Testing the Limits: Modifying the Masked Language Model of ADAPET

David Jensen
dmjensen

Anson Jones
ahjones

Morgan Nañez
mn9794

Abstract

The ability of ADAPET to outperform pattern-exploiting training models such as PET and iPET is exciting news for few-shot learning in the absence of task-specific data. In this paper, we further test and expand upon ADAPET in several ways. First, we test performance of ADAPET at various masking ratios confirming some results which they have, and also discovering that high masking ratios, despite having higher loss rates, also perform surprisingly well, particularly at higher numbers of batches. In addition, we investigate different masking methods from those used in ADAPET. By adding a part-of-speech language model to the pre-processing phase, we were able to identify and subsequently mask out different parts of speech across several data sets. We discovered that masking out an entire part-of-speech category did not improve the model across any of the three data sets studied upon. We test masking out proper nouns, adjectives, adverbs, and filler words, all of which decreased the finally accuracy of the model compared to the accuracy obtained by the current masking policy. Finally, we also perform a new ablation, seeing exactly how performance fares as we decrease the number of labeled examples used in ADAPET. We discovered that when reduced the number of labeled examples to 28, we outperform ADAPET's original model.

1 Introduction

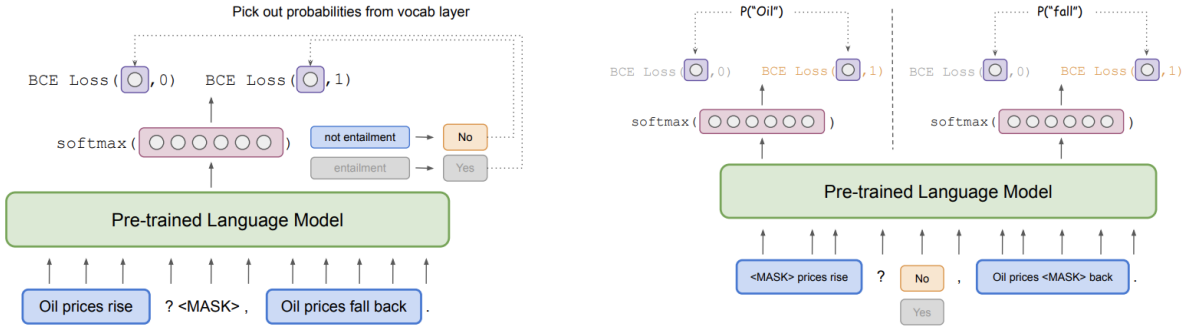
One of the problems which plagues the development of pre-trained language models (LMs), is the lack of easily-accessible labeled data. This has led to the development of Pattern-Exploiting Training (PET; Schick and Schütze, 2021), which exploits patterns to give good results for few-shot learning. PET performs well (74% accuracy) on the SuperGLUE benchmark (Wang et al., 2019) using 32 labeled and $\sim$ 9k unlabeled, task-specific examples per task. This paper proposes a variation to PET called ADAPET (Tam et al., 2021). ADAPET combines 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 applying transformers to 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.

2 Background and Related Works

2.1 Masked Language Model Objective

This paper uses cloze-style questions, a task of masking out random words and attempting to guess the masked-out word, which was first proposed in Taylor (1953) and applied later to the development of the masked language model objective (MLM) in BERT (Devlin et al., 2019). 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. While ADAPET's implementation of masking generally uses a variable (i.e. up to and including) 10.5% masking ratio, we experiment with a higher masking ratio of 40%, motivated by Wettig et al. (2022) results which show the efficacy of higher masking rates, particularly on larger models.

In addition, we also implement a new method of masking words. By implementing a part-of-speech model in the pre-processing phase, we can target specific types of words (Adverbs, Adjectives, Proper Nouns, or Filler Words) to mask out. This method represents one of our central additions to the original codebase.



(a) Decoupling Label Loss

(b) Label Conditioning

Figure 1: Description of training using cloze-style questions from Tam et al. (2021). 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.

3 Data

ADAPET was evaluated and compared to iPET, PET, and GPT-3 on SuperGLUE, a benchmark for training and evaluating a model's performance on natural language understanding tasks. SuperGLUE is based on GLUE (The General Understanding Evaluation **?**), a model-agnostic benchmark, but SuperGLUE is tailored to consist of more difficult tasks. SuperGLUE consists of eight tasks, taken from various text sources online and in print. PET, the basis for ADAPET, reformulates these tasks into close-style questions to match the format used in pre-training, using a pattern and a verbalizer. The pattern is a formula that converts the sentence or sentence pair from the original SuperGLUE task into a question with masked out tokens, while the verbalizer maps tokens (words in the sentence) to the classes originally defined in the task.

