Machine Learning HW3

Hongzun Liu
2022210866
Department of Computer Science and Technology
Tsinghua University
liuhz22@mails.tsinghua.edu.cn

1 Classification Task

1.1 Fine-tuning GLM-RoBERTa-Large by prompting

1.1.1 Dataset and Prompt Template

I choose `QASC`(Question Answering via Sentence Composition) for text classification. `QASC` is the first multi-hop reasoning dataset to offer two desirable properties: (a) the facts to be composed are annotated in a large corpus, and (b) the decomposition into these facts is not evident from the question itself.

I use prompt template `is_correct_1` provided by `promptsource`. Samples from this prompt template and dataset are shown in Appendix A.

1.1.2 Experiment Setting

After experiments on **GLM-RoBERTa-Large** with various hyperparameters, I use the following hyperparameters as Table 1. All our models are trained and tested on 1 NVIDIA A100 (80GB) GPU.

Table 1: Hyperparameters for GLM-RoBERTa-Large's Classification Task

Hyperparams	**Values**
Fine-tune Epoch	3
Batch Size	64
Learning Rate	1e-5
Optimizer	Adam($\epsilon = 10^{-6}, \beta = (0.9, 0.98), weight\_decay = 0.1$)
Gradient Clipping	1.0
Scheduler	Linear w/ warmup
Warmup Steps	256

1.1.3 Result

Testset results of proposed methods are presented in Table 2.

Table 2: Testset Results on GLM-RoBERTa-Large's Classification Task

Metrics	NLL↓	Accuracy↑
Results	0.0171	0.9934

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

1.2 Grid Search for Hyperparameters and Analysis

1.2.1 Effect of Learning Rate

In this seciton, we experimented performances of different learning rates. Results on test dataset are shown in Table 3. We can see that the accuracy on test dataset reaches its highest

Table 3: Testset Results on GLM-RoBERTa-Large's Classification Task using different Learning Rate

Max Learning Rate	NLL↓	Accuracy↑
1e-4	0.0977	**1.0000**
1e-5	**0.0170**	0.9934
1e-6	0.0173	0.9956

score when maximum learning rate is `1e-4` during fine-tuning, where all test's are correct. However, the negative log-likelihood reaches its best score when maximum learning rate is `1e-5`. Considering common fine-tuning methods performs poorly when learning rate is lager than `1e-4`, we choose `1e-5` as our maximum learning rate to avoid overfitting problem.

1.2.2 Effect of Batch Size

In this seciton, we experimented performances of different training batch size. Results on test dataset are as Table 4. We can see that when training batch size gets larger (128),

Table 4: Testset Results on GLM-RoBERTa-Large's Classification Task using different Batch Size

Batch Size	NLL↓	Accuracy↑
32	0.0135	0.9945
64	0.0171	0.9934
128	**0.0094**	**0.9967**

performance goes better. This is probably because the batch size during pre-training is 1024 or 2048 or 8192. The consistency between pre-training and fine-tuning may matter.

1.2.3 Effect of Weight Decay

In this seciton, we experimented performances of different `weight_decay` during optimization. Results on test dataset are as Table 5. We can see that performance doesn't change

Table 5: Testset Results on GLM-RoBERTa-Large's Classification Task using different Weight Decay

Weight Decay	NLL↓	Accuracy↑
0.1	0.0171	0.9934
0.01	0.0171	0.9934
0.001	0.0171	0.9934

when weight decay varies. This is a surprising result. This may be beacuse the fine-tuning samples are large enough to avoid overfitting problem. I choose the same weight decay(0.1) as the pre-training procedure for consistency.

1.2.4 Effect of Prompting

In this seciton, we experimented performances of **prompt-based** method and **head-based** method as well as **no-finetune** results. Results on test dataset are as Table 6. We can see that Prompt-based method achieves best performance on both NLL and Accuracy, outperforming traditional head-based method by +1% on accuracy. And prompting method

Table 6: Testset Results on GLM-RoBERTa-Large's Classification Tasks using different methods

Method	NLL↓	Accuracy↑
Prompt-based	**0.0170**	**0.9934**
No-finetune-prompt	2.3241	0.6782
Head-based	0.0461	0.9822

without fine-tuning also achieves 67% accuracy, showing GLM-RoBERTa-Large's strong zero-shot ability.

