EECS 182 Deep Neural Networks
Spring 2023 Anant Sahai

Discussion 11

1. Finetuning Pretrained NLP Models

In this problem, we will compare finetuning strategies for three popular architectures for NLP.

(a) **BERT** - encoder-only model

(b) **T5** - encoder-decoder model

(c) **GPT** - decoder-only model

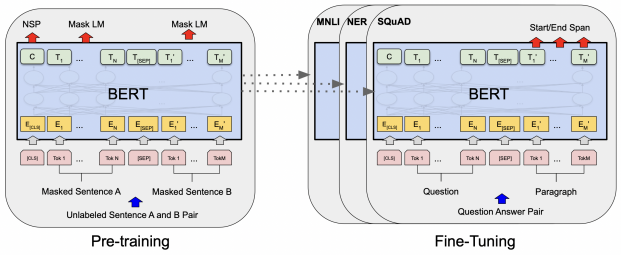

Figure 1: Overall pre-training and fine-tuning procedures for BERT.

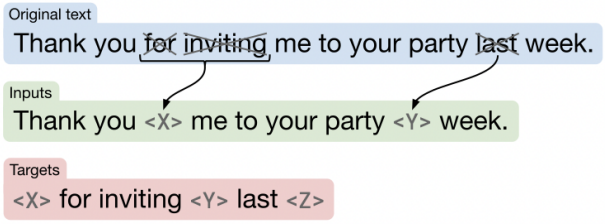

Figure 2: T5 Training procedure

(a) For each of the three models, state the objective used for pretraining.

(b) Consider the MNLI (Multi-Genre Natural Language Inference Corpus) task. It provides a passage and a hypothesis, and you must state whether the hypothesis is an entailment, contradition, or neutral.

> **EXAMPLE:**
> **Passage**: At the other end of Pennsylvania Avenue, people began to line up for a White House tour.
> **Hypothesis**: People formed a line at the end of Pennsylvania Avenue.
> **Classification**: entailment

(i) With each of the 3 models, state whether it is possible to use the model for this task with no finetuning or additional parameters. If so, state how.

(ii) With each of the 3 models, state how you would use the model for this task if you were able to add additional parameters and/or finetune existing parameters.

(c) Compare and contrast the ways we use pretrained representations in BERT to the way we use pretrained autoencoder representations.

2. Vision Transformer

Vision transformers (ViTs) apply transformers to image data by following the following procedure:

(a) **Split image into patches** - The original ViT paper split images into a 16x16 grid of patches.

(b) **Convert each patch into a single vector** - In the original paper, they flattened the patch and applied a linear projection.

(c) **Stack the patches into a sequence, concatenate a CLS token, and add in positional embeddings.** Absolute learned positional embeddings are most common here.

(d) **Pass the sequence through a transformer as usual.**

Below is a diagram of ViT for supervised image classification.

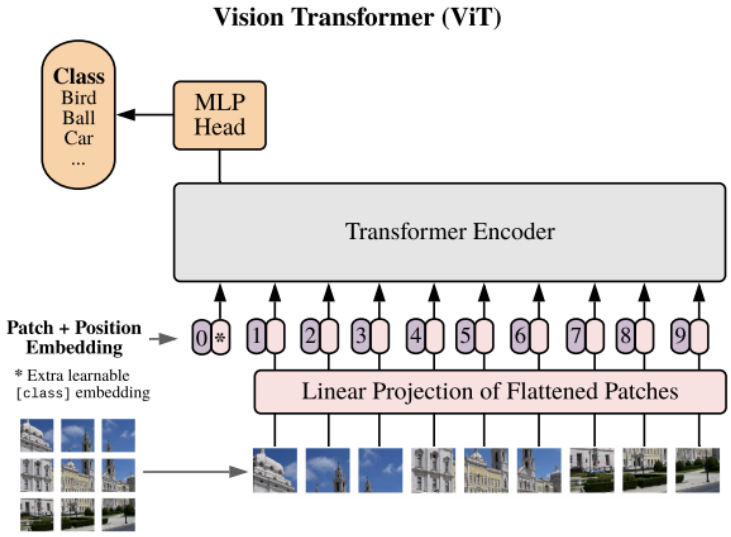

Figure 3: Vision Transformer (Link to Google blog)

(a) Does it matter which order you flatten the sequence of patches?

(b) What is the complexity of the vision transformer attention operation? Assume you have an image of size HxW and patches of size PxP. Only consider the time of the attention operation, not the time to produce queries, keys, and values. Queries, keys, and values are each size D.

(c) What is the receptive field of one sequence item after the first layer of the transformer? How does this compare to a conv net, and what are the pros and cons of this?

(d) If we forgot to include positional encodings, could the model learn anything at all? State one task where a model could perform well without positional encoding, and one task where it would do poorly.

(e) If you wanted to add a few conv layers into this achitecture, how would you incorporate them?

(f) How would you use this architecture to do GPT-style autoregressive generation of images?

(g) So far the vision transformer training procedure has been fully-supervised. Yet we know that most of the available data are unlabelled. How can we do BERT-style self-supervised **representation** learning with vision transformers?

- *Hint 1*: Think about how BERT is trained. How should the input image be modified? What should be the target?)
- *Hint 2*: ViT in this question only has an encoder. For BERT-style training, you will need a decoder. Why is this the case?

Contributors:

- Olivia Watkins.
- Anant Sahai.
- Kevin Li.
- CS 182/282A Staff from previous semesters.

