NLP Final Project - GLUE

Group 5: Harish Ram, Zhuohan Yu, Sisi Zhang

Table of Contents

GLUE - General Language Understanding Evaluation

A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING - Consists of 9 tasks:

Single-Sentence

- CoLA
- SST-2

Similarity and Paraphrase

- MRPC
- STS-B
- QQP

Inference

- MNLI
- QNLI
- RTE
- WNLI

Dataset Description: Single Sentence Tasks

Sentence: Our friends won't buy this analysis, let alone the next one we propose.

Label: 0/1

Task: Grammatical Acceptability

CoLA

Train len: 8K Dev len: 1K

Sentence: hide new secretions from the parental units

Label: 0/1

Task: Sentiment Analysis

SST-2

Train len: 67K Dev len: 872

Dataset Description: Similarity and Paraphrase

S1: The DVD-CCA then appealed to the state Supreme Court

S2: The DVD CCA appealed that decision to the U.S. Supreme Court .

Label: 0/1

Task: Semantic Equivalence

MRPC

Train len: 3668 Dev len: 408

S1: A plane is taking off

S2: An air plane is taking off

Label: 0.0 - 5.0

Task: Semantic Similarity

STS-B

Train len: 5749 Dev len: 1500

Q1: What can one do after MBBS?

Q2: What can i do after my MBBS?

Label: 0/1

Task: Semantic Equivalence

QQP

Train len: 363K Dev len: 40K

Dataset Description: Inference Tasks

P: How do you know? All this is their information again.

H: This information belongs to them.

Label: entailment, neutral, contradiction

Task: Textual Entailment

MNLI

Train len: 302K Dev len: 10K

Q: When did the third Digimon series begin?

S: Unlike the two seasons before it and most of the seasons that...

Label: entailment / not-entailment

Task: Question-Answering

QNLI

Train len: 105K Dev len: 5K

Dataset Description: Inference Tasks

S1: No Weapons of Mass Destruction Found in Iraq Yet.

S2: Weapons of Mass Destruction Found in Iraq.

Label: entailment / not-entailment

Task: Textual Entailment

RTE

Train len: 2490 Dev len: 277

S1: I stuck a pin through a carrot. When I pulled the pin out, it had a hole.

S2: The carrot had a hole.

Label: 0/1

Task: Textual Entailment (Pronoun Replacement)

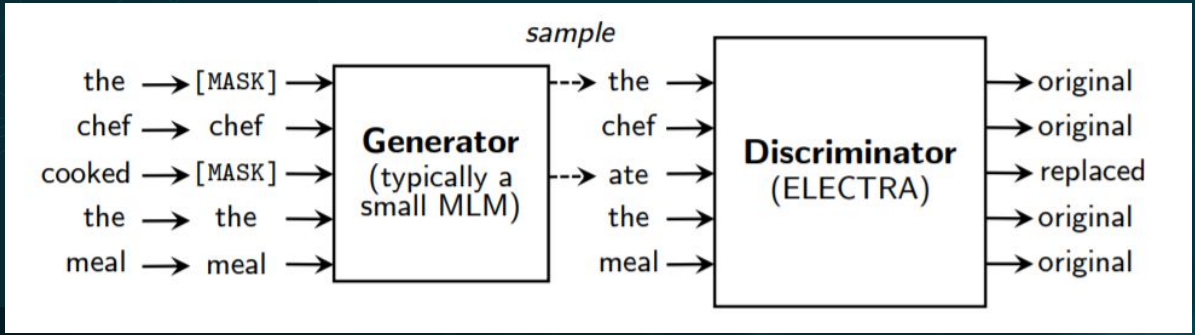
WNLI

Train len: 635 Dev len: 71

Model Description

ELECTRA

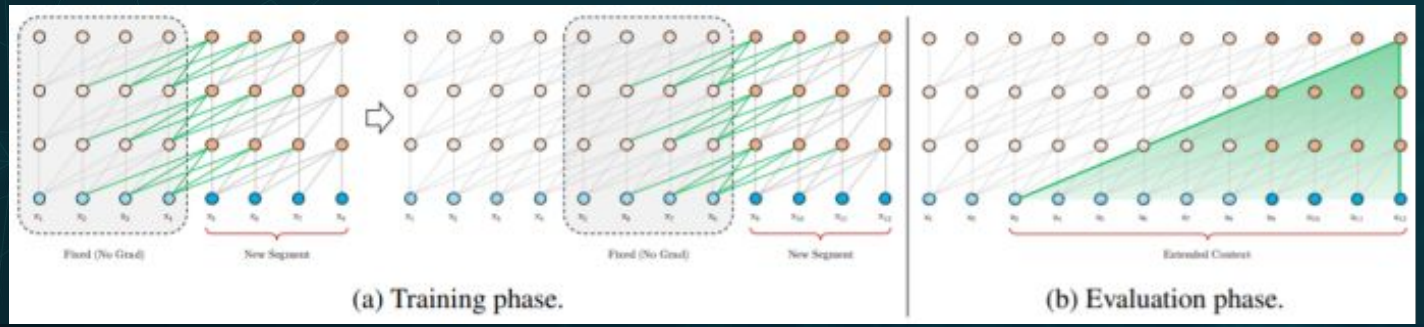
- Replaced Token Detection
- GAN and MLE
- Weight Sharing