1.3 Different Prompts and Baselines

1.3.1 More Prompts

I use two more prompt templates, `is_correct_2` on `QASC` and `what_is_the_missing_first_step` on `GLUE/sst2`, besides `what_might_be_the_first_step_of_the_process` on `QASC` mentioned before.

Samples from these two more templates and dataset are shown in Appendix A.

1.3.2 Experiment Result on Different Prompts

Table 7: Testset Results on Different Prompts using GLM-RoBERTa-Large

Prompt	NLL↓	Accuracy↑
`is_correct_1`	0.0171	0.9934
`is_correct_2`	0.0979	1.0000
`happy or mad`	2.4981	0.5113

From Table 7, we can see that GLM-RoBERTa-Large's performance varies significantly with different prompt template. It achieves 1.000 accuracy on template `is_correct_2` but only reach 0.5113 on template `happy or mad`, which is close to a poor random result.

1.3.3 Comparasion with other GLMs

I choose `Flan-T5` and `BART` as two GLM-like baselines.

`Flan-T5` is just better at everything than `T5`(Transfer Text-to-Text Transformer). For the same number of parameters, these models have been fine-tuned on more than 1000 additional tasks covering also more languages.

`BART` is a transformer encoder-decoder (seq2seq) model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder. BART is pre-trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text.

We use facebook/bart-large and google/flan-t5-large as our baseline for text **classification** tasks. Both models are fine-tuned on the same datasets with hyperparameters proposed by their original paper.

1.3.4 Experiment Result on Different GLMs

From Table 8, we can see that BART achieves best accuracy scores compared with other popular GLMs with similar scale. Considering both Flan-T5 and BART are Transformer Encoder-Decoder models, with a encoder to perform understanding and a decoder to perform generation. However, GLM uses only a Decoder model. So it's not surprising that BART can do better on language understanding tasks such as classification, as they have much larger parameters and more precisely, a separate Encoder. But GLM also achieves similar score with fewer parameters and a more simple architecture.

Table 8: Testset Results on Different Models on dataset `QASC` and prompt `is_correct_1`

Model	Params	NLL↓	Accuracy↑
Flan-T5-Large	780M	–	0.9696
BART-Large	–	0.2037	**0.9946**
GLM-RoBERTa-Large	335M	**0.0171**	0.9934

2 Generation Task

2.1 Fine-tuning GLM-RoBERTa-Large by prompting

2.1.1 Dataset and Prompt Template

I choose `WIQA`, the first large-scale dataset of "What if...?" questions over procedural text, for text generation. `WIQA` contains a collection of paragraphs, each annotated with multiple influence graphs describing how one change affects another, and a large(40k) collection of "What if...?" multiple-choice questions derived from these.

I use prompt template `what_might_be_the_first_step_of_the_process` provided by `promptsource`. A sample from this prompt template and dataset are shown in Appendix B.

2.1.2 Experiment Setting

After experiments on **GLM-RoBERTa-Large** with various hyperparameters, I use the following hyperparameters as table 9. All our models are trained and tested on 1 NVIDIA A100 (80GB) GPU.

Table 9: Hyperparameters for GLM-RoBERTa-Large's Generation Task

Hyperparams	**Values**
Fine-tune Epoch	3
Batch Size	64
Learning Rate	1e-6
Optimizer	Adam($\epsilon = 10^{-6}, \beta = (0.9, 0.98), weight\_decay = 0.1$)
Gradient Clipping	1.0
Scheduler	Linear w/ warmup
Warmup Steps	2000
Generation Policy	Beam Search Multinomial Sampling
Beam Size	1
Top-k	1

2.1.3 Result

Testset results of proposed methods are presented as table 10.

Table 10: Testset Results on GLM-RoBERTa-Large's Generation Task

Metrics	PPL↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑
Results	56.76	29.56	19.96	12.03	7.86

2.2 Grid Search for Hyperparameters and Analysis

2.2.1 Effect of Learning Rate

In this seciton, we experimented performances of different learning rates. Results on test dataset are as table 11. We can see that the accuracy on test dataset reaches its highest

Table 11: Testset Results on GLM-RoBERTa-Large's Generation Task using different Learning Rate

Max Learning Rate	PPL↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑
1e-4	676.80	11.59	4.75	2.93	2.03
1e-5	**46.71**	27.11	15.13	7.07	0.00
1e-6	56.76	**29.56**	**19.96**	**12.03**	**7.86**

score when maximum learning rate is `1e-6` during fine-tuning, where all BLEU scores are best. However, the perplexity reaches its best score when maximum learning rate is `1e-5`. BLEU scores change strongly when learning rate varies. We believe small learning rates can retain GLM-RoBERTa-Large's original strong ability of language modeling. Finally, we choose `1e-6` as our maximum learning rate to avoid overfitting problem.

