0.1.1 Contrastive Learning

Contrastive Learning is a method to learn representations of data without the need for labels, and is therefore a popular method in self-supervised learning. Here, samples (e.g., images) are compared with each other in representation space, typically using some distance metric, with cosine similarity being the usual choice. The goal is to learn an abstract representation of the input modality, such as images.

Originally used in computer vision models like MoCo [1] and SimCLR [2], the idea is that a representation of an image should be similar, or very close, to the augmented versions of the same image. Provided the augmentation is not too drastic (e.g., crop size too large), the high-level semantics of the image will remain the same after augmentation, even though pixel-level information do not. The goal of the image model is then to maximize the cosine similarity between the original image and its augmented versions.

However, this alone is not sufficient, as the model will collapse to a trivial solution by simply returning the same representation for all inputs, which is the simplest way to maximize the cosine similarity between the original image and its augmented versions, because the representation produced for an image would always be the same. To prevent this, negative samples are introduced. Negative samples are other images that do not contain the same content as the original image, and the cosine similarity between the original image and these negative samples should therefore be minimized (a cosine similarity of 0 indicates to similarity between the input vectors). This prevents the model from collapsing to a constant representation, as it would not minimize the cosine similarity and thus not minimize the loss. A simple yet expressive visualization can be found in [3].

This concept can be extended from unimodal to multimodal applications, such as image and text, which is used in this work. In this case, we aim to maximize the cosine similarity between an image and its corresponding text (i.e., caption) and vice versa. Augmentation is not needed, as we always have pairs: one image and one text. Negative samples for images are captions of other images, and vice versa. In this setting, the model learns to produce similar representations for an image and its caption, describing the same real-world concept, and dissimilar representations for an image and caption that are unrelated.

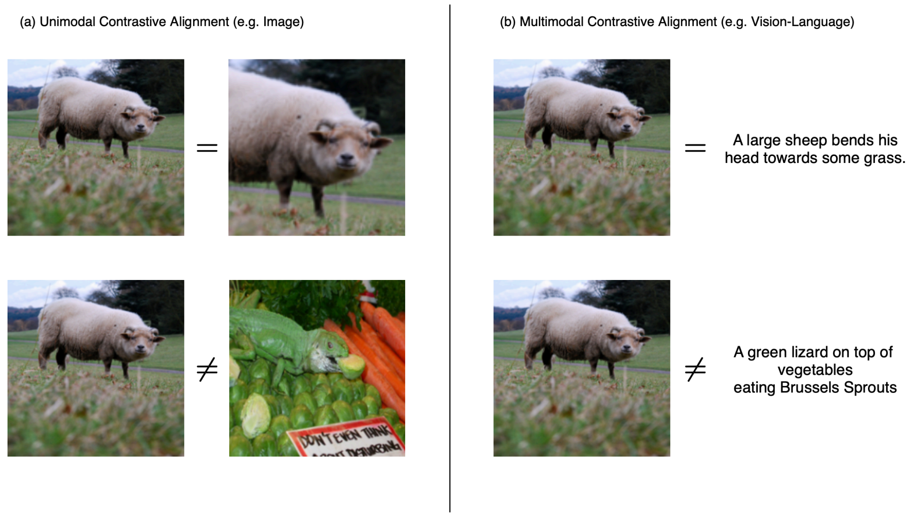
(a) Unimodal Contrastive Alignment (e.g. Image)
(b) Multimodal Contrastive Alignment (e.g. Vision-Language)
=
≠
=
A large sheep bends his head towards some grass.
≠
A green lizard on top of vegetables eating Brussels Sprouts
DON'T EVEN THINK

Figure 1: Contrastive Learning aims to align the same (or similar) real-world concepts in representation space, while pushing different concepts apart. For each input, a global representation is generated, usually through the cls token, and compared with the representation of other samples. Multimodal Contrastive Learning (b) requires existing pairs, e.g. image-text, while for the unimodal case (a) pairs are synthetically created by augmenting the input. Image-Text pairs in the figure have been taken from the COCO train set [4].

