GLUE Tasks

Group 3:
Abdulaziz Gebril, Yuan Meng, Yuchen Ma

Table of Contents

1. Introduction

- General Language Understanding Evaluation (GLUE) benchmark
  - Natural language understanding systems
  - Single-Sentence Similarity and Paraphrase Tasks, and Inference Tasks
- Widely used benchmark
- Solve the same problem with different teams worldwide to get more inspiration

2. Description of the dataset - Corpus of Linguistic Acceptability

The Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2018)

- 10,657 sentences from 23 linguistics publications

Dataset Preview

Subset: cola

Split: train

sentence (string)	label (class label)	idx (int)
Bill drank from the hose.	acceptable	36
This metal hammers flat easily.	acceptable	37
They made him president.	acceptable	38
They made him angry.	acceptable	39
They caused him to become angry by making him.	unacceptable	40
They caused him to become president by making him.	unacceptable	41
They made him to exhaustion.	unacceptable	42
They made him into a monster.	acceptable	43
The trolley rumbled through the tunnel.	acceptable	44
The wagon rumbled down the road.	acceptable	45
The bullets whistled past the house.	acceptable	46

Figure from Hugging Face datasets viewer, https://huggingface.co/datasets/glue/viewer/

2. Description of the dataset - The Stanford Sentiment Treebank

The Stanford Sentiment Treebank(SST-2) (Socher et al., 2013) - 9,645 sentences

Dataset Preview

Subset: sst2

Split: train

sentence (string)	label (class label)	idx (int)
hide new secretions from the parental units	negative	0
contains no wit , only labored gags	negative	1
that loves its characters and communicates something rather beautiful about human nature	positive	2
remains utterly satisfied to remain the same throughout	negative	3
on the worst revenge-of-the-nerds clichés the filmmakers could dredge up	negative	4
that 's far too tragic to merit such superficial treatment	negative	5
demonstrates that the director of such hollywood blockbusters as patriot games can still turn out a small , personal film with an emotional wallop .	positive	6
of saucy	positive	7
a depressed fifteen-year-old 's suicidal poetry	negative	8
are more deeply thought through than in most ` right-thinking ' films	positive	9

Figure from Hugging Face datasets viewer, https://huggingface.co/datasets/glue/viewer/

2. Description of the dataset - Microsoft Research Paraphrase Corpus

Microsoft Research Paraphrase Corpus for Similarity and Paraphrase Task (MRPC) (Dolan & Brockett, 2005) - 5,800 pairs of sentences

Dataset Preview

Subset: mrpc

Split: train

sentence1 (string)	sentence2 (string)	label (class label)	idx (int)
Amrozi accused his brother , whom he called " the witness " , of deliberately distorting his evidence .	Referring to him as only " the witness " , Amrozi accused his brother of deliberately distorting his evidence .	equivalent	0
Yucaipa owned Dominick 's before selling the chain to Safeway in 1998 for $ 2.5 billion .	Yucaipa bought Dominick 's in 1995 for $ 693 million and sold it to Safeway for $ 1.8 billion in 1998 .	not_equivalent	1
They had published an advertisement on the Internet on June 10 , offering the cargo for sale , he added .	On June 10 , the ship 's owners had published an advertisement on the Internet , offering the explosives for sale .	equivalent	2
Around 0335 GMT , Tab shares were up 19 cents , or 4.4 % , at A $ 4.56 , having earlier set a record high of A $ 4.57 .	Tab shares jumped 20 cents , or 4.6 % , to set a record closing high at A $ 4.57 .	not_equivalent	3
The stock rose $ 2.11 , or about 11 percent , to close Friday at $ 21.51 on the New York Stock Exchange .	PG & E Corp. shares jumped $ 1.63 or 8 percent to $ 21.03 on the New York Stock Exchange on Friday .	equivalent	4
Revenue in the first quarter of the year dropped 15 percent from the same period a year earlier .	With the scandal hanging over Stewart 's company , revenue the first quarter of the year dropped 15 percent from the same period a year earlier .	equivalent	5
The Nasdaq had a weekly gain of 17.27 , or 1.2 percent , closing at 1,520.15 on Friday .	The tech-laced Nasdaq Composite .IXIC rallied 30.46 points , or 2.04 percent , to 1,520.15 .	not_equivalent	6