2.2.2 Effect of Batch Size

In this seciton, we experimented performances of different training batch size. Results on test dataset are as table 12. We can see that GLM's BLEU score achieves best when batch

Table 12: Testset Results on GLM-RoBERTa-Large's Generation Task using different Batch Size

Batch Size	PPL↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑
16	102.56	24.81	15.13	10.19	6.25
64	**56.76**	**29.56**	**19.96**	**12.03**	**7.86**
128	74.32	27.29	16.13	8.65	4.72

size is 64. Too large batch size may lead to a local optimum, while too small batch size may cause a unstable gradient during fine-tuning.

2.2.3 Effect of Top-k

In this seciton, we experimented performances of different `top-k` during generation. Results on test dataset are as table 13. We can see that GLM's BLEU score varies with different

Table 13: Testset Results on GLM-RoBERTa-Large's Generation Task using different Top-k during inference

Top-k	PPL↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑
1	**56.76**	**29.56**	**19.96**	**12.03**	**7.86**
10	87.92	25.31	17.44	9.13	6.72
100	120.43	24.09	16.95	8.56	6.33

Top-k during inference. GLM achieves best BLEU scores when top-k is . With a larger top-k, generation in each step can sample from a larger space, leading to a more general but not so specific result, so larger top-k may have poor performance in Question-Answering Generation tasks.

2.2.4 Effect of Fine-tuning

In this seciton, we experimented performances of **prompt-based** method and **no-finetune** method. Results on test dataset are as as table 14. We can see that with further fine-tuning, GLM-RoBERTa-Large can achieve much higher BLEU scores on test dataset, indicating the necessity of fine-tuning.

Table 14: Testset Results on GLM-RoBERTa-Large's Generation Task with or without fine-tuning

Method	PPL↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑
Prompt-based	56.76	**29.56**	**19.96**	**12.03**	**7.86**
No-finetune-prompt	–	2.96	1.52	0.81	0.47

2.3 Different Prompts and Baselines

2.3.1 More Prompts

I use two more prompt templates, `what_might_be_the_last_step_of_the_process` and `what_is_the_missing_first_step`, besides `what_might_be_the_first_step_of_the_process` mentioned before. Samples from these two more templates and dataset are shown in Appendix B.

2.3.2 Comparasion with other GLMs

Similarly, we use facebook/bart-large and google/flan-t5-large as our baseline for text **generation** tasks. Both models are fine-tuned on the same datasets with hyperparameters proposed by their original paper.

2.3.3 Experiment Result on Different Prompts

Table 15: Testset Results on Different Prompts using GLM-RoBERTa-Large

Prompt	PPL↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑
`first_step`	56.76	**29.56**	**19.96**	**12.03**	**7.86**
`last_step`	39.66	22.67	12.91	7.85	3.70
`missing_first_step`	**38.47**	29.04	18.70	11.86	7.10

From table 15, we can see that GLM-RoBERTa-Large's performance varies significantly with different prompt template on generation tasks. It achieves 7.86 BLEU-4 score on template `first_step` but only reach 3.70 on template `last_step`. This may be becacuse GLM-RoBERTa-Large is more good at understanding texts at the beginning, but fails to understand those at the end.

2.3.4 Experiment Result on Different GLMs

Table 16: Testset Results on Different Models on dataset `WIQA` and prompt `first_step`

Model	Params	PPL↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑
Flan-T5-Large	780M	137.75	27.93	17.10	10.99	6.48
BART-Large	–	**50.65**	26.74	17.63	11.58	**7.86**
GLM-RoBERTa-Large	335M	56.76	**29.56**	**19.96**	**12.03**	**7.86**

From table 16, we can see that GLM-RoBERTa-Large achieves best BLEU scores compared with other popular GLMs such as Flan-T5 and BART with similar scale. This result shows "Autoregressive Blank Infilling" technique is promising, especially in natural language generation tasks.

