Contrastive Self-Supervised Learning

Subjects: Others

Contributor: Ashish Jaiswal

Self-supervised learning has gained popularity because of its ability to avoid the cost of annotating large-scale datasets. It is capable of adopting self-defined pseudolabels as supervision and use the learned representations for several downstream tasks.

Keywords: contrastive learning ; self-supervised learning

1. Introduction

The advancements in deep learning have elevated the field to become a core component in most intelligent systems. The ability to learn rich patterns from the abundance of data available today has made the use of deep neural networks (DNNs) a compelling approach in the majority of computer vision (CV) tasks such as image classification, object detection, image segmentation, and activity recognition, as well as natural language processing (NLP) tasks such as sentence classification, language models, machine translation, etc. However, the supervised approach to learning features from labeled data has almost reached its saturation due to intense labor required in manually annotating millions of data samples. This is because most of the modern computer vision systems (that are supervised) try to learn some form of image representations by finding a pattern between the data points and their respective annotations in large datasets. Works such as GRAD-CAM [1] have proposed techniques that provide visual explanations for decisions made by a model to make them more transparent and explainable.

Traditional supervised learning approaches rely heavily on the amount of annotated training data available. Even though there is a plethora of data available, the lack of annotations has pushed researchers to find alternative approaches that can leverage them. This is where self-supervised methods play a vital role in fueling the progress of deep learning without the need for expensive annotations and learn feature representations where data provide supervision.

Supervised learning not only depends on expensive annotations, but also suffers from issues such as generalization error, spurious correlations, and adversarial attacks[2]. Recently, self-supervised learning methods have integrated both generative and contrastive approaches that have been able to utilize unlabeled data to learn the underlying representations. A popular approach has been to propose various pretext tasks that help in learning features using pseudolabels. Tasks such as image-inpainting, colorizing grayscale images, jigsaw puzzles, super-resolution, video frame prediction, audio-visual correspondence, etc. have proven to be effective for learning good representations.

Generative models gained their popularity after the introduction of Generative Adversarial Networks (GANs) [3] in 2014. The work later became the foundation for many successful architectures such as CycleGAN[4], StyleGAN[5], PixelRNN [6], Text2Image[7], DiscoGAN[8], etc. These methods inspired more researchers to switch to training deep learning models with unlabeled data in an self-supervised setup. Despite their success, researchers started realizing some of the complications in GAN-based approaches. They are harder to train because of two main reasons: (a) non-convergence—the model parameters oscillate a lot and rarely converge, and (b) the discriminator gets too successful that the generator network fails to create real-like fakes due to which the learning cannot be continued. Furthermore, proper synchronization is required between the generator and the discriminator that prevents the discriminator converging and the generator diverging.

Unlike generative models, contrastive learning (CL) is a discriminative approach that aims at grouping similar samples closer and diverse samples far from each other as shown in Figure 1. To achieve this, a similarity metric is used to measure how close two embeddings are. Especially, for computer vision tasks, a contrastive loss is evaluated based on the feature representations of the images extracted from an encoder network. For instance, one sample from the training dataset is taken and a transformed version of the sample is retrieved by applying appropriate data augmentation techniques. During training, referring to Figure 2, the augmented version of the original sample is considered as a positive sample, and the rest of the samples in the batch/dataset (depends on the method being used) are considered negative samples. Next, the model is trained in a way that it learns to differentiate positive samples from the negative ones. The differentiation is achieved with the help of some pretext task (explained in Section 2). In doing so, the model learns quality representations of the samples and is used later for transferring knowledge to downstream tasks. This idea is advocated by an interesting experiment conducted by Epstein[9] in 2016, where he asked his students to draw a dollar bill with and without looking at the bill. The results from the experiment show that the brain does not require complete information of a visual piece to differentiate one object from the other. Instead, only a rough representation of an image is enough to do so.