Figure from Hugging Face datasets viewer, https://huggingface.co/datasets/glue/viewer/

2. Description of the dataset - Semantic Textual Similarity Benchmark

Semantic Textual Similarity Benchmark (STS-B) (Cer et al., 2017) - 8,628 sentence pairs

Dataset Preview

Subset: stsb

Split: train

sentence1 (string)	sentence2 (string)	label (float)	idx (int)
A plane is taking off.	An air plane is taking off.	5	0
A man is playing a large flute.	A man is playing a flute.	3.8	1
A man is spreading shreded cheese on a pizza.	A man is spreading shredded cheese on an uncooked pizza.	3.8	2
Three men are playing chess.	Two men are playing chess.	2.6	3
A man is playing the cello.	A man seated is playing the cello.	4.25	4
Some men are fighting.	Two men are fighting.	4.25	5
A man is smoking.	A man is skating.	0.5	6
The man is playing the piano.	The man is playing the guitar.	1.6	7
A man is playing on a guitar and singing.	A woman is playing an acoustic guitar and singing.	2.2	8
A person is throwing a cat on to the ceiling.	A person throws a cat on the ceiling.	5	9
The man hit the other man with a stick.	The man spanked the other man with a stick.	4.2	10

Figure from Hugging Face datasets viewer, https://huggingface.co/datasets/glue/viewer/

2. Description of the dataset - Quora Question Pairs

Quora Question Pairs (QQP) collecting question pairs from the website Quora

Dataset Preview

Subset: qqp

Split: train

question1 (string)	question2 (string)	label (class label)	idx (int)
How is the life of a math student? Could you describe your own experiences?	Which level of prepration is enough for the exam jlpt5?	not_duplicate	0
How do I control my horny emotions?	How do you control your horniness?	duplicate	1
What causes stool color to change to yellow?	What can cause stool to come out as little balls?	not_duplicate	2
What can one do after MBBS?	What do i do after my MBBS ?	duplicate	3
Where can I find a power outlet for my laptop at Melbourne Airport?	Would a second airport in Sydney, Australia be needed if a high-speed rail link was created between Melbourne and Sydney?	not_duplicate	4
How not to feel guilty since I am Muslim and I'm conscious we won't have sex together?	I don't beleive I am bulimic, but I force throw up atleast once a day after I eat something and feel guilty. Should I tell somebody, and if so who?	not_duplicate	5
How is air traffic controlled?	How do you become an air traffic controller?	not_duplicate	6
What is the best self help book you have read? Why? How did it change your life?	What are the top self help books I should read?	duplicate	7
Can I enter University of Melbourne if I couldn't achieve the guaranteed marks in Trinity College Foundation?	University of the Philippines: If I take a second BFA in the UP College of Fine Arts, can I be exempted from gen. ed. or core subjects?	not_duplicate	8

Figure from Hugging Face datasets viewer, https://huggingface.co/datasets/glue/viewer/

2. Description of the dataset - MultiNLI Matched

MultiNLI Matched (MNLI) (Williams et al., 2018) - 433,000 sentence pairs

Dataset Preview

Subset: mnli

Split: train

premise (string)	hypothesis (string)	label (class label)	idx (int)
Conceptually cream skimming has two basic dimensions - product and geography.	Product and geography are what make cream skimming work.	neutral	0
you know during the season and i guess at at your level uh you lose them to the next level if if they decide to recall the the parent team the Braves…	You lose the things to the following level if the people recall.	entailment	1
One of our number will carry out your instructions minutely.	A member of my team will execute your orders with immense precision.	entailment	2
How do you know? All this is their information again.	This information belongs to them.	entailment	3
yeah i tell you what though if you go price some of those tennis shoes i can see why now you know they're getting up in the hundred dollar range	The tennis shoes have a range of prices.	neutral	4
my walkman broke so i'm upset now i just have to turn the stereo up real loud	I'm upset that my walkman broke and now I have to turn the stereo up really loud.	entailment	5
But a few Christian mosaics survive above the apse is the Virgin with the infant Jesus, with the Archangel Gabriel to the right (his companion Michael…	Most of the Christian mosaics were destroyed by Muslims.	neutral	6
(Read for Slate 's take on Jackson's findings.)	Slate had an opinion on Jackson's findings.	entailment	7
Gays and lesbians.	Heterosexuals.	contradiction	8

