Improving and Simplifying Pattern Exploiting Training

David Jensen, Anson Jones, Morgan Nañez

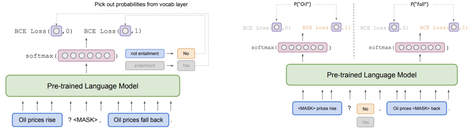
Background

One of the problems which plagues the development of pre-trained language models (LMs), is the lack of easily-accessible labeled data. This has led to the development of **Pattern-Exploiting Training (PET)**, which exploits patterns to give good results for **few-shot learning**. PET performs well (74% accuracy) on the SuperGLUE benchmark using 32 labeled and $\sim$ 9k unlabeled, task-specific examples(per task). This paper proposes a variation to PET called **ADAPET** [1]. ADAPET seeks to combine the main benefit of PET, namely few labeled examples, with one of the benefits of earlier few-shot language models such as GPT-3, namely few task-specific data-points. Both ADAPET and PET reframe language understanding tasks as cloze-style questions, and use BERT's underlying technique of using transformers on a masked language model. ADAPET further distinguishes itself by using a reinforcement learning loss model (decoupling label losses using a binary cross-entropy loss model), and by adding a loss component which conditions on the label rather than on the input.

Methods

MLM Objective

This paper use cloze-style questions, in which the model has to predict the original word at masked locations. During the training phase, masking is done randomly throughout the text, while during testing, the masked out word is always the answer to a task-specific question.



(a) Decoupling Label Loss (b) Label Conditioning

Figure 1. Figure describing training using cloze-style questions [1]. The two additions in ADAPET. Blue boxes are inputs for an entailment task. In task a, the model predicts correct and incorrect labels from the masked word. In task b, the label in the input may be correct or incorrect, and the model must predict the original token from the correct label, and not the original token from the incorrect label.

Patterns and Verbalizers

A pattern describes how to formulate the cloze style question; *p* Question: *q*. It tells you were to add the mask. The verbalizer describes what the out should be. For example, true or false, yes or no. Below is an example of patterns and verbalizer for BoolQ task.

Pattern : p. Question: q? Answer: ___.
Verbalizer: `yes/no`

Pattern :
p. Based on the previous passage, q? ___.
Verbalizer: `yes/no`

Figure 2. Example of what the verablizer and pattern for BoolQ would look like.

Decoupling Label Loss

For decoupled loss, we treat all masked words the same. Essentially, there is no distinction between tokens corresponding to labels and regular tokens. This enables the model to have some probability (and therefore gradient for training) for non-label tokens, which was not the case with PET.

Label Conditioning

Label conditioning is essentially training the model in the opposite direction: asking it to guess the context given the correct label. Both losses from the label conditioning and decoupled label losses are summed together for the final loss.

Reproducing Baseline Model

ADAPET achieves 76.0% average performance accuracy on test data across several few-shot data sets, this slightly outperforms PET (74.0%)and iPET (75.4%) average performance accuracy. For the purposes of the project so far, we have focused on obtaining results using the **BoolQ** dataset. Any interesting findings will be studied across other datasets as well.

Method	BoolQ Acc.	CB Acc./F1	COPA Acc.	RTE Acc.	WiC Acc.	WSC Acc.	MultiRC EM/F1a	ReCoRD Acc./F1	Avg -
ALBERT	55.7	68.6 / 49.1	63.0	50.5	41.4	81.7	3.6 / 49.8	84.1/83.5	57.7
GPT-3 (LAB; SINGLE)	77.5	82.1 / 57.2	92.0 ♠	72.9	55.3 ♠♦	75.0	32.5 / 74.8	89.0 / 90.1 ♠♦	73.2
sPET (LAB; SINGLE)	76.9	87.5 / 85.4	89.0	67.1	49.7	82.7 ♠♦	31.2 / 74.6	85.0 / 91.9	74.2
ADAPET (LAB; SINGLE)	80.3 ♠	89.3 / 86.8 ♠	89.0	76.5 ♠♦	54.4	81.7	39.2 / 80.1 ♠♦	85.4 / 92.1	77.3 ♠♦
PET (LAB + UNLAB; ENSEMBLE)	79.4	85.1 / 59.4	95.0 ♦	69.8	52.4	80.1	37.9 / 77.3	86.0 / 86.5	74.1
iPET (LAB + UNLAB; ENSEMBLE)	80.6 ♦	92.9 / 92.4 ♠	95.0 ♦	74.0	52.2	80.1	33.0 / 74.0	86.0 / 86.5	76.8

Figure 3. Table describing ADAPET's baseline performance across different cloze-style question task compared to other LMs [1]

Adjusting Number of Labeled Examples

The paper compared performance of ADAPET to other few-shot LMs on a benchmark called SuperGLUE, a set of difficult language understanding tasks [1]. GPT-3 used 32 labeled examples and no unlabeled examples, but was outperformed by iPET and PET, which both required $\sim$ 9k additional unlabeled examples. ADAPET, however, outperformed all three using just the 32 labeled examples. Since labeled and unlabeled data can be difficult to collect, GPT-3 and ADAPET are useful in that they use minimal labeled data and no unlabeled data. One thing we are interested in investigating is how far we could push the meaning of 'minimal' labeled data and still outperform the other three models. We are looking at the drop-off curve of accuracy relative to the number of labeled data given for fine tuning, starting with 32 examples (what they used) and then 28, 24, and so on. Initial results imply that there is not a strong accuracy drop off between 32 and 24 datapoints (0.768 for 32, 0.774 for 28, 0.768 for 24). Next, we will corroborate these my averageing over multiple trials and continue the pattern with 20 and 16 datapoints.