As we can see, Flan-T5 and BART are both Transformer Encoder-Decoder models, with a encoder to perform understanding and a decoder to perform generation. However, GLM uses only a Decoder model, which takes fewer parameters to achieve a better performance.

A Datasets and Prompts Sample for Classification Task

Dataset QASC and Prompt is_correct_1

```
Sample 1:
Before:
Climate is generally described in terms of local weather conditions.

After:
If I tell you that Climate is generally described in terms of local
    weather conditions, and ask you the question "climate is generally
     described in terms of what?", is the correct answer "sand"?
    Answer: [MASK]

Answer: No
--------------------------------------------------
Sample 2:
Before:
Temperature and moisture is changing globally.

After:
If I tell you that Temperature and moisture is changing globally, and
    ask you the question "what is changing globally?", is the correct
    answer "the number of countries"? Answer: [MASK]

Answer: No
```

Dataset QASC and Prompt is_correct_2

```
Sample 1:
Before:
Climate is generally described in terms of local weather conditions.

After:
Do you think the right answer to the question "climate is generally
    described in terms of what?" is "occurs over a wide range", given
    that\n climate is generally described in terms of local weather
    conditions? Answer: [MASK]

Answer: No
--------------------------------------------------
Sample 2:
Before:
Temperature and moisture is changing globally.

After:
Do you think the right answer to the question "what is changing
    globally?" is "rapid growth", given that\n temperature and
    moisture is changing globally? Answer: [MASK]

Answer: No
```

Dataset GLEU/SST-2 and Prompt happy or mad

```
Sample 1:
Before:
it's a charming and often affecting journey.

After:
```

```
Someone sent me an email with the sentence "it's a charming and often affecting journey.". Do you think they are feeling good or bad? Answer: [MASK]

Answer: good
--------------------------------------------------
Sample 2:
Before:
unflinchingly bleak and desperate

After:
Someone sent me an email with the sentence "unflinchingly bleak and desperate ". Do you think they are feeling good or bad? Answer: [MASK]

Answer: bad
```

B Datasets and Prompts Sample for Generation Task

Dataset `WIQA` and Prompt `what_might_be_the_first_step_of_the_process`

```
Sample:
Before:
- Squirrel gains weight and fat
- Squirrel also hides food in or near its den
- Squirrels also grow a thicker coat as the weather gets colder
- Squirrel lives off of its excess body fat
- Squirrel uses its food stores in the winter.

After:
- Squirrel gains weight and fat
- Squirrel also hides food in or near its den
- Squirrels also grow a thicker coat as the weather gets colder
- Squirrel lives off of its excess body fat
- Squirrel uses its food stores in the winter.
What might be the first step of the process?
Answer: [Mask]

Answer:
Squirrels try to eat as much as possible
```

Dataset `WIQA` and Prompt `what_might_be_the_last_step_of_the_process`

```
Sample:
Before:
- Squirrel gains weight and fat
- Squirrel also hides food in or near its den
- Squirrels also grow a thicker coat as the weather gets colder
- Squirrel lives off of its excess body fat
- Squirrel uses its food stores in the winter.

After:
- Squirrel gains weight and fat
- Squirrel also hides food in or near its den
- Squirrels also grow a thicker coat as the weather gets colder
- Squirrel lives off of its excess body fat
- Squirrel uses its food stores in the winter.
What might be the last step of the process?
```

```
Answer: [Mask]

Answer:
Squirrel uses its food stores in the winter.
```

Dataset `WIQA` and Prompt `what_is_the_missing_first_step`

```
Sample:
Before:
- Squirrel gains weight and fat
- Squirrel also hides food in or near its den
- Squirrels also grow a thicker coat as the weather gets colder
- Squirrel lives off of its excess body fat
- Squirrel uses its food stores in the winter.

After:
What is the missing first step of the following process:
- Squirrel gains weight and fat
- Squirrel also hides food in or near its den
- Squirrels also grow a thicker coat as the weather gets colder
- Squirrel lives off of its excess body fat
- Squirrel uses its food stores in the winter.
Answer: [Mask]

Answer:
Squirrels try to eat as much as possible
```