Summary: Learning Transferable Visual Models From Natural Language Supervision Part II

Isaac Akintaro

16/03/2024

Key Definitions:

- **Pretraining:** Pretraining essentially involves training on a large general dataset before finetuning on a specific dataset. This pretraining gives the model a better direction on how to deal with new data, allowing it to generalise better.

1 Current Paradigm in Computer Vision

Most computer vision systems predict on a fixed set of object categories such as types of animals, food etc. Those building the model are aware of these categories before hand. If we want to have more object categories we would need additional labelled data. The issue with this, is that it does not generalise well to unseen images.

2 Proposed Future Paradigm for Computer Vision

A key idea from the advancement of text, the field of Natural Language Processing (NLP) is that training on a web scale collection of text beats high-quality crowd-labelled data sets. In computer vision you normally start with systems pretrained on high-quality crowd-labelled data sets such as ImageNet. Can we do something similar to text were computer vision systems are trained on web scale collection of images?

Well this paper takes a major step in that direction, by predicting the captions which go with an image. The pretrained model developed is trained on 400 million (image, text) pairs collected from the internet.