a general overview of the model architecture. The model consists of two parts for image and text processing. An input batch contains positive and negative examples. Positive pairs are images and descriptions that are aligned semantically in the dataset, all other samples in the batch are negative samples for this pair. Using the image and text models, we obtain embeddings of each sample in the batch, then project them onto the common space. Finally, we compute a similarity matrix of these embeddings which is then used for the loss function.

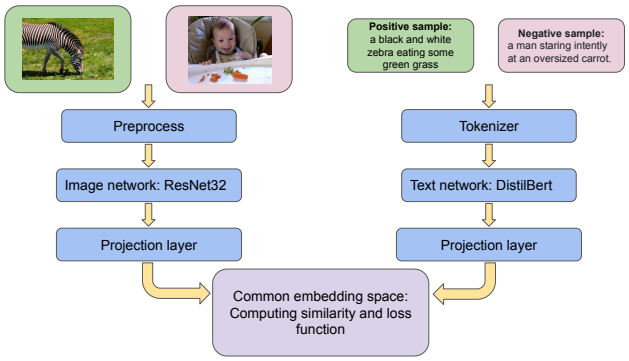


Figure 1: Model architecture.

We choose ResNet32 [40] pretrained on ImageNet [41] to obtain image embeddings. It is one of the best performing convolutional neural network model. The main advantage of the ResNet architecture is "skip connections". When working with deep convolutional neural networks, machine learning experts engage in stacking more layers. While the number of stacked layers can enrich the features of the model, a deeper network can lead to the degradation problem. In other words, as the number of layers of the neural network increases, the accuracy may get saturated and slowly degrade after some depth. As a result, the performance of the model deteriorates both on the training and testing data. This degradation is not a result of overfitting, it is the result of vanishing or exploding gradients. ResNet was created with the aim of tackling this problem. Deep residual networks make use of residual blocks to improve the accuracy of the models. The concept of "skip connections," which lies at the core of the residual blocks, is the strength of this type of neural network.

We take DistilBERT [21] for processing texts. DistilBERT is a small, fast, efficient, and light Transformer model trained by distilling BERT [42]. The BERT model is a bidirectional transformer pretrained using a combination of masked language modeling objective and next sentence prediction on a large corpus comprising the Toronto Book Corpus and Wikipedia. DistilBERT has 40% less parameters than BERT model, runs 60% faster while preserving over 95% of BERT's performances as measured on the GLUE language understanding benchmark.

To align text and image embeddings' size, we add fully-connected projection layers to both models. After putting embeddings into a common space, we apply l_2-normalization and then calculate a similarity matrix.

There are different options for utilizing model weights. Our model unites pretrained image and text models and added on the top projection layers. Thus, it is possible to freeze several parts of the pipeline and train others. In our experiments, we tried to freeze different parts of the model: freeze the image model and train the text model with both projection layers and vice versa, freeze both models and train only projection layers, and train all the weights. It turns out that the best training strategy for our model is to use all its weights for training, i.e. unfreeze both models. We adhere to this strategy throughout all the described experiments.

V. Dataset

We use the MS-COCO 2014 dataset [23]. It has 5 captions for each image of the training and validation sets and no caption provided for tests. To solve this problem and account for our computational resources, we use 5000 samples of the validation set for validation and another 5000 from the same set for testing. Also, we use the remaining samples of the validation set and add them to the train set.

During dataloading, we first process images to ensure that they are the same size. In the training phase, we take a random (224, 224) crop of the resized image and apply a horizontal flip. In the test phase, we resize and crop images deterministically. Second, to process input texts we use the fast implementation of the tokenizer for DistillBert [9]. Since MS-COCO has at least 5 captions to each image, we pick one randomly during training and fix one during testing.