Figure from Hugging Face datasets viewer, https://huggingface.co/datasets/glue/viewer/

2. Description of the dataset - Question NLI

Question NLI (QNLI) (Rajpurkar et al. 2016)

- more than 100,000 question-answer pairs from more than 500 articles

Dataset Preview

Subset: qnli

Split: train

question (string)	sentence (string)	label (class label)	idx (int)
When did the third Digimon series begin?	Unlike the two seasons before it and most of the seasons that followed, Digimon Tamers takes a darker and more realistic approach to its story…	not_entailment	0
Which missile batteries often have individual launchers several kilometres from one another?	When MANPADS is operated by specialists, batteries may have several dozen teams deploying separately in small sections; self-propelled air defence gun…	not_entailment	1
What two things does Popper argue Tarski's theory involves in an evaluation of truth?	He bases this interpretation on the fact that examples such as the one described above refer to two things: assertions and the facts to which they…	entailment	2
What is the name of the village 9 miles north of Calafat where the Ottoman forces attacked the Russians?	On 31 December 1853, the Ottoman forces at Calafat moved against the Russian force at Chetatea or Cetate, a small village nine miles north of Calafat, an…	entailment	3
What famous palace is located in London?	London contains four World Heritage Sites: the Tower of London; Kew Gardens; the site comprising the Palace of Westminster, Westminster Abbey, and St…	not_entailment	4
When is the term 'German dialects' used in regard to the German language?	When talking about the German language, the term German dialects is only used for the traditional regional varieties.	entailment	5
What was the name of the island the English traded to the Dutch in return for New Amsterdam?	At the end of the Second Anglo-Dutch War, the English gained New Amsterdam (New York) in North America in exchange for Dutch control of Run, an…	entailment	6

Figure from Hugging Face datasets viewer, https://huggingface.co/datasets/glue/viewer/

2. Description of the dataset - Recognizing Textual Entailment

Recognizing Textual Entailment (RTE) - combining the data from RTE1 (Dagan et al., 2006), RTE2 (Bar Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009)

Dataset Preview

Subset: rte

Split: train

sentence1 (string)	sentence2 (string)	label (class label)	idx (int)
No Weapons of Mass Destruction Found in Iraq Yet.	Weapons of Mass Destruction Found in Iraq.	not_entailment	0
A place of sorrow, after Pope John Paul II died, became a place of celebration, as Roman Catholic faithful gathered in downtown Chicago to mark…	Pope Benedict XVI is the new leader of the Roman Catholic Church.	entailment	1
Herceptin was already approved to treat the sickest breast cancer patients, and the company said, Monday, it will discuss with federal regulators the…	Herceptin can be used to treat breast cancer.	entailment	2
Judie Vivian, chief executive at ProMedica, a medical service company that helps sustain the 2-year-old Vietnam Heart Institute in Ho Chi Minh City…	The previous name of Ho Chi Minh City was Saigon.	entailment	3
A man is due in court later charged with the murder 26 years ago of a teenager whose case was the first to be featured on BBC One's Crimewatch.…	Paul Stewart Hutchinson is accused of having stabbed a girl.	not_entailment	4
Britain said, Friday, that it has barred cleric, Omar Bakri, from returning to the country from Lebanon, where he was released by police after being…	Bakri was briefly detained, but was released.	entailment	5
Nearly 4 million children who have at least one parent who entered the U.S. illegally were born in the United States and are U.S. citizens as a result,…	Three quarters of U.S. illegal immigrants have children.	not_entailment	6
Like the United States, U.N. officials are also dismayed that Aristide killed	[illegible]	[illegible]	[illegible]

Figure from Hugging Face datasets viewer, https://huggingface.co/datasets/glue/viewer/

