GLUE Benchmark

By Nigel Martis, Rohan Paul, Bradley Reardon, Jongchan Kim

Table of Contents

Introduction

- The General Language Understanding Evaluation (GLUE) benchmark
- Collection of resources for training, evaluating, and analyzing natural language understanding systems
- Nine natural language processing (NLP) tasks
    - Sentence or sentence-pair data
- Assess model performance
- Compare performance on global leaderboard
- We tested the DeBERTa and RoBERTa transformer models

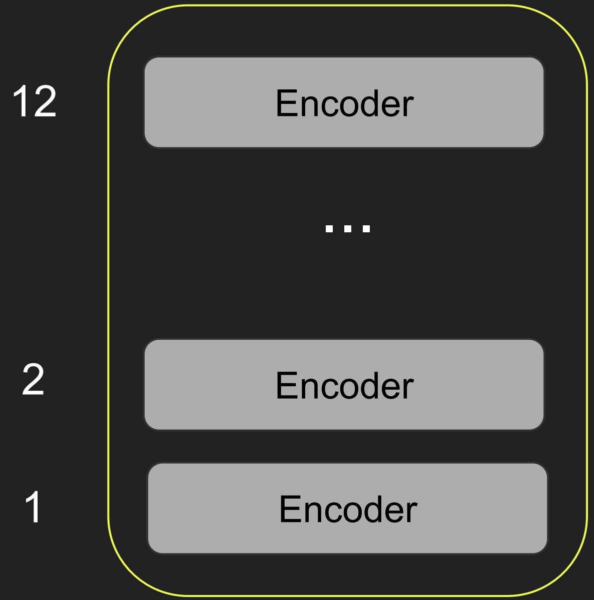
Dataset Overview

- 9 tasks
- Train set
- Test set
- Validation set
- Sentences and sentence pairs

Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**	Accuracy / F1
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**	Pearson / Spearman
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**	Accuracy
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**	Accuracy

BERT

Use Only Encoder part of transformer of 12 layers.



RoBERTa

a retraining of BERT with improved training methodology, 1000% more data and compute power.

- Removes the NSP(Next Sentence Prediction) task
- Dynamic masking : the masked token changes
- Larger batch size, more data

RoBERTa

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

DeBERTa

DeBERTa (Decoding-enhanced BERT with disentangled attention) improves the BERT and RoBERTa models using two novel techniques.

1) **Disentangled attention mechanism**

Each word is represented using two vectors for contents, positions

- 2 Key : Key (content), Key(position)
- 2 Query: Query(content), Query(position)
- 1 Value