VI. Training Details

We initialize ResNet32 with ImageNet weight as well as use pretrained DistillBERT. The entire models are finetuned during training. We use Adam as an optimizer without weight regularization and set the initial learning rate to 10^{-4}. We schedule

the learning rate with a reduction by 0.5 on 1st, 4th, and 10th epoch, which during experiments performed better than cosine annealing [43] and gradual warmup [44] schedulers. The number of training epochs is 30. The best model is chosen by the best validation loss. The size of embeddings is 128. The batch size is 64. The parameters of the tested loss functions are fixed during training. We use a margin parameter of 0.2 for maximum triplet loss and a temperature parameter of 0.07 for SimCLR, which are values recommended by the authors of the losses.

VII. Evaluation

We evaluate our model on both image to text and text to image retrieval tasks. Our testing dataset consists of 5000 images and 25000 respective captions.

In text to image retrieval, for each caption we calculate the embedding similarity between the caption and each of the 5000 images. We then calculate the ranking based on that. Based on the ranking of the positive image (the one matched to the caption in the dataset) we calculate the mean rank, median rank, and recall@k, where $k = 1, 5, 10$.

The metrics for text to image retrieval are defined as follows:

$$\text{recall@}K_{T2I} = \frac{|\{i, j : \text{rank}(C_{i,j}, I_i) < K\}|}{N},$$
$$\text{MR}_{T2I} = \frac{\sum_{i,j} \text{rank}(C_{i,j}, I_i)}{N},$$
$$\text{MedR}_{T2I} = \text{median}_{i,j}(\text{rank}(I_i, C_{i,j})),$$

where $C_{i,j}$ is the j-th caption assigned to i-th image and I_i is the i-th image.

For image to text retrieval we calculate the embedding similarity ranking between the image and 25000 captions, and take the rank of the highest ranked out of the five positive captions. We calculate the same metrics as for the text to image retrieval task. The metrics for this task are defined as follows:

$$\text{recall@}K_{I2T} = \frac{|\{i : \min_j(\text{rank}(I_i, C_{i,j})) < K\}|}{N},$$
$$\text{MR}_{I2T} = \frac{\sum_i \min_j(\text{rank}(I_i, C_{i,j}))}{N},$$
$$\text{MedR}_{I2T} = \text{median}_i(\min_j(\text{rank}(I_i, C_{i,j}))).$$

Due to hardware limitations, we need to optimize the memory usage of the evaluation. We are unable to load all of the images in the dataset. Thus, in the image-to-text task, for a batch of captions we subsequently iterate over all the images and concatenate the similarity matrices. We calculate the cumulative ranking at the end of the iteration for the batch of captions. Similarly, we do it for the text-to-image task.

VIII. Results

First of all, we demonstrate the performance of our model in dependence of different hyperparameters based on the maximal triplet loss and MS-COCO. Though it is better to explore all combinations of hyperparameters, we start with the most important and hierarchically descend to others. Nevertheless, certain parameters such as an optimizer and parameters of the loss functions are left untouchable and set up to the values recommended in the previous works.

The first experiments are related to the choice of model parameters which should be frozen. Here, we obtain results for the fixed learning rate mentioned in Section VI, the batch size is 128, and the embedding size is 1024 as in [45]. It worth noting that all base models are initialized with pretrained weights, and when a model is frozen, projection layers are trainable. The comparison in Table I claims that the best setting is when both vision and text branches are unfrozen and trained.

Table I: The results of experiments with freezing of model parts. DB - DistilBERT, RN - ResNet32. The best result is in bold.

Setting	I2T Recall@1 (↑)	T2I Recall@1 (↑)
RN & DB frozen	6.2	4.7
RN unfrozen, DB frozen	8.2	5.8
RN frozen, DB unfrozen	8.9	7.2
RN & DB unfrozen	**11.5**	**8.1**

Next, we explore the importance of the embedding size. The both models are unfrozen here, the rest hyperparameters are left the same. We investigate embedding sizes of 64, 128, 512, 1024, 2048. Figure 2 shows that the embedding size is less important hyperparameter. Moreover, the best embedding size of 128 for our models differ from one of 1024 in [45]. This choice makes our model more compact and helpful with overfitting.

