Paper Review:
What makes for good views for contrastive learning

Jiaxin Hu

May 1, 2021

1 Summary

Contrastive learning aims to train the encoders to find the representations of the inputs, which maximizes the mutual information or minimizes the contrastive loss when the inputs are from the same distribution. Contrastive learning can be considered as a dimension reduction problem (Hadsell et al., 2006), and the goal is to learn a mapping which is invariant to a transformation (cropping the images, generating the images in different color channels) of the data. The learned representations are used in the downstream tasks.

Recently, (Tian et al., 2019) proposes a contrastive multi-view learning which finds the invariant mapping across multiple transformations (views). The views can be generated by data augmentations of the original inputs or can be the data collected by different sensors like vision, sounds, and touch. A natural question is that how many views should we use, and how to select the good views for the downstream tasks.

Why this question is important? Here is a possible explanation. In some sense, multi-view data address the overfitting problem by increasing the sample size. For example, we have $n = 10$ original data point $X_i, i = 1, ..., 10$, and we find 5 views of the data. Then we have $5n = 50$ data points $X_i^{(j)}, j = 1, ..., 5$ and the corresponding sufficient representations $Z_i^{(j)}$. Note that $X_i^{(j)}$ may only have partial information of X_i. Basically, we expect that the representations $Z_i^{(j)}$ encoded from $X_i^{(j)}$ are close, since $X_i^{(j)}$ are from the same X_i. The metric to measure the "closeness" is mutual information. Mutual information between views can be considered as the intersection of the information of the views, where each view contains partial information of the original input. It is possible that the mutual information between two views is 0 or very large with sufficient encoders. If the mutual information is 0, $Z_i^{(j_1)}$ and $Z_i^{(j_2)}$ may be totally irrelevant; if mutual information is very large, we may have $Z_i^{(j_1)} = Z_i^{(j_2)}$ and thus we back to the case with only 10 data points. Both cases are not desirable, and thus the choices of $X_i^{(j)}$ are critical!

Today's paper (Tian et al., 2020) tackles this question. Therefore, the goal of this paper is no longer to find a good representation $Z_i^{(j)}$ with given $X_i^{(j)}$ but to find the proper choice of $X_i^{(j)}$! According to the paper, the ideal choice of $X_i^{(j)}$ is that the $Z_i^{(j)}$ can show the underlying structure of X_i which is needed for the downstream task. That means $I(Z^{(j)}) = I(X, y)$, where $I(Z^{(j)})$ refers to the mutual information of $Z_i^{(j)}$, y is the labels in the downstream task, and $I(X, y)$ is the information in X_i required by downstream task. This principle is called "InfoMin" principle, and corresponding optimization problem is stated in section 4.2.2.

2 Questions

1. This paper presents techniques for minimizing mutual information between views dependent on the downstream task. Is it possible for these techniques in contrastive learning to be applied in a task independent manner? How about in a multi-task setting? Explain and argue why this is or isn't the case.

   I believe it would be hard without the supervision of the downstream task because we do not how close the $Z_i^{(j)}$ should be. In multi-task setting, I believe the InfoMin principle can be generalized by

   $$I(Z^{(j)}) = \sum_{m=1}^{M} I(X, y_m),$$

   where y_m is the labels for the m-th downstream task.

2. This paper highlights that data augmentation resulted in an improvement in downstream classification accuracy. The authors argue this is because data augmentation minimizes mutual information across views rather than the model being shown more data. Is this a reasonable claim? Explain why or why not.

   As far as I am concerned, data augmentation is a way to generate mutli-view of the original data, i.e., generate $X_i^{(j)}$ from X_i. Therefore, if we choose the augmentations $X_i^{(j)}$ based on the InfoMin principle, we can obtain a good performance. However, the statement "data augmentation minimizes mutual information across views" is not always true. Because the mutual information across views depends on the way you implement the data augmentation.

3. What is your personal takeaway from this paper?

   My takeaway includes the following points

   - The choice of views determines the "closeness" (mutual information) of the sufficient representations. The representations which are too close or too far away from each other are not desirable.
   - InfoMin principle implies that the best choice of views has representations whose mutual information is equal to that between the original data and labels from downstream task.
   - Numerical studies imply that the views chosen by InfoMin has better performance in downstream task.

4. Other questions.

   - Notice that the mutual information just describes the amount of information, not the content of the information. Why the InfoMin principle $I(Z^{(j)}) = I(X, y)$ ensures the information in $Z_i^{(j)}$ covers the task -relevant information in X?

References

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Tian, Y., Krishnan, D., and Isola, P. (2019). Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. (2020). What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*.