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

Pattern :
`p. Based on the previous passage, q? ___.`

Verbalizer: `yes/no`

Figure 2: From Tam et al. (2021), two examples of what a verbalizer and pattern for BoolQ would look like.

BoolQ (Clark et al., 2019) is a task where the model must answer a binary question (q) based on a short passage (p). These questions are either yes/no or true/false and an example pattern is "(p). Based on the previous passage, (q)? __ ". A description of the examples are given in Figure 2.

CB (de Marneffe et al., 2019) is a textual entailment task where the model is asked to determine if a hypothesis (h)is in line with a short passage of premise (p). An example pattern is "(h)? __, (p)". This task gives options for entailment, contradiction, and neutrality (yes, no and maybe respectively).

COPA (Roemmele et al.) is a cause and effect task, where the model is asked to determine the logical cause out of two options (c1 and c2) for a given premise (p). An example pattern is "(c1) or (c2)? (p), so __."

MultiRC (Khashabi et al., 2018)) is a multiple choice answer task. The model is given a question (q) and asked to select from a list of answers (e). An example pattern is "(p). Based on the previous passage, (q)? Is (e) a correct answer? __."

4 Baseline Reproduction and Masking Percent Ratio

This section provides some baseline testing and model reproduction as well as an explication on the study of the masking percent ratio.

4.1 Baselines

	Predicted	Reproduced
BoolQ (Acc.)	79.4%	79.5% (**+0.1%**)
CB (Acc.)	91.1%	91.1%

Table 1: Best performance on BoolQ and CB datasets using 1000 batches at $\alpha = 10.5\%$

For their final results, ADAPET trained its mod-

els using a variable masking ratio $\alpha = 10.5\%$ on 1000 batches. In the interest of time, we reproduced the values for the given α value on only two data sets, BoolQ and CB, using 1000 batches.

The model output values every 50 batches (for 250 batch case) and every 250 batches (for 1000 batch case). For these values, we chose the highest valued output value, although there was sometimes variation, and accuracy values could sometimes decrease even with increased batch size. As we can see, at high batch values, it was sometimes possible to do even better than Tam et al. (2021) with their given parameters.

4.2 Masking Ratio Ablations

What was most interesting however, was studying the performance across different masking ratios. This time, we only ran the models using 250 batches, but as we found that this didn't make a sizeable difference on the results. Table 2 summarizes the results. In addition to reproducing the masking ablations with fewer batches, we additionally took the opportunity to kick up the value of α substantially to 40%. This choice was made in light of work done in Wettig et al. (2022) which showed that it was possible to outperform the standard $\sim 15\%$ masking rate in MLMs with a masking rate of 40% or even higher. We decided to put this to the test and found that, at least for BoolQ, this was true.

We were unsure why it was that for the CB dataset, the same didn't hold. This might have something to do with the qualitative differences between the two datasets. BoolQ is clearly a harder dataset to perform well on than CB, as we can see from the accuracy values, so it is possible that masking out more words in BoolQ wouldn't negatively impact performance due to the heavily supervised nature of ADAPET. We are also not sure about how big the model needs to be in order for a higher masking ratio to improve performance, as Wettig et al. (2022) suggests a larger model might be better suited to a higher α, It is possible that the benefits of a high masking ratio would not apply in this scenario.

In conclusion, however, we reproduced the baseline performance of ADAPET, but it looks best to run the simulation for 1000 batches, as opposed to 250. Further steps might study the effect of a larger masking ratio on different data sets when run on a larger number of batches.

5 Modifications

5.1 Masking by Part-of-speech

We also investigate different masking methods for cloze tasks. The paper current randomly masks 10.5 percent of all inputs. the authors of the paper examine what happens when important words are masked out. The determine importance by using TFDIF and associating relevance with frequency. They found that using TFDIF as an approximation for masking out words hurts performance. We further explore purposeful masking methods, by integrating a 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, adverbs and filler words. For the POS tagging task, we used NLTK's pre-trained part of speech tagger package. For filler words, we masked out words that were tagged as 'IN', which are prepositions, and 'TO'. The intuition is that for certain tasks, some words may not hold much value is determining an answer, while other may be very useful.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]*

Adverb Masking, Detected words: ['often', 'also', 'mainly', 'often']

- **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]*

Adjective Masking, Detected words: ['Brazilian', 'German', 'Green', 'small', 'such']

- **Output:** *[CLS] onyx – [MASK][MASK] onyx was often used as plinths for art deco sculp-*

Masking Ratio	**BoolQ** (1000 batches) Acc.	**BoolQ** (250 batches) Acc.	**CB** (1000) Acc./F1	**CB** (250) Acc./F1
10%	80.0%	77.0%	89.3%/86.8%	89.3%/86.8%
10.5%	79.4%	77.9%	91.1%/88.1%	91.1%/88.1%
15%	78.9%	77.6%	87.5%/80.0%	87.5%/78.7%
40%	–	78.3%.	–	89.3%/85.2%

Table 2: Masking ratio ablation on BoolQ and CB datasets using 250 batches (best output value selected), compared to the values from Tam et al. (2021)

tures created in the 1920s and 1930s. the [MASK] sculptor ferdinand preiss . . .