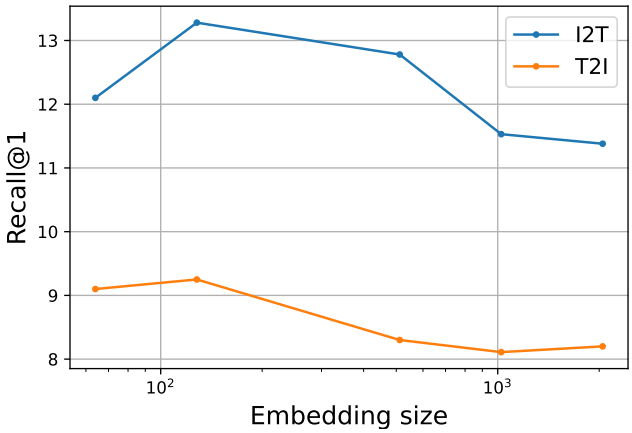


Figure 2: The results in dependence of different embedding sizes for image-to-text and text-to-image retrieval tasks.

Now, we move to the batch size. We fix the embedding size of 128 for this case and investigate batch sizes of 16, 32, 64, 128. The further increasing of the batch size is impossible because of hardware restrictions. We can see from Figure 3 that the changes are miserable but the best choice is 64 and also differ from our original setting.

The left hyperparameters are the learning rate and schedulers. We explore them jointly and extensively. The presenting of all possible options is complicated. Thus, we skip this step saying only that our best results are achieved when the initial learning rate is 10^{-4} and reduced by 0.5 on 1st, 4th, and 10th epoch.

Table II and Table III shows the final evaluation results for image-to-text retrieval and text-to-image retrieval tasks. For both cases the SimCLR obtains the best results among tested losses. Only metric where it is outperformed is the mean rank, in which the sum triplet loss performs better. This is expected since the sum triplet loss sums similarities over images and captions, therefore optimizing this loss implicitly optimizes mean of similarities.

There is also and observable difference in scores between

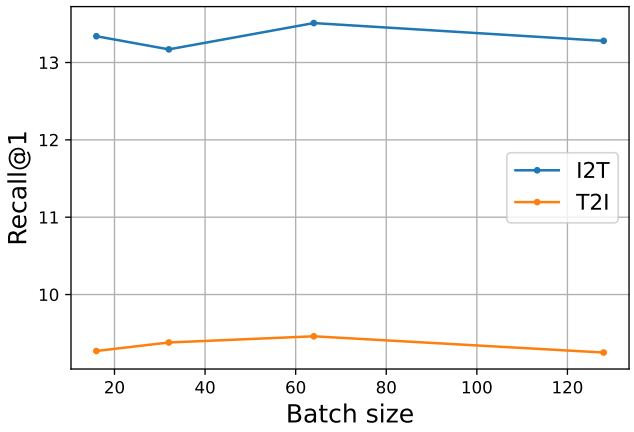


Figure 3: The results in dependence of different batch sizes for image-to-text and text-to-image retrieval tasks.

Table II: Image-to-text retrieval results

Loss Type	Metrics				
	r@1 ↑	*r@5* ↑	*r@10* ↑	*MR* ↓	*MedR* ↓
Sum Triplet	16.5	39.0	52.5	**41.7**	9.0
Max Triplet	15.9	39.4	52.4	42.8	9.0
SimCLR	**19.7**	**43.7**	**56.4**	55.8	**7.0**
Barlow Twins	2.0	7.6	12.0	397.7	111.0

image-to-text and text-to-image retrieval tasks. Image-to-text retrieval scores better results than text-to-image task, because the metrics for this task are calculated based on the best of five positive captions from the dataset.

The Barlow Twins loss showed the worst performance. We suppose it is due to the fact that we optimised the hyperparameters for the max triplet loss and used this parameters for other losses including Barlow Twins. We could have obtained better results if we optimized them with regard to the Barlow Twins loss.

