EECS 182 Deep Neural Networks
Fall 2022 Anant Sahai

# Discussion 10

## 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

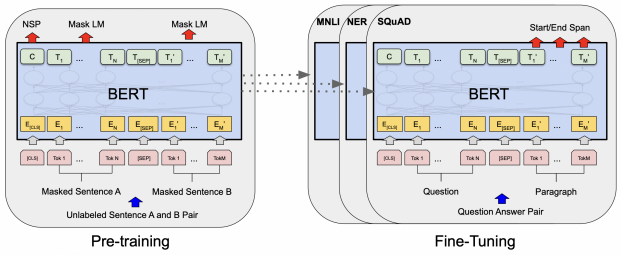


**Figure 1:** Overall pre-training and fine-tuning procedures for BERT.

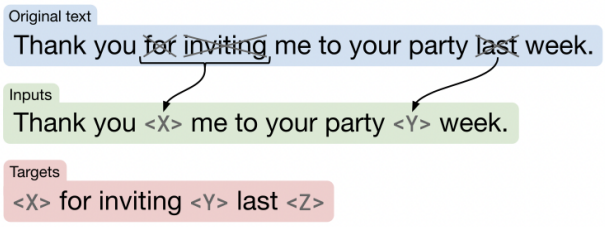


**Figure 2:** T5 Training procedure

(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) Next, consider the SQuAD question-answering task. It provides a passage and asks a question about it. The answer is a span within the task.

**EXAMPLE:**

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Discuss how you could use each of the three models for this task. (You may consider frozen or finetuned methods.)

(d) 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.**

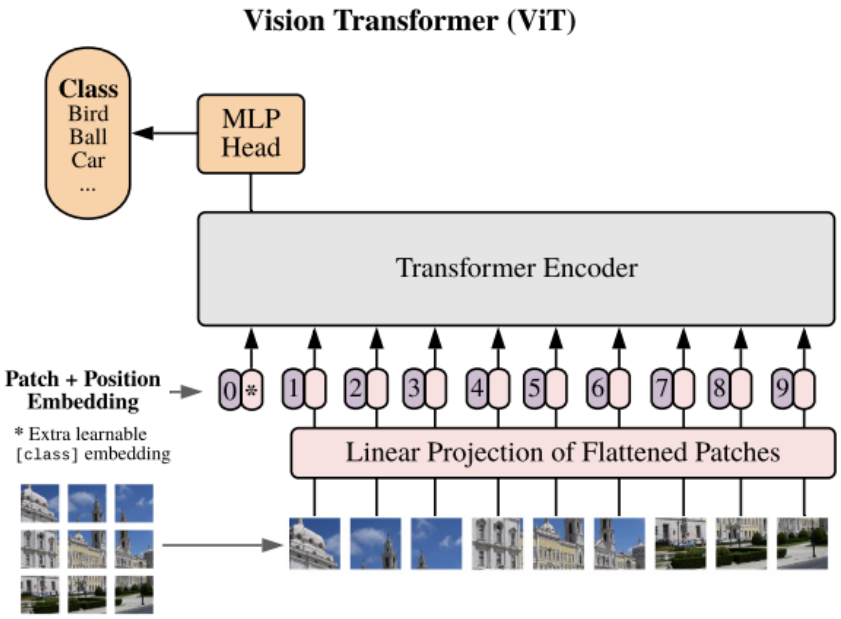


**Figure 3:** Vision Transformer https://ai.googleblog.com/2020/12/transformers-for-image-recognition-at.html

(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?

**Contributors:**

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

  