- **Question:** *is there[MASK] a thing as[MASK] onyx?*
- **Answer:** *yes. [SEP]*

Proper Noun Masking, Detected Words: ['Onyx', 'Ferdinand', 'Preiss', 'Austria', 'Answer']

- **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]*

Filler Word Masking, Detected Words: ['as', 'in', 'on', 'of', 'that', 'for']

- **Output:** *[CLS] onyx – brazilian green onyx was often used [MASK] plinths MASK] art deco sculptures created [MASK] the 1920s and 1930s. the german sculptor ferdinand preiss . . .*
- **Question:** *is there such a thing [MASK] green onyx?*
- **Answer:** *yes. [SEP]*

For this part of our project, we ran each model, with the new masking technique on 250 batches, collecting training loss and development loss every 50 batches. We kept the rest of the parameters the same as the original paper, such as learning rate, warm-up ratio, and weight decay. Due to time and resource constraints, we focused this study on only three datasets, BoolQ, CB, and MultiRC. Results from the paper's own ablation studies are performed on the first pattern for each task. For the following three datasets, we followed suit and only ran the study with their respective first patterns.

5.1.1 Results

	BOOLQ	**CB**	**MultiRC**
Original	80.3	89.3	80.1
NNP	75.5	77.0	78.9
JJ	75.0	73.7	80.0
RB	76.1	78.2	78.2
FILLER	78	76.8	78.5

Table 3: The abbreviations above are as follows: NNP is Proper Noun, JJ is Adjective, RB is Verb, and Filler are prepositions and 'to'. All metrics are reported after being training for 250 batches.

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** did not take away too much context from the input text. Masking adverbs performed generally about the same across data sets
2. When masking **proper nouns**, we noticed too much important information was lost. The model gathered zero context about any specific noun, making answering questions about said proper nouns difficult to do.
3. Masking **adjectives** seemed to give the worse results for BoolQ and MultiRC, and slightly better results for CB. By removing all the adjectives, the model is not able to create deep contextual relationships about the objects in the input.
4. Masking **filler** words tended to have the positive results across the board. Masking the filler words may reduce some noise, but it also teaches the model to create relationships to those missing words, which not entirely beneficial to our task.

BOOLQ Results Masking adverbs didn't affect overall accuracy as batches increased. It stayed relatively close to 75% the whole time. Loss continued to decrease, which leads us to believe that while the model was learning better on the training data, it was not creating useful dependencies for the test data set. 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. While the accuracy increased as batch number increased, we noticed that past 50 batches, the increase is marginal. At 50 epochs, a development accuracy of 74.7 was reached.

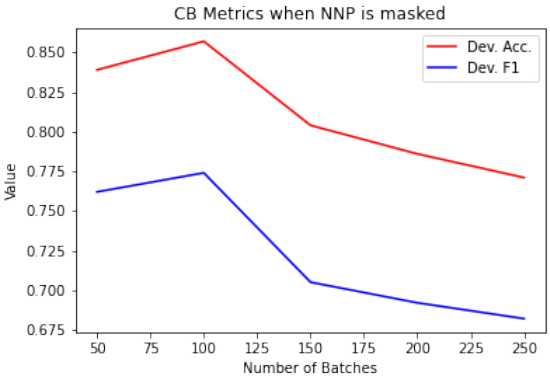
Due to the nature of the task, it is hard for the model to learn about specific nouns and their associations. For example, any question about [MASK] would be hard to discern. Masking adjectives actually decreased in test accuracy throughout the entire training process. After 50 batches, the model had a development accuracy of 75, which decreased to 70 by 250 batches. We are not sure where this decline comes from, but we think that since the training accuracy was 100, that perhaps the model because to over fit and preform worse on the testing data set. When masking filler words, there was a steady increase in development accuracy, up to 78 percent. This masking method proved to be the best for this task.