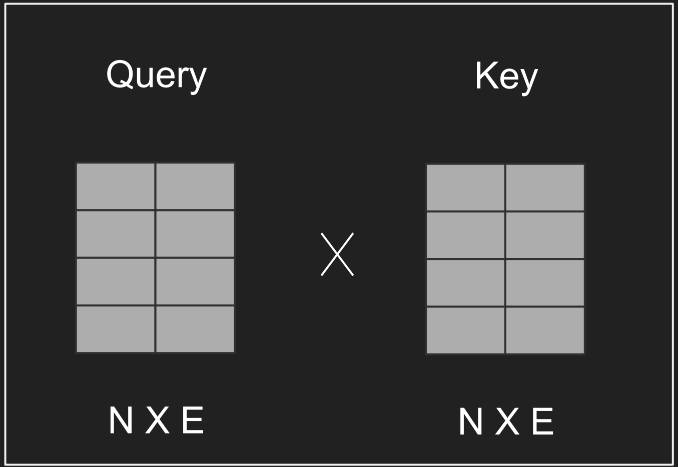


Fig1. Query X Key for Attention Score

N : number of tokens in a sentence
E : dimension of embedding

DeBERTa

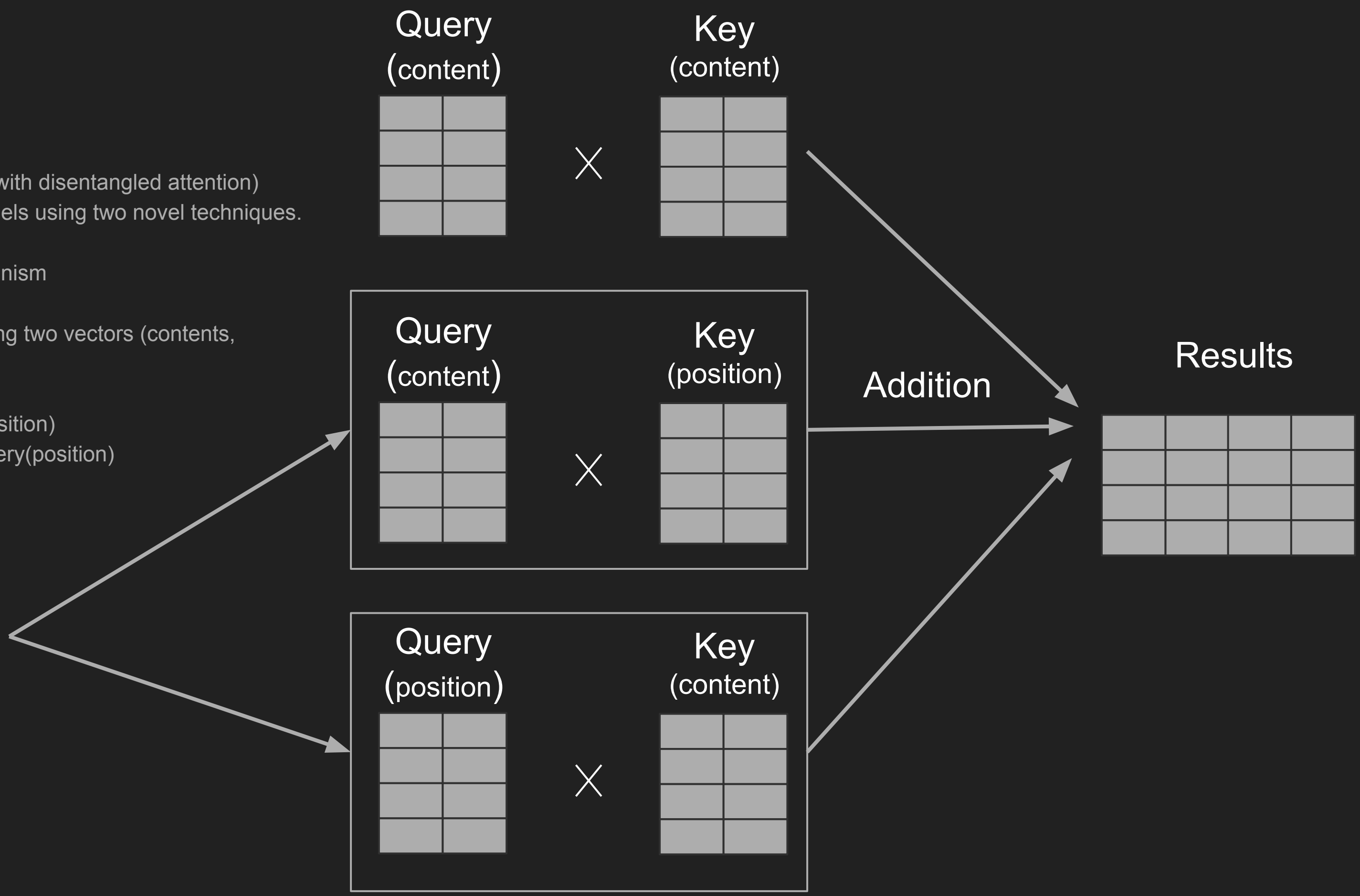
DeBERTa (Decoding-enhanced BERT with disentangled attention) improves the BERT and RoBERTa models using two novel techniques.

