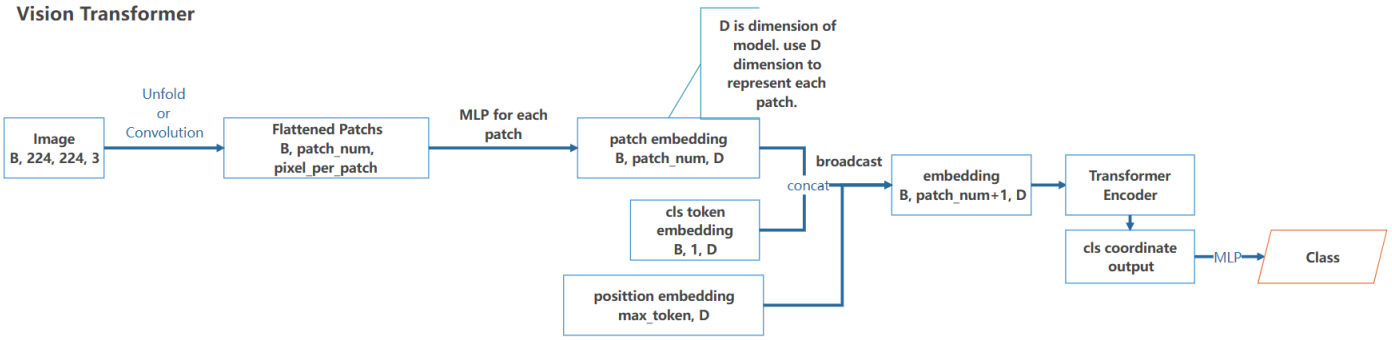
1 月

ViT: Vision Transformer

An image is worth 16x16 words: Transformers for image recognition at scale

ViT：仅仅使用一个标准的 Transformer 来做图片识别，探究一个标准的 Transformer 能否在图像领域中取得一个好的效果。做法是将图片分割成一个一个 Patch，这样降低了模型复杂度，通过一个 MLP 将每个 patch 进行转换，也就是每个 Patch 使用长度为 D 的向量表示，用来抽取语义信息，然后加上位置编码，这是从 Transformer 中借鉴过来的。ViT 最后做的是一个分类问题，也就是说我们最后输出的是一个标号，从分布来说是从图片到标号，而不是 Seq2Seq（Auto 模型），所以我们无需解码器，ViT 将 BERT 的 cls token 引入过来，作为开头，通过 Attention 机制，这个 cls token 能够聚集其他 token 上的信息，同时由于它是可以学习的，因此可以学习到整个数据集的统计特性。因此最后我们只需要取出 cls 对应的输出，经过一个 MLP 映射到我们类别的概率分布上即可。基于图片等价 patch 思想，后续 MAE 的出现真正引领了 Image 领域内大规模数据集的 Pre-train。

Vision Transformer: Image Recognition just using a standard Transformer. The approach is to divide an image into smaller patches, reducing the complexity of the model. A MLP is used to convert each patch into a vector representation of length D, which is used to extract semantic information. The Position encoding, borrowed from Transformer, is then added to these vectors. ViT ultimately performs a classification task, where the output is a label, rather than a sequence-to-sequence task as in an autoregressive model. Therefore, ViT does not require a decoder, and instead introduces the BERT "cls" token, which is used as the starting point for the attention mechanism. The cls token can gather information from other tokens and learn the statistical properties of the entire dataset. The final output is obtained from the cls token and is mapped to the probability distribution of the categories using an MLP. The idea of patch-equivalent images in ViT, along with the subsequent emergence of MAE (Masking Attention Estimation), has truly led the way in pre-training large-scale datasets in the image domain.



学习过的模型：

CNN: AlexNet*, VGG*, NiN*, GoogLet*, ResNet*

RNN: RNN*, GRU, LSTM

Seq2Seq relative: Transformer, Vision Transformer*, BERT, MAE

Generative Model: AE*, VAE*(还差点细节…数学上和代码上还差点), GAN*

Object Detection: SSD, YoLo v5, YoLo v7

Semantic segmentation: FCN*, UNet

*表示复现过.

工作

1. 沐神的《动手学深度的学习》完结。
2. 正在刷一遍 Linux 的视频，配合书，重新回顾一遍。
3. 了解了 VAE 在异常检测方面的应用。
4. 学习了一些新的模型，Transformer，VAE, MAE，其中有的模块 torch 内置了，就没太专注于复现。（偷懒了）