

0.1.1 Contrastive Learning

In settings where masking discrete tokens and predicting them based on a set of possible tokens, as in language models, is not possible, contrastive learning can be used as a self-supervised method. This is especially useful in vision models, as images are continuous, so there is no discrete set of possible tokens to predict.

Contrastive learning, or the contrastive loss, is a method to learn representations of data without the need for labels, and used in computer vision models like MoCo [1], SimCLR [2], and CLIP [3].

In computer vision, contrastive learning exploits the fact that the high-level semantics of an image are invariant to small (or moderate) changes in pixel-level information. This is achieved by augmenting the input image, e.g., by cropping, rotating, or flipping it. 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 global representations of two augmented versions of the same image. In Transformers, the global representation is usually the [CLS] token returned by the final layer of the model, which is a vector that can be compared with the [CLS] token of another image using the cosine similarity. The augmented versions are often referred to as a different *view* of the same image [4], as shown in Figure 1.

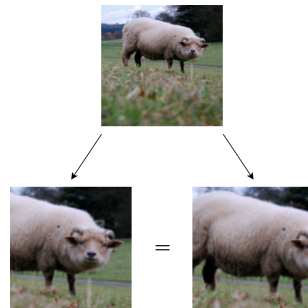


Figure 1: Adding small translations to an image, e.g. a random crop, as illustrated in the figure, will retrain high-level semantic features while changing pixel-level information. The content of the image stays the same, and the same should therefore hold for the representations produced by the model. Image in the figure has been taken from the COCO train set [5].

However, this alone is not sufficient, as the model will collapse to a trivial solution by simply returning the same representation for all inputs, as demonstrated in the papers MoCo [1] and SimSiam [4]. Producing the same representation for all inputs 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, therefore maximizing the cosine similarity (a value of 1). 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 no 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 [6]. This makes self-supervised training of image models possible, and the learned representations represent the high-level semantics of the images, learned without the need for labels.

An implementation and mathematical formulation of the contrastive loss will be introduced in (TODO: cite vision language contrast) on the example of vision-language 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] 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.
- [4] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15745–15753. doi: 10.1109/CVPR46437.2021.01549.
- [5] 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.
- [6] 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/>