Model Description

XLNet

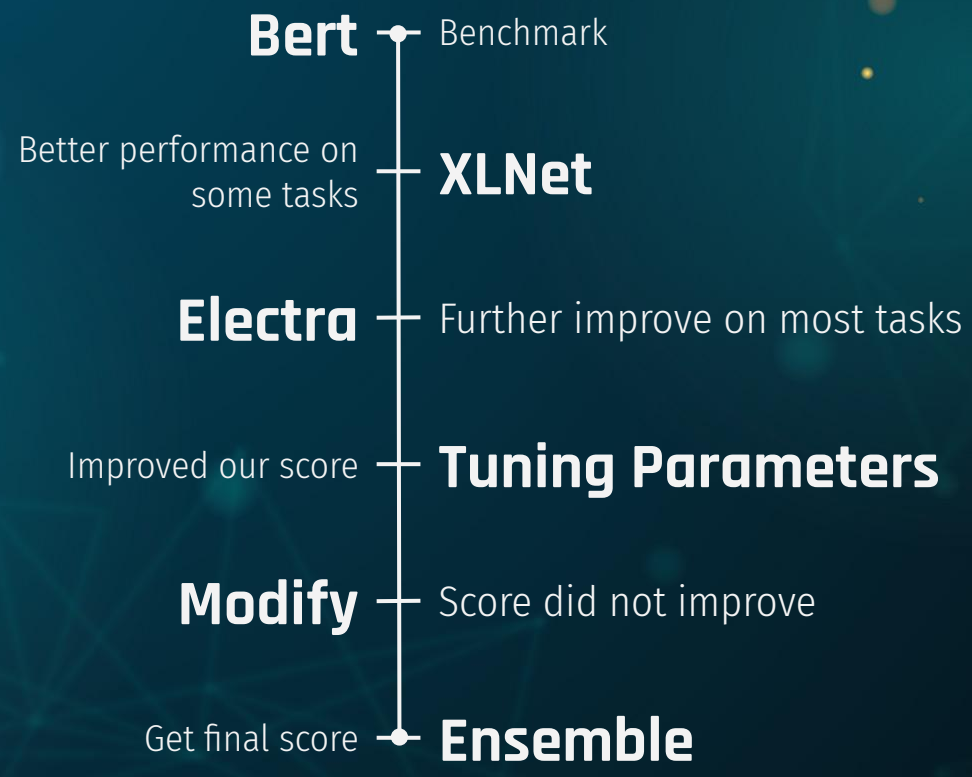
- Transformer XL
- Permutation Modeling
- Two-Stream Self-Attention



Experimental Setup

```
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset if training_args.do_train else None,
    eval_dataset=eval_dataset if training_args.do_eval else None,
    compute_metrics=compute_metrics,
    tokenizer=tokenizer,
    data_collator=data_collator,
)
```

Experimental Setup



Metrics

Matthew’s Correlation Coefficient - CoLA

1	0.704362
0	0.295638

Pearson Spearman correlation coefficient - STS-B

Accuracy - SST-2, MNLI, QNLI, RTE, WNLI

F1/Accuracy - MRPC , QQP

1	0.671166
0	0.328834

0	0.630673
1	0.369327

Hyperparameter Tuning

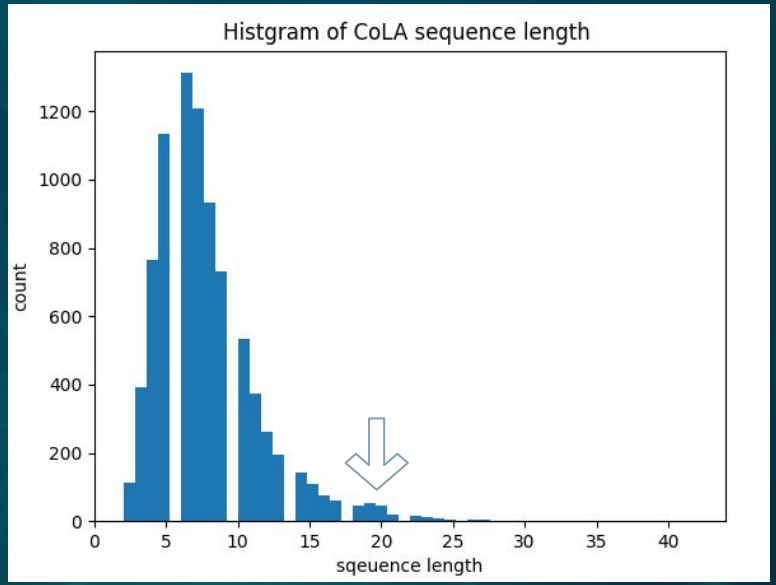
Batch Size - We tried 16, 20, **32**, 64 (higher score and less running time)