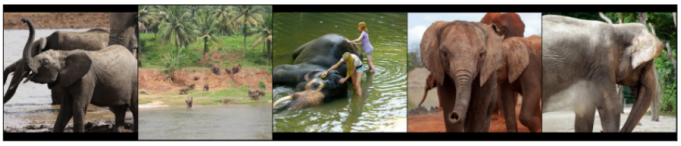
These results are not comparable to the current state-of-the-art, but the presented results compare the performance of the losses and illustrate the potential of using other contrastive losses for multi-modal retrieval. Such results may also be related to the choice of the model architecture. The most recent and best models may exhibit much better results with our loss functions.

Now, we conduct sample analysis of our results. In Figure 4, we test what images will be returned when we input a caption from the dataset. As we see all images contains elephants (the first one is even baby elephant), movement from ”playing” word was captured and in the three best pictures there is also water.

In the next example (Figure 5), we tested what would be the output for a caption which does not have an appropriate image in the dataset. As there is no hamster in the dataset, our model

Table III: Text-to-image retrieval results

Loss Type	Metrics				
	r@1 ↑	*r@5* ↑	*r@10* ↑	*MR* ↓	*MedR* ↓
Sum Triplet	13.2	36.0	50.3	**37.9**	10.0
Max Triplet	12.8	34.5	49.0	39.0	11.0
SimCLR	**15.7**	**39.8**	**54.2**	58.5	**9.0**
Barlow Twins	1.9	7.9	13.6	154.9	62.0



"Baby elephant plays in the water"

Figure 4: Text-to-image retrieval with a caption in the dataset.

"Big hamster plays on the grass"

Figure 5: Text-to-image retrieval with a caption not in the dataset.

finds the closest semantically objects and on this example returns fluffy animals sitting on the grass.

The other way around we test what captions would be returned in image-to-text retrieval. Figure 6 presents predicted and original results for the image of an elephant taken from the MS-COCO dataset. Bolded captions are those which results belong to the originally assigned image captions from the dataset. Even the captions, which are not originally assigned to the searched image, are semantically accurate. Likewise to the previous case, we test the model with an image which does not have the best match in the dataset. Going after text-to-image experiment we choose image of hamster in the grass. As the output can only have captions from the dataset, the result of what the model finds is the closest to the image. Therefore our hamster is labeled as a bear or even a zebra. We think it is because model cannot distinguish size in the image and hamster has similar shape as bear. Mistaking it with zebra can be explained with both animals having stripes on the body. We also notice that typos in descriptions are the same as in the dataset.

IX. Extensions

Receiving promising results with changing the loss function, next step would be to apply and test it on SOTA in multi-modal retrieval. Furthermore we can experiment with optimizable loss parameters for loss functions espacially interesting can be changing the τ in Equation 2. To receive more reliable results it would be needed to test our solution also on different datasets like Flickr 30K. We can also test our approach with a multilingual model as in [28], [29] or add image-to-image and caption-to-caption similarity on augmented data as in [30], [31].

X. Conclusion

The new (in the domain) SimCLR loss showed much better performance then already known loss functions. This is because smooth loss functions are easier for optimization. On the other hand, other works focus on applying in the model RNN and ResNet based architectures. There is a possibility that transformers may not show the best results in a combination with inappropriate loss functions. We also notice that the choice of right hyper-parameters significantly affect the final performance in our setting.

Predicted

- **The elephants are standing beside each other near the water.**
- A very big pretty elephant laying down in the water.
- A young elephant walks with adult elephants along a dirt path
- A small baby elephant walking with other larger elephants
- **A group of elephants walking in muddy water**

Original

- A herd of elephants walking through a lake filled with water.
- A family of elephants washing up at a watering hole
- **A group of elephants walking in muddy water**
- Group of elephants walking in muddy water today.
- **The elephants are standing beside each other near the water**

Figure 6: Image-to-text retrieval with an image from the dataset. The predicted captions matched with original ones in bold.