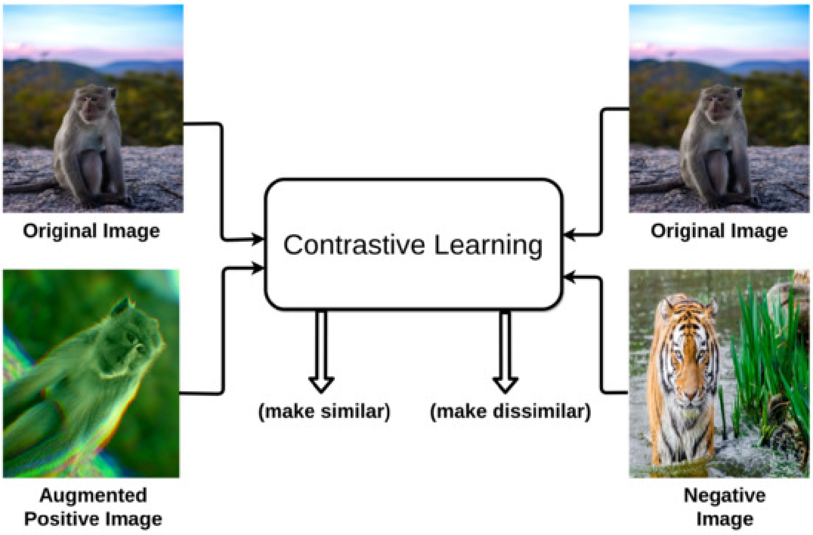

Figure 1. Basic intuition behind contrastive learning paradigm: push original and augmented images closer and push original and negative images away.

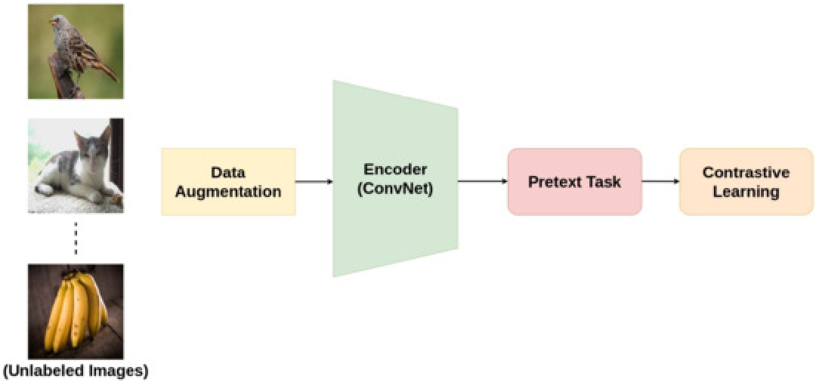

Figure 2. Contrastive learning pipeline for self-supervised training.

Most of the earlier works in this area combined some form of instance-level classification approach [10][11][12] with contrastive learning and were successful to some extent. However, recent methods such as SwAV[13], MoCo [14], and SimCLR[15] with modified approaches have produced results comparable to the state-of-the-art supervised method on ImageNet [16] dataset as shown in Figure 3. Similarly, PIRL [17], Selfie[18], and the work in[19] are some papers that reflect the effectiveness of the pretext tasks being used and how they boost the performance of their models.

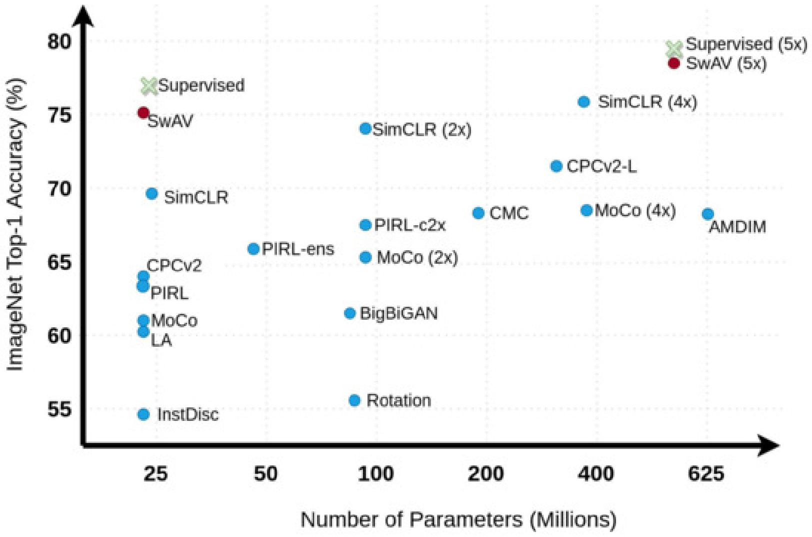

Figure 3. Top-1 classification accuracy of different contrastive learning methods against baseline supervised method on ImageNet.