Masking Ratio Ablation

One of the secondary studies which this paper performed was aimed at finding the best masking ratio α to use in creating cloze-style questions. Their findings below demonstrated that a value of $\alpha = 10.5\%$ provided the best results across scores in 4 different datasets.

Masking Ratio	BoolQ Acc.	CB Acc./F1	RTE Acc.	MultiRC EM / F1a
15% (FIXED)	80.7	91.1/87.7	70.8	35.8/79.1
10.5% (FIXED)	80.1	89.3/85.0	72.9	35.8/79.1
10% (FIXED)	79.9	81.1/87.5	69.0	33.9/78.4
7.5% (FIXED)	78.3	85.7/79.8	74	36.9/78.8

Figure 4. Results of ADAPET trained on different masking ratios [1]. "FIXED" refers to training with a fixed masking ratio α, while "VARIABLE" (not shown) means training with a variable masking ratio $\leq \alpha$.

We decided that we would like to take this a little bit further. While our first goal is to reproduce the performance on various masking ratios, other studies have shown that higher masking rates, such as 40% or even 80% can stack up to or outperform, lower masking rates [2]. In addition, while performing the testing, we noticed that different batch sizes would have different dev. accuracies, with higher accuracies not necessarily corresponding to larger batch sizes. Focusing only on the "FIXED" case, Figure 5 summarizes these results.

	50	100	150	200	250
10.5%	77.9	77.7	72.5	75.0	72.7
15%	74.4	77.6	69.4	75.2	75.9
40%	76.0	78.2	77.9	78.3	77.1

Figure 5. Dev. Accuracies of independent masking ratio ablation on BoolQ. Horizontal axis is **batch size**, while vertical axis is **masking ratio**.

As we can see, larger batch sizes seem to favor higher masking ratios. This is an interesting result, and somewhat falls in line with past studies that have found a correlation between better performance of higher masking rates and larger model sizes [2].

Masking Method Using POS

We also investigate different masking methods for cloze tasks. The paper current randomly masks 10 percent of all inputs. We further explore purposeful masking methods, by integrating an Part-Of-Speech model into the pre-processing phase to identify words to mask out. For preliminary results, we began by masking out all proper nouns, adjectives, and adverbs. For the POS tagging task, we used NLTK's pre-trained part of speech tagger package. An example of what the new MLM tasks is below.

Original Text: "[CLS] onyx – brazilian green onyx was often used as plinths for art deco sculptures created in the 1920s and 1930s. the german sculptor ferdinand preiss used brazilian green onyx for the base on the majority of his chryselephantine sculptures. green onyx was also used for trays and pin dishes – produced mainly in austria – often with small bronze animals or figures attached.

question : is there such a thing as green onyx? answer : yes.[SEP]"

POS: **Adverb**, Detected words: ['often', 'also', 'mainly', 'often']

New Masked Output: [CLS] onyx – brazilian green onyx was[MASK] used as plinths for art deco sculptures created in the 1920s and 1930s. the german sculptor ferdinand preiss . . .

question : is there such a thing as green onyx? answer : yes. [SEP]

POS: **Adjective**, Detected Words: ['Brazilian', 'German', 'Brazilian', 'Green', 'small', 'such']

New Masked Output: [CLS] onyx –[MASK][MASK] onyx was often used as plinths for art deco sculptures created in the 1920s and 1930s. the[MASK] sculptor ferdinand preiss . . .

question : is there[MASK] a thing as[MASK] onyx? answer : yes.[SEP]

POS: **Proper Nouns**, Detected Words: ['Onyx', 'Ferdinand', 'Preiss', 'Austria', 'Answer']

New Masked Output: [CLS] [MASK] – brazilian green [MASK] was often used as plinths for art deco sculptures created in the 1920s and 1930s. the german sculptor [MASK] [MASK] . . .

question : is there such a thing as green [MASK]? answer : yes.[SEP]

Masked POS	Dev Acc
Original	80.3
NNP	75.5
JJ	75.0
RB	76.1

Figure 6. POS results on BoolQ

As you can see, preliminary POS masking results are not up to par with the original randomized 10.5% masking technique. Some trends we noticed were:

1. Masking **adverbs** didn't affect overall accuracy as batches increased. It stayed relatively close to 75% the whole time.
2. When masking **proper nouns**, as the model trained more, the test accuracy increased, which is expected but not something that was observed in all cases.
3. Masking **adjectives** actually decreased in test accuracy throughout the entire training process.

We would like to test masking 'Filler' words, such as 'to', 'in', and 'a.' We also want to extrapolate current and future masking experiments to other cloze-style tasks such determining contradicting statements and evaluating hypothesis.

Sources

[1] D. Tam, R. R. Menon, M. Bansal, S. Srivastava, and C. Raffel, "Improving and Simplifying Pattern Exploiting Training", Sept. 28, 2021.

[2] A. Wettig, T. Gao, Z. Zhong, and D. Chen, "Should You Mask 15% in Masked Language Modeling?", Feb. 16, 2022.