1) Disentangled attention mechanism

Each word is represented using two vectors (contents, positions)

- 2 Key : Key (content), Key(position)
- 2 Query: Query(content), Query(position)
- 1 Value

Matrix modification using Distance matrix :choose elements based on relative distance between tokens



DeBERTa **Disentangled attention mechanism**

Algorithm 1 Disentangled Attention

Input: Hidden state $\boldsymbol{H}$, relative distance embedding $\boldsymbol{P}$, relative distance matrix $\boldsymbol{\delta}$. Content projection matrix $\boldsymbol{W_{k,c}}$, $\boldsymbol{W_{q,c}}$, $\boldsymbol{W_{v,c}}$, position projection matrix $\boldsymbol{W_{k,r}}$, $\boldsymbol{W_{q,r}}$.

$\boldsymbol{K_c} = \boldsymbol{HW_{k,c}}, \boldsymbol{Q_c} = \boldsymbol{HW_{q,c}}, \boldsymbol{V_c} = \boldsymbol{HW_{v,c}}, \boldsymbol{K_r} = \boldsymbol{PW_{k,r}}, \boldsymbol{Q_r} = \boldsymbol{PW_{q,r}}$

$\boldsymbol{A_{c\to c}} = \boldsymbol{Q_c K_c^{\intercal}}$

for $i = 0, ..., N-1$ **do**

  $\boldsymbol{\tilde{A}_{c\to p}}[i,:] = \boldsymbol{Q_c}[i,:]\boldsymbol{K_r^{\intercal}}$

end for

for $i = 0, ..., N-1$ **do**

  for $j = 0, ..., N-1$ **do**

    $\boldsymbol{A_{c\to p}}[i,j] = \boldsymbol{\tilde{A}_{c\to p}}[i, \boldsymbol{\delta}[i,j]]$

  end for

end for

for $j = 0, ..., N-1$ **do**

  $\boldsymbol{\tilde{A}_{p\to c}}[:,j] = \boldsymbol{K_c}[j,:]\boldsymbol{Q_r^{\intercal}}$

end for

for $j = 0, ..., N-1$ **do**

  for $i = 0, ..., N-1$ **do**

    $\boldsymbol{A_{p\to c}}[i,j] = \boldsymbol{\tilde{A}_{p\to c}}[\boldsymbol{\delta}[j,i], j]$

  end for

end for

$\boldsymbol{\tilde{A}} = \boldsymbol{A_{c\to c}} + \boldsymbol{A_{c\to p}} + \boldsymbol{A_{p\to c}}$

$\boldsymbol{H_o} = \mathrm{softmax}(\frac{\boldsymbol{\tilde{A}}}{\sqrt{3d}})\boldsymbol{V_c}$

Output: $\boldsymbol{H_o}$

DeBERTa

DeBERTa (Decoding-enhanced BERT with disentangled attention) improves the BERT and RoBERTa models using two novel techniques.

2) Enhanced mask decoder

BERT : incorporates absolute positions in the input layer (positional encoding)

DeBERTa : incorporates them in right after all the Transformer layers but before the softmax layer

Results - RoBERTa

RoBERTa					
Task	Optimizer	Epoch	Batch Size	Learning Rate	Metrics
CoLA	AdamW	4	32	0.00005	Acc: 0.84, Mat Cor: 0.63
SST-2	AdamW	4	64	0.00005	Acc: 0.94
MRPC	AdamW	4	8	0.00005	Acc: 0.89, F1: 0.92
STS-B	AdamW	4	8	0.00005	Pearson: 0.91, Spearman: 0.90
QQP	AdamW	1	8	0.00005	Acc: 0.89
MNLI-m	AdamW	4	64	0.00005	Acc: 0.88
MNLI-mm	AdamW	4	64	0.00005	Acc: 0.87
QNLI	AdamW	4	8	0.00005	Acc: 0.92
RTE	AdamW	4	8	0.00005	Acc: 0.52
WNLI	AdamW	4	8	0.00005	Acc: 0.56