2. Description of the dataset - Winograd NLI

Winograd NLI (WNLI) (Levesque et al., 2011) -150 Winograd schemas

Dataset Preview

Subset: wnli

Split: train

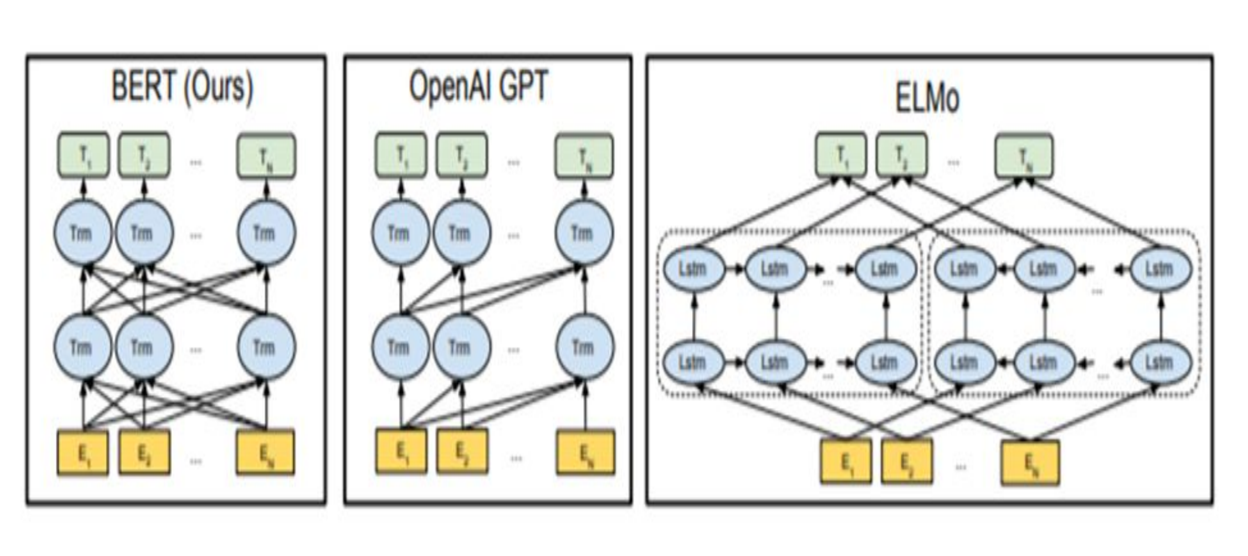
sentence1 (string)	sentence2 (string)	label (class label)	idx (int)
I stuck a pin through a carrot. When I pulled the pin out, it had a hole.	The carrot had a hole.	entailment	0
John couldn't see the stage with Billy in front of him because he is so short.	John is so short.	entailment	1
The police arrested all of the gang members. They were trying to stop the drug trade in the neighborhood.	The police were trying to stop the drug trade in the neighborhood.	entailment	2
Steve follows Fred's example in everything. He influences him hugely.	Steve influences him hugely.	not_entailment	3
When Tatyana reached the cabin, her mother was sleeping. She was careful not to disturb her, undressing and climbing back into her berth.	mother was careful not to disturb her, undressing and climbing back into her berth.	not_entailment	4
George got free tickets to the play, but he gave them to Eric, because he was particularly eager to see it.	George was particularly eager to see it.	not_entailment	5
John was jogging through the park when he saw a man juggling watermelons. He was very impressive.	John was very impressive.	not_entailment	6
I couldn't put the pot on the shelf because it was too tall.	The pot was too tall.	entailment	7

Figure from Hugging Face datasets viewer, https://huggingface.co/datasets/glue/viewer/