2. Contrastive Learning in NLP

Contrastive learning was first introduced by Mikolov et al. [20] for natural language processing in 2013. The authors proposed a contrastive learning-based framework by using co-occurring words as semantically similar points and negative sampling [21] for learning word embeddings. The negative sampling algorithm differentiates a word from the noise distribution using logistic regression and helps to simplify the training method. This framework results in huge improvement in the quality of representations of learned words and phrases in a computationally efficient way. Arora et al. [22] proposed a theoretical framework for contrastive learning that learns useful feature representations from unlabeled data and introduced latent classes to formalize the notion of semantic similarity and performs well on classification tasks using the learned representations. Its performance is comparable to the state-of-the-art supervised approach on the Wiki-3029 dataset. Another recent model, CONtrastive Position and Ordering with Negatives Objective(CONPONO)[23], discourses coherence and encodes fine-grained sentence ordering in text and outperforms BERT-Large model despite having the same number of parameters as BERT-Base.

Contrastive Learning has started gaining popularity on several NLP tasks in recent years. It has shown significant improvement on NLP downstream tasks such as cross-lingual pretraining[24], language understanding[25], and textual representations learning[26]. INFOXLM [24], a cross-lingual pretraining model, proposes a cross-lingual pretraining task based on maximizing the mutual information between two input sequences and learns to differentiate machine translation of input sequences using contrastive learning. Unlike TLM [27], this model aims to maximize mutual information between machine translation pairs in cross-lingual platform and improves the cross-lingual transferability in various downstream tasks, such as cross-lingual classification and question answering. Table 1 shows the recent contrastive learning methods on NLP downstream task.

Table 1. Recent contrastive learning methods in NLP along with the datasets they were evaluated on and the respective downstream tasks.

Model	Dataset	Application Areas
Distributed Representations [20]	Google internal	Training with Skip-gram model
Contrastive Unsupervised [22]	Wiki-3029	Unsupervised representation learning
CONPONO[23]	RTE, COPA, ReCoRD	Discourse fine-grained sentence ordering in text
INFOXLM[24]	XNLI and MLQA	Learning cross-lingual representations
CERT [25]	GLUE benchmark	Capturing sentence-level semantics
DeCLUTR [26]	OpenWebText	Learning universal sentence representations

Most of the popular language models, such as BERT [28] and GPT[29], approach pretraining on tokens and therefore may not capture sentence-level semantics. To address this issue, CERT [25] that pretrains models on the sentence level using contrastive learning was proposed. This model works in two steps: (1) creating augmentation of sentences using back-translation, and (2) predicting whether two augmented versions are from the same sentence or not by fine-tuning a pretrained language representation model (e.g., BERT and BART). CERT was also evaluated on 11 different natural language understanding tasks in the GLUE benchmark where it outperformed BERT on seven tasks. DeCLUTR [26] is self-supervised model for learning universal sentence embeddings. This model outperforms InferSent, a popular sentence encoding method. It has been evaluated based on the quality of sentence embedding on the SentEval benchmark.

3. Discussions and Future Directions

Although empirical results show that contrastive learning has decreased the gap in performance with supervised models, there is a need for more theoretical analysis to form a solid justification. For instance, a study by Purushwalkam et al.[30] reveals that approaches like PIRL[17] and MoCo[14] fail to capture viewpoint and category instance invariance that are crucial components for object recognition. Some of these issues are further discussed below.

3.1. Lack of Theoretical Foundation

In an attempt to investigate the generalization ability of the contrastive objective function, the empirical results from Arora et al.[22] [77] show that architecture design and sampling techniques also have a profound effect on the performance. Tsai et al. [31] provide an information-theoretical framework from a multi-view perspective to understand the properties that encourage successful self-supervised learning. They demonstrate that self-supervised learned representations can extract task-relevant information (with a potential loss) and discard task-irrelevant information (with a fixed gap). Ultimately, these findings propel such methods towards being highly dependent on the pretext task chosen during training. This affirms the need for more theoretical analysis on different modules in a contrastive pipeline.

3.2. Selection of Data Augmentation and Pretext Tasks

PIRL[17] emphasizes on methods that produce consistent results irrespective of the pretext task selected, but works like SimCLR[32] and MoCo-v2 [33], and Tian et al. [19] demonstrate that selecting robust pretext tasks along with suitable data augmentations can highly boost the quality of the representations. Recently, SwAV[13] beat other self-supervised methods by using multiple augmentations. It is difficult to directly compare these methods to choose specific tasks and transformations that can yield the best results on any dataset.