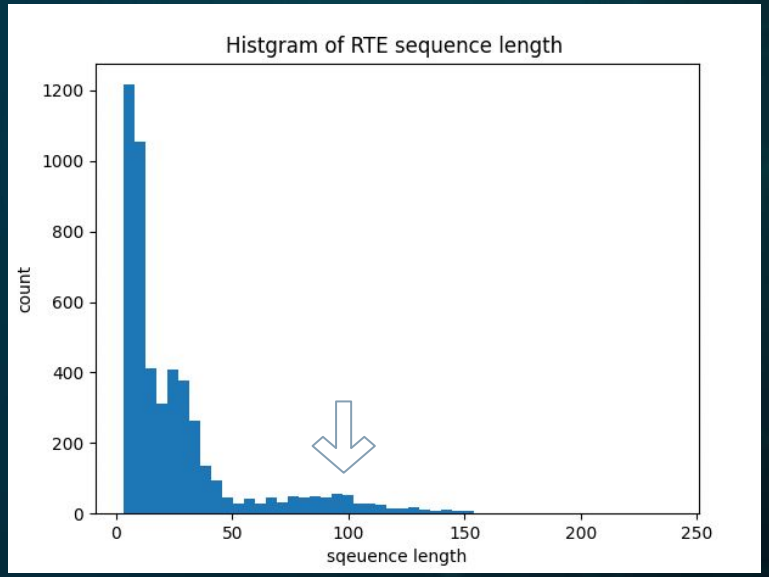
Maximum Sequence Length - Use histogram to decide

Learning Rate - 2e-4, **2e-5**, 2e-6 (same running time with higher score)

Number of Epochs - **3**, 5, 10 (22 mins , 38 mins, 41 mins with similar score)

Maximum Sequence Length





***** eval metrics *****		
epoch	=	3.0
eval_loss	=	0.4242
eval_matthews_correlation	=	0.6531
eval_runtime	=	0:00:01.88
eval_samples	=	1043
eval_samples_per_second	=	553.952
eval_steps_per_second	=	69.576

***** eval metrics *****		
epoch	=	3.0
eval_loss	=	0.4386
eval_matthews_correlation	=	0.6585
eval_runtime	=	0:00:09.22
eval_samples	=	1043
eval_samples_per_second	=	113.122
eval_steps_per_second	=	14.208

20 vs 128

***** eval metrics *****		
epoch	=	3.0
eval_accuracy	=	0.7184
eval_loss	=	0.5568
eval_runtime	=	0:00:01.87
eval_samples	=	277
eval_samples_per_second	=	147.575
eval_steps_per_second	=	18.647

***** eval metrics *****		
epoch	=	3.0
eval_accuracy	=	0.7256
eval_loss	=	0.5489
eval_runtime	=	0:00:13.54
eval_samples	=	277
eval_samples_per_second	=	20.457
eval_steps_per_second	=	2.585

100 vs 128

Benchmark

BERT Benchmark | XLNet 2nd best | ELECTRA Best

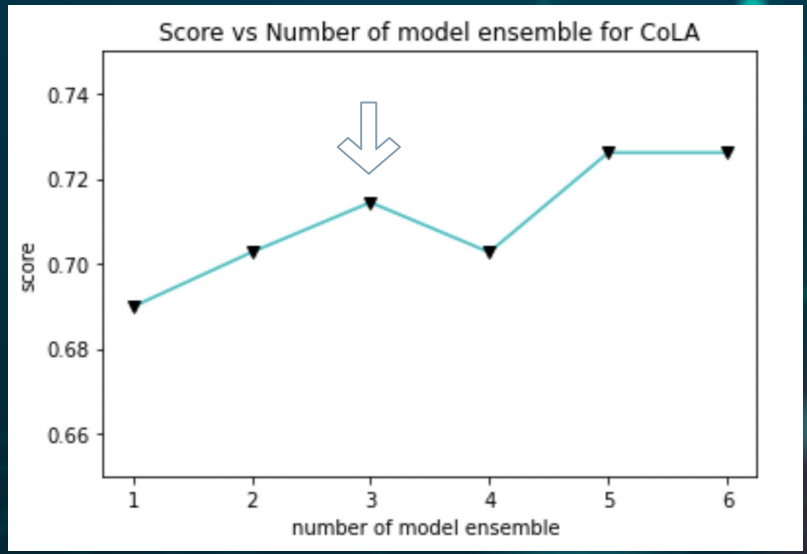
Task	Metric	Bert-base-cased	Xlnet-base-cased	Electra-base-discriminator
CoLA	Matthews corr	57.01	31.99	65.85
SST-2	Accuracy	92.43	94.15	95.41
MRPC	F1/Accuracy	89.46/84.8	89.76/85.29	90.88/86.76
STS-B	Pearson/Spearman corr.	88.82/88.5	86.73/86.56	90.31/90.35
QQP	Accuracy/F1	90.71/87.49	90.83/87.74	91.79/89.06
MNLI	Matched acc./Mismatched acc.	83.74/84.11	86.38/87.01	88.82/88.67
QNLI	Accuracy	90.32	92.02	92.77
RTE	Accuracy	64.98	64.98	72.56
WNLI	Accuracy	38.03	29.58	35.21