3. Description of the NLP model

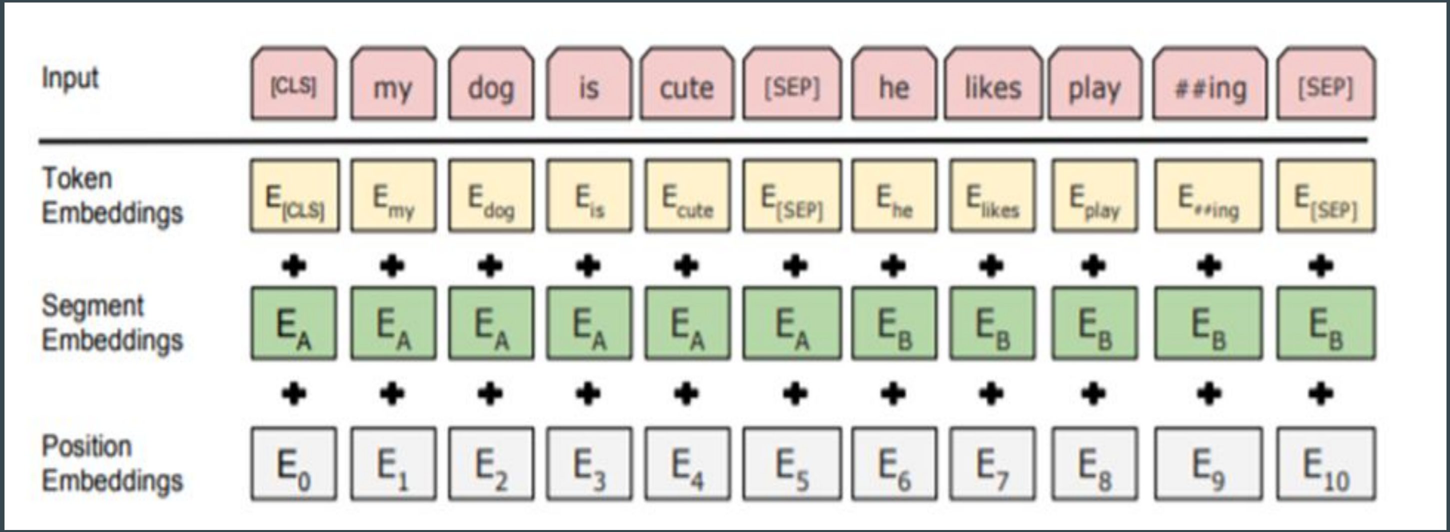
- BERT (Bidirectional Encoder Representations From Transformers)
    - Paper published by Google AI
    - BERT is pre-trained on both left and right context
    - Uses Masked language model and next sentence prediction objectives
    - Trained on Book corpus and English Wikipedia which contains 800 Million and 2,500 Million words.

BERT Vs OPENAI GPT Vs ELMO



https://arxiv.org/pdf/1810.04805.pdf

BERT Input Representation



https://arxiv.org/pdf/1810.04805.pdf

3. Description of the NLP model (Cont.)

- RoBERTa (Robustly Optimized BERT Pre Training Approach)
  - Paper published by Facebook and University of Washington.
  - Showed that BERT was significantly undertrained.
  - Longer Training, bigger batches, removing NSP, Dynamic MLM
  - Trained on Book corpus and English Wikipedia as BERT, In addition, RoBERTa was pretrained on CC-News, OpenWebText and Stories.

RoBERTa With and Without NSP

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6
$\text{BERT}_{\text{BASE}}$	88.5/76.3	84.3	92.8	64.3
$\text{XLNet}_{\text{BASE}}$ (K = 7)	–/81.3	85.8	92.7	66.1
$\text{XLNet}_{\text{BASE}}$ (K = 6)	–/81.0	85.6	93.4	66.7

https://arxiv.org/pdf/1907.11692.pdf

Static Vs Dynamic Masking

Masking	SQuAD 2.0	MNLI-m	SST-2
reference	76.3	84.3	92.8
Our reimplementation:			
static	78.3	84.3	92.5
dynamic	78.7	84.0	92.9

https://arxiv.org/pdf/1907.11692.pdf

3. Description of the NLP model (Cont.)

- CustomBERT & CustomRoBERTa
  - Used both pretrained BERT & RoBERTa and added two additional layers.

  - Bidirectional LSTM and output layer for different GLUE tasks

CustomBERT

```python
class CustomBERTModel(nn.Module):
    def __init__(self, num_labels):
        super(CustomBERTModel, self).__init__()
        self.bert = BertModel.from_pretrained("bert-base-uncased")
        self.hidden_size = self.bert.config.hidden_size
        self.lstm = nn.LSTM(self.hidden_size, self.hidden_size, batch_first=True,bidirectional=True)
        self.clf = nn.Linear(self.hidden_size*2, num_labels)

    def forward(self, **batch):
        sequence_output, pooled_output = self.bert(batch['input_ids'],batch['attention_mask'])[:2]
        # sequence_output has the following shape: (batch_size, sequence_length, self.hidden_size)
        lstm_output, (h,c) = self.lstm(sequence_output) ## extract the 1st token's embeddings
        hidden = torch.cat((lstm_output[:,-1, :self.hidden_size],lstm_output[:,0, self.hidden_size:]),dim=-1)
        hidden = F.dropout(hidden,0.1)
        linear_output = self.clf(hidden.view(-1,self.hidden_size*2)) ### only using the output of the last LSTM cell to perform classification
        return linear_output
```