3.3. Proper Negative Sampling during Training

During training, an original (positive) sample is compared against its negative counterparts that contribute towards a contrastive loss to train the model. In cases of easy negatives (where the similarity between the original sample and a negative sample is very low), the contribution towards the contrastive loss is minimal. This limits the ability of the model to converge quickly. To get more meaningful negative samples, top self-supervised methods either increase the batch sizes [15] or maintain a very large memory bank[17]. Recently, Kalantidis et al. [34] proposed a few hard negative mixing strategies to facilitate faster and better learning. However, this introduces a large number of hyperparameters that are specific to the training set and are difficult to generalize for other datasets.

3.4. Dataset Biases

In any self-supervised learning task, the data provide supervision. In effect, the representations learned using self-supervised objectives are influenced by the underlying data. Such biases are difficult to minimize with the increasing size of the datasets.

References

1. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
2. Liu, X.; Zhang, F.; Hou, Z.; Wang, Z.; Mian, L.; Zhang, J.; Tang, J. Self-supervised learning: Generative or contrastive. arXiv, 2020; arXiv:2006.08218.
3. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. arXiv, 2014; arXiv:1406.2661.
4. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
5. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.

6. Oord, A.V.d.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel recurrent neural networks. arXiv, 2016; arXiv:1601.06759.

7. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. arXiv, 2016; arXiv:1605.05396.

8. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. arXiv, 2017; arXiv:1703.05192.

9. Epstein, R. The Empty Brain. 2016. Available online: https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer (accessed on 1 November 2020).

10. Bojanowski, P.; Joulin, A. Unsupervised learning by predicting noise. arXiv, 2017; arXiv:1704.05310.

11. Dosovitskiy, A.; Fischer, P.; Springenberg, J.T.; Riedmiller, M.; Brox, T. Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE Trans. Pattern Anal. Mach. Intell. 2015, 38, 1734–1747.

12. Wu, Z.; Xiong, Y.; Yu, S.X.; Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3733–3742.

13. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. arXiv, 2020; arXiv:2006.09882.

14. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.

15. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. arXiv, 2020; arXiv:2002.05709.

16. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.

17. Misra, I.; Maaten, L.V.D. Self-supervised learning of pretext-invariant representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6707–6717.

18. Trinh, T.H.; Luong, M.T.; Le, Q.V. Selfie: Self-supervised pretraining for image embedding. arXiv, 2019; arXiv:1906.02940.

19. Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; Isola, P. What makes for good views for contrastive learning. arXiv, 2020; arXiv:2005.10243.

20. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 2013, 26, 3111–3119.

21. Gutmann, M.U.; Hyvärinen, A. Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. J. Mach. Learn. Res. 2012, 13, 307–361.

22. Arora, S.; Khandeparkar, H.; Khodak, M.; Plevrakis, O.; Saunshi, N. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. arXiv, 2019; arXiv:1902.09229.

23. Iter, D.; Guu, K.; Lansing, L.; Jurafsky, D. Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models. arXiv, 2020; arXiv:2005.10389.

24. Chi, Z.; Dong, L.; Wei, F.; Yang, N.; Singhal, S.; Wang, W.; Song, X.; Mao, X.L.; Huang, H.; Zhou, M. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. arXiv, 2020; arXiv:2007.07834.

25. Fang, H.; Wang, S.; Zhou, M.; Ding, J.; Xie, P. CERT: Contrastive Self-supervised Learning for Language Understanding. arXiv, 2020; arXiv:2005.12766.

26. Giorgi, J.M.; Nitski, O.; Bader, G.D.; Wang, B. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. arXiv, 2020; arXiv:2006.03659.

27. Lample, G.; Conneau, A. Cross-lingual Language Model Pretraining. arXiv, 2019; arXiv:1901.07291.

28. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv, 2018; arXiv:1810.04805.

29. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. 2018. in progress.

30. Purushwalkam, S.; Gupta, A. Demystifying Contrastive Self-Supervised Learning: Invariances, Augmentations and Dataset Biases. arXiv, 2020; arXiv:2007.13916.

31. Tsai, Y.H.H.; Wu, Y.; Salakhutdinov, R.; Morency, L.P. Self-supervised Learning from a Multi-view Perspective. arXiv, 2020; arXiv:2006.05576.
32. Chen, T.; Zhai, X.; Ritter, M.; Lucic, M.; Houlsby, N. Self-supervised gans via auxiliary rotation loss. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12154–12163.
33. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved Baselines with Momentum Contrastive Learning. arXiv, 2020; arXiv:2003.04297.
34. Kalantidis, Y.; Sariyildiz, M.B.; Pion, N.; Weinzaepfel, P.; Larlus, D. Hard Negative Mixing for Contrastive Learning. arXiv, 2020; arXiv:2010.01028.