Implementation

In the multimodal case, used in this thesis, contrastive learning and contrastive loss are computed at the batch level. The multimodal model creates representations for all images and captions within the batch. In most cases, the CLS token is used as the global representation of an image ([`I_CLS`]) and text ([`T_CLS`]), respectively.

Then, the cosine similarity between the representations of all possible image-text pairs in the batch is computed. This can be done efficiently by first normalizing each embedding and then performing matrix multiplication on the normalized representations. For a batch size of 256, each image has 255 negative samples (i.e., captions of other images) and one positive sample (i.e., its own caption), and vice versa. This can be interpreted as a classification problem with 256 classes, where the model has to predict the correct class (i.e., the positive sample) out of 256 classes/representations, where each class is one caption or image, respectively. The result of the matrix multiplication is a 256x256 matrix of logits, where the diagonal contains the cosine similarity between the positive samples (i.e., the correct class). Applying a row-wise softmax-normalization yields a probability distribution for each image, where each caption in the batch has a probability of being the correct caption for the image, and vice versa. The cross-entropy loss is then used to calculate the loss for the image scores and caption scores, respectively. An illustration of multimodal contrastive learning can be found in Figure 2.

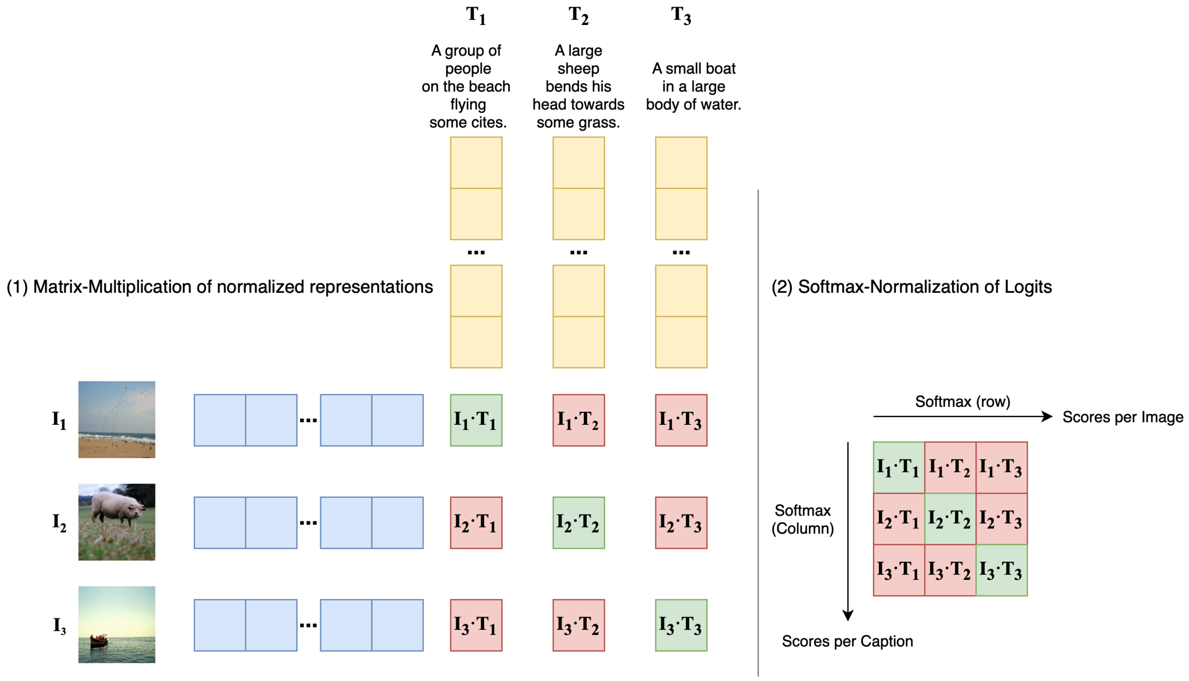
T₁
T₂
T₃
A group of people on the beach flying some cites.
A large sheep bends his head towards some grass.
A small boat in a large body of water.
(1) Matrix-Multiplication of normalized representations
(2) Softmax-Normalization of Logits
I₁
I₂
I₃
I₁·T₁
I₁·T₂
I₁·T₃
I₂·T₁
I₂·T₂
I₂·T₃
I₃·T₁
I₃·T₂
I₃·T₃
Softmax (row)
Scores per Image
Softmax (Column)
Scores per Caption