4. Experimental setup - data preprocessing

- Auto Tokenizer
- Data Collator With Padding
- Remove & Rename Columns
  - attention_mask
  - input_ids
  - labels
  - token_type_ids
- Data Loader
  - Train & validation

Datasets	Column 1	Column 2
CoLA	Sentence	None
SST-2	Sentence	None
MRPC	Sentence1	Sentence 2
STS-B	Sentence1	Sentence 2
QQP	Question 1	Question 2
MNLI	Premise	Hypothesis
QNLI	Question	Sentence
RTE	Sentence1	Sentence 2
WNLI	Sentence1	Sentence 2

4. Experimental setup - data modeling

- Customize transformer models
  - Load pre-trained models
    - Sequence output shape: (batch size, sequence length, hidden size)
  - Add LSTM layer
    - Batch first and bidirectional
  - Define hidden layer and drop outs
    - Concatenate the LSTM outputs
  - Attach final linear layer
    - Output feature = number of labels
- Optimizer: AdamW
- Criterion: Cross-Entropy Loss & Mean Squared Error
- Number of training steps: number of epochs * length of train dataloader

4. Experimental setup - model evaluation

Task	Metric
CoLA	Matthew's Corr
SST-2	Accuracy
MRPC	F1 & Accuracy
STS-B	Pearson & Spearman
QQB	F1 & Accuracy
MNLI	Accuracy
QNLI	Accuracy
RTE	Accuracy
WNLI	Accuracy

5. Hyper-parameters

Task	Batch size	Epoch	Learning rate
CoLA	64	30	5e^5
SST-2	64	5	
MRPC	16	5	
STS-B	8	5	
QQP	32	5	
MNLI	32	5	
QNLI	32	5	
RTE	16	5	
WNLI	4	30	

6. Results

Task	Batch size	Epoch	Learning rate	BERT	RoBERTa
CoLA	64	30	5e^5	Matthew's Corr: 0.5931	Matthew's Corr: 0.6131
SST-2	64	5		Accuracy: 0.9243	Accuracy: 0.938
MRPC	16	5		F1: 0.9091 Accuracy: 0.8701	F1:0.927 Accuracy: 0.8995
STS-B	8	5		Pearson: 0.8780 Spearman: 0.8757	Pearson:0.90488 Spearman Corr: 0.902
QQP	32	5		F1: 0.8789 Accuracy: 0.9095	F1: 0.885 Accuracy:0.9144
MNLI	32	5		Accuracy: 0.8338	Accuracy: 0.865
QNLI	32	5		Accuracy: 0.9028	Accuracy: 0.9218
RTE	16	5		Accuracy: 0.6245	Accuracy:0.711
WNLI	4	30		Accuracy: 0.5493	Accuracy: 0.577

6. Results (Cont.)

From the result, we can find that the BERT algorithm and RoBERTa algorithm are good for SST-2 task, MPRC task, QQP task and QNLI task especially. For these tasks, we all get high scores from our models. Comparing the results we get from the BERT model and RoBERTa mode, we can find that for each task, RoBERTa mode always has a better result, although the difference is not much. As mentioned, to improve the training procedure, RoBERTa removes the Next Sentence Prediction (NSP) task from BERT’s pre-training and introduces dynamic masking so that the masked token changes during the training epochs. Moreover, RoBERTa uses larger text for pre-training. Therefore, we can conclude that RoBERTa model outperforms the BERTA model.

7. Summary and Conclusion

In the project, we learn how to build the architecture for Bert and RoBerta and analyze natural language through nine tasks. For each task, we use different metrics to evaluate their performance. When we run the code, we also met some problems due to large size of the task, so we change hyperparameters and run in GPU. From metrics, we conclude that Bert and Roberta both work well for nine tasks and Roberta works better. There are still many transformers which have similar functions. It is also interesting for us to explore in the future.

8. References

1. GLUE benchmark, https://gluebenchmark.com/tasks
2. Hugging Face, https://huggingface.co/models
3. BERT: Pre-training of Deep Bidirectional Transformers For Language Understanding. https://arxiv.org/pdf/1810.04805.pdf
4. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/pdf/1907.11692.pdf