Ensemble-CoLA

	electra-0	electra-1	electra2-0	electra2-1	xlnet-0	xlnet-1	bert-0	bert-1	bert2-0	bert2-1	bert3-0	bert3-1	target
0	-3.015364	2.818316	-2.163078	2.018426	-2.792709	3.361839	-2.782302	3.292792	-2.141454	1.960070	-2.280861	2.351163	1
1	-3.272816	3.013487	-2.871374	2.883802	-3.019989	3.219829	-2.536592	3.144536	-2.195974	1.860165	-2.379374	2.406841	1
2	-2.484975	2.380282	-2.376473	2.256080	-1.700235	1.687799	-2.348618	2.650339	-2.242263	1.838615	-2.820595	2.764201	1
3	-3.138503	2.965038	-2.718466	2.658213	-2.994277	3.023010	-2.814234	3.324904	-2.432509	1.973530	-2.748433	2.449374	1
4	-2.365048	2.197786	-0.875185	0.664087	-2.657328	2.292092	2.783062	-3.252907	0.454222	-0.853928	1.091676	-0.697155	0

Xlnet	Bert2	Bert1	Bert3	Electra2	Electra1	Ensemble
0.4676	0.5754	0.5955	0.5981	0.6631	0.6774	0.7261
	0.5754	0.5955	0.5981	0.6631	0.6774	0.7261
		0.5955	0.5981	0.6631	0.6774	0.7027
			0.5981	0.6631	0.6774	0.7144
				0.6631	0.6774	0.7027
					0.6774	0.6898

We decide to ensemble 3 models due to time concern



Ensemble Result

Logistic Regression | Random Forest Classification | Linear Regression

	electra-0	electra-1	xlnet-0	xlnet-1	bert-0	bert-1	target
0	-3.015364	2.818316	-2.792709	3.361839	-2.782302	3.292792	1
1	-3.272816	3.013487	-3.019989	3.219829	-2.536592	3.144536	1
2	-2.484975	2.380282	-1.700235	1.687799	-2.348618	2.650339	1
3	-3.138503	2.965038	-2.994277	3.023010	-2.814234	3.324904	1
4	-2.365048	2.197786	-2.657328	2.292092	2.783062	-3.252907	0

Task	Metric	Benchmark(Bert)	XLnet	Electra	B+X+E
CoLA	Matthews corr	59.55	46.76	67.74	74.04
SST-2	Accuracy	92.20	93.81	94.95	94.85
MRPC	F1/Accuracy	90.16/86.03	91.36/87.99	91.46/88.24	94.92/92.68
STS-B	Pearson/Spearman corr.	85.19/85.13	88.56/88.34	90.74/90.57	90.57
QQP	Accuracy/F1	90.79/87.58	90.99/87.93	91.79/89.06	92.41/90.11
MNLI	Matched acc./Mismatched acc	0.8882/0.8867	0.8673/0.8646	0.8397/0.8436	89.22
QNLI	Accuracy	90.44	92.02	93.03	94.00
RTE	Accuracy	56.68	66.43	81.23	83.92
WNLI	Accuracy	56.34	53.52	56.34	46.66

Improvements

Large models

Histogram - max seq length

Model class with transformers as layer

Averaging predictions

Using test set to make predictions and upload to glue

References

Huggingface. (n.d.). Transformers/examples/pytorch/text-classification at master · Huggingface/Transformers. GitHub. Retrieved December 9, 2021, from https://github.com/huggingface/transformers/tree/master/examples/pytorch/text-classification.

Huggingface. (n.d.). Transformers/examples/pytorch/text-classification at master · Huggingface/Transformers. GitHub. Retrieved December 9, 2021, from https://github.com/huggingface/transformers/tree/master/notebooks.

Clark, M. Luong, Q. V. Le, and C. D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In ICLR.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Questions

?