Summary of "Generative Pretraining from Pixels"

George Stoica

Abstract

"Generative Pretraining from Pixels"[1], written by a team from OpenAI, aims to apply techniques used in natural language processing in image classification tasks. Therefore, the paper's approach is to train a large scale Transformer in a generative task and the model is later used in classification tasks after finetuning. They match other top pretrained models on both CIFAR10 and ImageNet.

1 Overview

The pretraining task is done by using low resolution pictures from ImageNet, cutting the bottom half of the picture and asking the model to complete the picture pixel by pixel. It is similar to a language model, but it is used on pixel. It goes over the image pixels in order, from left to right and learns to generate images similar to the ground truth.

Previously, in image classification, the pretraining on a different dataset was done by also using a classification task. The idea of using a generative task for pretraining comes from natural language processing, in which huge amounts of text data were used to pretrain large scale models. This idea is also useful for image classification due to the fact that, the larger datasets that can be used for pretraining usually lack labels in order to do a proper supervised classification task so it may be better to use a generative task for pretraining on unlabeled data.

2 Differences between Transformers and Convolutions

There are several differences between the Transformers and Deep Neural Networks. A convolution neural network assumes that one pixel cares most about its immediate neighbourhood and the neighbourhood cares concentrically about the region near it until it covers the whole picture.

Transformers do not have this inductive prior of first considering close pixels and then the far away pixels. Transformers have a more general architecture, as everything is connected to everything, while in standard neural networks there are no connections between neurons on the same layer and between layers that are not successive. Being more general allows it to learn things differently and approximate complex functions better because it does not carry that much inductive bias due to its architecture. Using Transformers for image classification task has proved that it creates internal representation similar to convolutional layers from CNNs on its own[2], and this is the reason it may work better. However, Transformers usually need more data and more resources in order to train.

Regarding the generative task, a convolutional neural network is specifically designed to understand that in order to predict one pixel, they need to pay more attention to the close neighbourhood of that pixel. However, a Transformer does not have the information about the spatial relations between the pixels, it needs to learn that by itself.

3 Approach

In order to do the pretraining task the paper's approach is to take an image, to downscale it, to reduce the three color channels to a single one (that indexes the color representation) and finally to unroll the

pixels (to squeeze the pixel matrix to a single dimension). After preprocessing the image, they either do autoregressive generative pretraining or use the BERT model.

The autoregressive generative pretraining is inspired from GPT-2 and using this approach the Transformer receives a list of pixels and predicts the next pixel using attention[3]. The model only knows the previous pixels and this resembles the language models which receive the sentence and predict the next word of the sentence.

The BERT model receives a list of pixels from which some pixels are crossed out and it needs to predict the eliminated pixels. The difference from the autoregressive model is that the autoregressive approach uses only the previous pixels in order to predict the next one while BERT is bidirectional, it also knows some of what follows next.

This approach does not need labels, it can be done as an unsupervised task. After the pretraining on large datasets is done, the research shows that the generative task helps for classification. The linear probe is used to asses how good the internal representation is, and it consists of adding a classification layer somewhere on top of one of the layers of the model and use it for classification. Finetuning consists of putting the classification layer on top of the model and also training the whole model for the classification task.

4 Results

After finishing the generative pretraining task, the model manages to reach a 96.3% accuracy on CIFAR10 on a linear probe and 72% accuracy on ImageNet also on a linear probe. For CIFAR10, the pretraining is done on ImageNet, and for ImageNet the pretraining is done on a wider collection of images from the internet.

This results are obtained after attaching the classification layer on top of a model trained for generative task, and the results are good considering the fact that the model was not trained on a classification task and that it was not even trained on CIFAR10, but on ImageNet. Another interesting thing is that this accuracy is achieved after attaching the classification head on top of one of the intermediate layers. Attaching it to the end of the model slightly decreases accuracy.

This may be due to the fact that the model has not been trained on a classification task so it has a better global representation somewhere in the middle. In the later layers, because they need to generate the next pixel of the image the focus is more on the exact features of that pixel, while in the intermediate layers higher level representation of the global information about the image can be found.

After transfer learning and full finetuning, the model achieves 99% accuracy on CIFAR10 which is comparative to the current state of the art. On ImageNet they do not achieve state of the art accuracy, and this could be due to the downscale they apply for training the large scale model in reasonable time, which does not affect images from CIFAR10 that much.

The paper also claims that while for linear probe, better accuracy is achieved in intermediate layers, for finetuning, it's better to put the classification layer on top of the model and train it. This is expected because the latter layers would be affected by the classification task and the network has a bigger depth.

5 Conclusion

The research shows that useful features can be extracted during pretraining using a generative task and huge amounts of data, and all of this can be transfer learned and used for classification tasks for a smaller dataset. Moreover, a better representation of the global information of an image seems to be located in the intermediate layers of a Transformer. Nevertheless, using techniques previously used in natural language processing, such as attention, can be useful in image processing.

References

[1] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*,

pages 1691–1703. PMLR, 2020.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.