HW 11

Chen Yuanteng
3039725444

1. Pretraining and Finetuning

(a).

(1) Sentiment Analysis: Bert, its bidirectional nature and ability to capture context make it effective at understanding the sentiment expressed in a piece of text. Fine-tuning Bert for sentiment analysis using task-specific data is likely to yield a good result.

(2) Summarization:

GPT-2 is a natural language generation model that excels in generating coherent and fluent text → fine-tune to learn to generate concise summaries of longer texts.

(3) Named Entity Recognition:

Bert → its ability to capture both left and right context helps in identifying and classifying named entities within a text.

(4) Translation.
Both Bert and GPT-2 can be used for translation.

(b). Early layers of convolutional neural network extract low-level features of image. Cifar-10-C is low level feature shifts. Therefore, fine-tuning only the first block of the pretrained model outperforms the traditional approach of adjusting the task-head.

2. Prompting Languague Models.
(a).
(ii) When temperature = 0, the model is deterministic and outputs the greedy argmax answer every time you generate with the same prompt. When temperature is higher, results are different every time, and the model is more likely to have wried, creative, and nonsensical outputs.

(iii). for most tasks, larger model outperforms the smaller model.

(b).

(i). for simple prompt and simple QA prompt, many of the errors are the model outputting invalid solutions.

for QA Instruction and the two Fewshot variants failures are mostly choosing the incorrect answer.

(ii). the accuracy with incorrect labels in the prompt is similar to or slightly worse than with clean prompts.

(iii)

The model is on average more confident when it's correct.

(iv). GPT-2 model is much smaller and trained on much less data.

(v) On pluralize task, the soft prompt significantly outperforms hard prompts.

(vi). when the model is uncertain which token comes next, the most likely token is often a token which commonly occurs throughout most text corpuses.

3. Soft-Prompting Language Model.

(a). We should include tokens 50-71 (reasoning, answer and newline)

(b)

this soft prompt consists of 5 vectors prepended to the sequence at the input, the token embedding is in dimention E.

So 5E parameters are being trained.

(c)

(i). True, because the masking is autoregressive, the prompt representations will be same for each data point.

(ii). True, if we apply k to both hard, then the embedding of the best possible hard prompt is contained within the set of soft prompts, the best soft-prompt will be at least as good.

(iii) False, full finetuning, especially on a small dataset may hurt generalization and encourage over-fitting.

(iv) False, the parameters of the model remain unchanged expect for new embedding. and when apply to task B, soft prompt will not be attached to inputs.