MultiRC Results When masking proper nouns, the model was able to achieve a final development accuracy of 78.9. While this is high, at around 100 batches, the model had a development accuracy of 79.1. This slight decrease might be due to noise or overfitting. When adjectives were masked, the model achieved a development accuracy of 80. It slowly increased at the number of batches increased. When adverbs were masked the model had an accuracy of 78. When fillers were masked the model maintained an accuracy near 77 throughout all the batches.

Because the nature of the MultiRC data set, masking adjectives may have preformed the best, nearly identical to the original paper, because the questions tended to be geared towards more objective, noun based questions. This would also explain the lower performance when proper nouns were masked.

CB Results The result on the CB data set were perhaps the most perplexing. In all cases, masking out any part-of-speech lead to an interesting phenomena where as the number of batches increased, there was a decline in test accuracy and F1 scores.

Figure 3: CB Metric Analysis



As you can see from above, when proper nouns we masked, the model had a development accuracy of 83.9, which decreased to 77.0. This is one of the greatest differences between starting and ending development accuracy's we see throughout the study. Perhaps because the task is not binary this hurts the model. This trend is seen in other masking variations as well. When adjectives are masked we see a difference of 10 points between starting accuracy and ending accuracy. When masking adverbs or fillers we noticed only a 8 point difference. We are very shocked by this result and admittedly do not have a strong intuition as to why these results are so staggering. Our best guess is that this task is highly sensitive to masking and perhaps ADAPET does not perform well on it as a whole. The original paper also reports low development accuracy for its use involving MultiRC.

5.1.2 Shortcomings

While the idea of masking certain parts-of-speech has appeal, we noticed that overall, the models tend to perform worse across all different masking variations. When masking proper nouns, instead of masking every single one, we believe that masking half or two-thirds of them would be more beneficial to the model accuracy. This is because when the model is presented with any proper noun, it will now be able to add marginal context and have a better understanding of the nouns relation to its context. In theory, this would improve the model when asked questions concerning proper nouns.

We are hesitant to say that masking filler words adds any real value to the model. Instead of masking these filler words, we propose removing them entirely as they have the potential to carry extra

noise.

5.2 Number of Labeled Examples

One of the original paper's motivations was to perform similarly to models like GPT-3, PET, and iPET, without the constraints of data that PET and iPET require and without the model complexity (number of parameters) of GPT-3. In their performance comparison tests, GPT-3 used 32 labeled examples and no unlabeled examples, while iPET and PET outperformed but used 9 thousand additional unlabeled examples. ADAPET, however, outperformed all three using just the 32 labeled examples. The paper reported around 76.0% accuracy compared to GPT-3's roughly 71.9% accuracy. We were interested to see by what margin we could reduce the number of datapoints and still outperform GPT-3, and see what the 'drop off' curve looks like as data gets removed.

The original paper did not specify how many batches they ran to get the data on this particular graph but in other places they used 250 batches for all ablation experiments and occasionally 1000 batches when comparing models. Due to time constraints, we chose to run 250 batches so the direct comparison of this data to GPT-3 is unclear.

We tested performance using BoolQ with pattern 1, and our accuracy with 32 labeled examples (76.8) was lower than what they reported (79.4). We then tested the model with 28 examples, 24 examples, 20 examples, and 16 examples. Since there are so few labeled examples in total, we averaged three trials together to get a more accurate number. For each trial, we removed examples at random from the original 32. It was quite surprising that the dev accuracy increased from 32 to 28 examples, although past that the results were much more as expected. It seemed that there was a relatively sharp dropoff between 24 and 20 examples, with the dev accuracy dropping nearly 4%.

	Dev Accuracy
32 (Original)	76.8
28	77.6
24	77.5
20	73.6

6 Conclusion

Our project expands and explore upon the original ADAPET model. We replicate the masking ratio ablation study and find results confirming the papers conclusion. In addition, we also note discover that higher masking ratios perform better than expected when the number of batches is large. We also explore modifying the masking policy to depend upon part of speech. We examine what happens, across three different data sets, when adjectives, proper nouns, filler words, and adverbs are completely masked in the input. We find that across all variations, that the original masking policy works best.For the MultiRC task, we achieve very similar results to what the original paper reported. We also explored using a different number of labeled examples in the training phase and found that 28 examples outperformed 32 labeled examples for the BoolQ data set. Overall we found the original model to be sufficient, but are curious as to other ways the model can be improved, using the conclusions from our own ablation studies.

Acknowledgments

7 Appendix

References

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. page 6.

Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and Simplifying Pattern Exploiting Training. *arXiv:2103.11955 [cs]*.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2022. Should You Mask 15% in Masked Language Modeling? *arXiv:2202.08005 [cs]*.