Results - DeBERTa

DeBERTa					
Task	Optimizer	Epoch	Batch Size	Learning Rate	Metrics
CoLA	AdamW	4	32	0.00005	Acc: 0.90, Mat Cor: 0.67
SST-2	AdamW	4	64	0.00005	Acc: 0.95
MRPC	AdamW	4	8	0.00005	Acc: 0.85, F1: 0.89
STS-B	AdamW	4	8	0.00005	Pearson: 0.90, Spearman: 0.90
QQP	AdamW	1	8	0.00005	Acc: 0.90
MNLI-m	AdamW	4	64	0.00005	Acc: 0.89
MNLI-mm	AdamW	4	64	0.00005	Acc: 0.88
QNLI	AdamW	4	8	0.00005	Acc: 0.92
RTE	AdamW	4	8	0.00005	Acc: 0.55
WNLI	AdamW	4	8	0.00005	Acc: 0.56

Results Summary

- DeBERTa outperformed RoBERTa in:
  - COLA
  - SST-2
  - STSB
  - QQP
  - MNLI matched
  - MNLI mismatched
  - RTE
- Equal performance in:
  - QNLI
  - WNLI
- Lower performance in:
  - MRPC

Transformer Model Comparison		
Model	Task	Score
DeBERTa	CoLA	Acc: 0.90, Mat Cor: 0.67
RoBERTa	CoLA	Acc: 0.84, Mat Cor: 0.63
DeBERTa	SST-2	Acc: 0.95
RoBERTa	SST-2	Acc: 0.94
DeBERTa	MRPC	Acc: 0.85, F1: 0.89
RoBERTa	MRPC	Acc: 0.89, F1: 0.92
DeBERTa	STS-B	Pearson: 0.90, Spearman: 0.90
RoBERTa	STS-B	Pearson: 0.91, Spearman: 0.90
DeBERTa	QQP	Acc: 0.90
RoBERTa	QQP	Acc: 0.89
DeBERTa	MNLI-m	Acc: 0.89
RoBERTa	MNLI-m	Acc: 0.88
DeBERTa	MNLI-mm	Acc: 0.88
RoBERTa	MNLI-mm	Acc: 0.87
DeBERTa	QNLI	Acc: 0.92
RoBERTa	QNLI	Acc: 0.92
DeBERTa	RTE	Acc: 0.55
RoBERTa	RTE	Acc: 0.52
DeBERTa	WNLI	Acc: 0.56
RoBERTa	WNLI	Acc: 0.56

Top 3 Current Leaderboard Results vs. Ours

URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI
	91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9
	91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9
	91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2
Our DeBERTa		67	95	85, 89	90, 90	90	89	88	92	55	56

Conclusion

- Model Performance:
  - DeBERTa outperformed the RoBERTa model in our experiments.
  - Our scores on validation sets were close to the test scores on the GLUE benchmark leaderboard for the majority of tasks.
- Areas for Improvement:
  - Testing other models
  - Hyperparameter tuning
  - Ensembling models
- Limitations:
  - Time constraints
  - Computing power

Reference

[1] He, Pengcheng, et al. "Deberta: Decoding-enhanced bert with disentangled attention." arXiv preprint arXiv:2006.03654 (2020).

[2] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

1. https://gluebenchmark.com/
2. https://huggingface.co/docs/transformers/model_doc/deberta
3. https://huggingface.co/docs/transformers/model_doc/roberta
4. https://github.com/huggingface/transformers/blob/5bfcd0485ece086ebcbed2d008813037968a9e58/examples/run_glue.py#L128
5. https://www.kaggle.com/aryansakhala/bert-pytorch-cola-classification#Evaluate-on-Test-Set