Closeup of a brown bear sitting in a grassy area
A large bear that is sitting on grass.
A big burly grizzly bear is show with grass in the background
A black bear stands in the wild amongst dead grass
A zeebra running on a grass feild in a park

Figure 7: Image-to-text retrieval with an image not in the dataset.

To sum up, our method is able to design an embedding space where texts and images with similar sense are close to each other achieving the desired solution.

References

[1] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011, pp. 689–696. [Online]. Available: https://icml.cc/2011/papers/399_icmlpaper.pdf

[2] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao, "Deep image-to-video adaptation and fusion networks for action recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 3168–3182, 2019.

[3] F. Chen and Z. Luo, "Audio-text sentiment analysis using deep robust complementary fusion of multi-features and multi-modalities," 04 2019.

[4] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 2333–2338.

[5] M. Haldar, M. Abdool, P. Ramanathan, T. Xu, S. Yang, H. Duan, Q. Zhang, N. Barrow-Williams, B. C. Turnbull, B. M. Collins *et al.*, "Applying deep learning to airbnb search," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1927–1935.

[6] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[7] J. Xu, X. He, and H. Li, "Deep learning for matching in search and recommendation," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 1365–1368.

[8] U. Chaudhuri, B. Banerjee, and A. Bhattacharya, "Siamese graph convolutional network for content based remote sensing image retrieval," *Computer Vision and Image Understanding*, vol. 184, 04 2019.

[9] J. Hitschler and S. Riezler, "Multimodal pivots for image caption translation," *CoRR*, vol. abs/1601.03916, 2016. [Online]. Available: http://arxiv.org/abs/1601.03916

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[11] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 157–166. [Online]. Available: https://doi.org/10.1145/2647868.2654948

[12] W. Chen, Y. Liu, W. Wang, E. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew, "Deep image retrieval: A survey," *arXiv preprint arXiv:2101.11282*, 2021.

[13] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[15] M. Bastan, A. Ramisa, and M. Tek, "T-vse: Transformer-based visual semantic embedding," *arXiv preprint arXiv:2005.08399*, 2020.

[16] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Hierarchical multimodal lstm for dense visual-semantic embedding," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[17] H. Fei, T. Yu, and P. Li, "Cross-lingual cross-modal pretraining for multimodal retrieval," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 3644–3650.

[18] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, "Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data," 2020.

[19] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," 2019.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[22] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," *arXiv preprint arXiv:2103.03230*, 2021.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[24] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013. [Online]. Available: https://proceedings.neurips.cc/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf

[25] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, "Camp: Cross-modal adaptive message passing for text-image retrieval," 2019.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[27] H. Fei, T. Yu, and P. Li, "Cross-lingual cross-modal pretraining for multimodal retrieval," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 3644–3650. [Online]. Available: https://aclanthology.org/2021.naacl-main.285

[28] A. Jain, M. Guo, K. Srinivasan, T. Chen, S. Kudugunta, C. Jia, Y. Yang, and J. Baldridge, "Mural: Multimodal, multitask retrieval across languages," 2021.

[29] A. Mohammadshahi, R. Lebret, and K. Aberer, "Aligning multilingual word embeddings for cross-modal retrieval task," *arXiv preprint arXiv:1910.03291*, 2019.

[30] I. Calixto, Q. Liu, and N. Campbell, "Multilingual multi-modal embeddings for natural language processing," 2017.

[31] S. Gella, R. Sennrich, F. Keller, and M. Lapata, "Image pivoting for learning multilingual multimodal representations," 2017.

[32] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," 2019.

[33] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019.

[34] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[35] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," 2021.

[36] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[37] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 2. Bejing, China: PMLR, 22–24 Jun 2014, pp. 595–603. [Online]. Available: https://proceedings.mlr.press/v32/kiros14.html

[38] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[39] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2015.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[43] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[44] P. Goyal, P. Dollár, R. B. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch SGD: training imagenet in 1 hour," *CoRR*, vol. abs/1706.02677, 2017. [Online]. Available: http://arxiv.org/abs/1706.02677

[45] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," 2018.