Figure 2: Contrastive Learning is performed using Matrix-Multiplication of normalized representations (1), usually done using the cls token. The diagonal of the resulting matrix contains the cosine similarity between positive samples. The softmax operation along the rows yields a probabilty distribution for each image over all captions, and the softmax operation along the columns vice versa. The cross-entropy loss is then used to calculate the loss for the image scores and caption scores, respectively. The final loss is the mean of both losses. Image-Text pairs in the figure have been taken from the COCO train set [4].

Problem

The performance of contrastive learning is highly dependent on the number of negative samples available. For instance, with a batch size of two, the model only needs to differentiate between one caption that belongs to the image and one that does not (a negative sample), and vice versa. This task is significantly simpler than with 255 negative samples or more, where there might be captions that are semantically similar to the image, but do not belong to it.

The results improve with an increased number of negative examples [1], [5], as the task becomes more challenging. More negative samples are usually achieved by using larger batch sizes [1], [6], [5]. However, this typically requires higher VRAM GPUs, or multiple GPUs, which is costly.

0.1.2 Image-Text Retrieval

The goal of image-text retrieval (ITR) is to find the matching (most similar) caption for a given image, and vice versa. The process begins with embedding and normalizing a set of samples, such as images or captions, which become a set of keys. For some candidate image or text, called the query, the most similar key is retrieved, after the query is also embedded and normalized. Similar to contrastive learning, cosine similarity is used to compute the similarity between the query and all keys, which is, again, computed by matrix multiplication of the normalized embeddings. The similarities between the query and keys are then ranked, and the key with the highest similarity to the query is the retrieved sample. This method can be viewed as a form of semantic search, which has significant practical relevance in areas like recommendation systems, e.g. to find images based on a given text query. This is precisely what is learned through multimodal contrastive learning.

Image-Text Retrieval is a cheap and efficient way to benchmark the quality of the learned representations of a vision-language model, as it does not require any finetuning, just the embeddings produced by the model. The metric used for benchmarking is Rank@K (R@K), where K determines at which rank the paired/correct sample has to be in the ranking in order to be considered as a correct retrieval. We use R@1, R@5, and R@10, where R@1 is the normal accuracy, i.e., the paired sample has to be the most similar one. R@5 means that the paired sample has to be in the top 5 most similar samples, and for R@10, it has to be in the top 10 most similar samples, in order for the retrieval to be considered correct.

In this thesis, we use the 5K test set of MSCOCO [4], and the 1K test set of Flickr30k [7] for benchmarking, which is the standard benchmarking dataset for multimodal models like FLAVA [8], CLIP [6], VLMo [9], and BEiT-3 [5]. This provides us with an easy and cheap way to compare our model to state-of-the-art (SOTA) models.

Bibliography

[1] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," *arXiv preprint arXiv:1911.05722*, 2019.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," *arXiv preprint arXiv:2002.05709*, 2020.

[3] T. Chen and G. Hinton, "Advancing Self-Supervised and Semi-Supervised Learning with SimCLR." [Online]. Available: https://research.google/blog/advancing-self-supervised-and-semi-supervised-learning-with-simclr/

[4] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., in Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 740–755.

[5] W. Wang *et al.*, "Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19175–19186. doi: 10.1109/CVPR52729.2023.01838.

[6] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, M. Meila and T. Zhang, Eds., in Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763.

[7] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, Feb. 2014.

[8] A. Singh *et al.*, "FLAVA: A foundational language and vision alignment model," *CoRR*, 2021, [Online]. Available: https://arxiv.org/abs/2112.04482

[9] H. Bao *et al.*, "VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts," in *Advances in Neural Information Processing Systems*, 2022. [Online]. Available: https://openreview.net/forum?id=bydKs84JEyw