Combining Shallow and Deep Representations for Text-Pair Classification

Vincent Nguyen[1,2], **Sarvnaz Karimi**[2], **Zhenchang Xing**[1]
Australian National University[1] and CSIRO's Data61[2]
Vincent.Nguyen@anu.edu.au
https://ngu.vin

Background

Text-pair classification determines the class relationship between two pieces of text (e.g., two sentences).

Motivation

- Some methods in text-pair classification use transformer encoders with a dense fully connected layer over the classification token.
- Improvements to transformer encoders often involve scaling the model capacity through either dataset source or size, pretraining task changes and increasing model parameter count.
- However, the *classification token* is typically the only feature used for classification.
- Our work leverages more representations, from shallow model layers and deep model representations from the upper layers, for classification without significantly changing inference or training time, and model parameter count.

Datasets

MEDIQA

Natural Language Inference

Natural Language Inference (NLI) can be used to validate if the answer can be inferred from the question.

Premise	Hypothesis	Label
She was not able to speak , but appeared to comprehend well	Patient had aphasia	entailment
Had an ultimately negative esophagogastroduodenoscopy and colonoscopy	Patient has no pain	neutral
Aorta is mildly tortuous and calcified	the aorta is normal	contradiction

Recognising Question Entailment

Q1: Can you mail me patient information about Glaucoma, I was recently diagnosed and want to learn all I can about the disease.
Q2: How is glaucoma diagnosed?

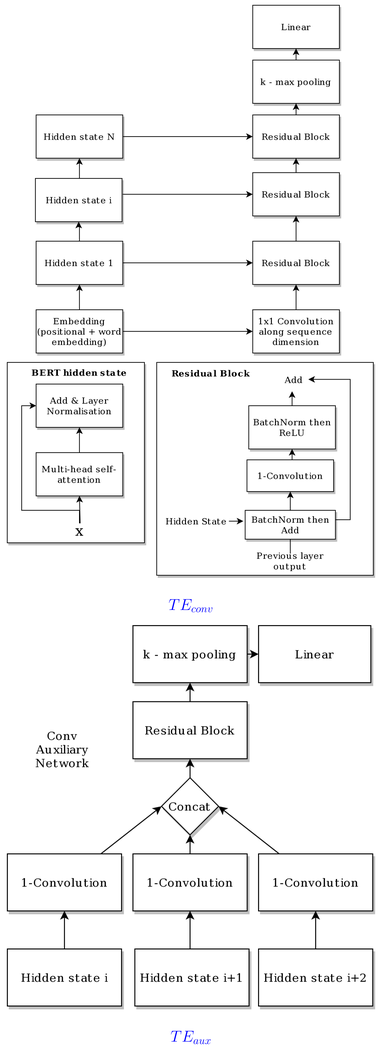
Stack Overflow

Original Question Conversion Error setting value for 'null Converter' - Why do I need a Converter in JSF?
Duplicate Question selectOneMenu with complex objects, is a converter necessary?
Negative Sample Conversion Error setting value '1' for 'null Converter'

General Domain

We use the SNLI dataset and Quora Duplicate Questions dataset from the general domain to assess generalisability.

Architecture



Main Problems

- Contemporary transformer-based language models do not leverage different levels of representation in a model.
- **Proposal:** Leverage shallow and deep representations from all layers in the encoder for use in the final prediction.

Results

		Method	A	P	R	F1
Open-Domain	SNLI	TE	0.798	0.599	0.609	0.604
		TE_{conv}	**0.869**†	**0.653**‡	**0.657**†	**0.657**
		TE_{aux}	0.830	0.624	0.633	0.627
		SOTA	0.923 (Liu et al., 2019b)			
	Quora	TE	0.811	0.739	0.755	0.747
		TE_{conv}	**0.880**‡	**0.842**‡	0.832‡	0.836‡
		TE_{aux}	0.879‡	0.811‡	**0.878**‡	**0.843**‡
		SOTA	0.923 (Yang et al., 2019)			
MediQA	NLI	TE	0.335	0.112	0.333	0.170
		TE_{conv}	**0.797**‡	**0.797**‡	**0.797**‡	**0.797**‡
		TE_{aux}	0.728‡	0.761‡	0.727‡	0.723‡
		SOTA	0.980 (Ben Abacha et al., 2019)			
	RQE	TE	0.557	0.278	0.500	0.358
		TE_{conv}	0.536	0.567‡	0.535	0.490‡
		TE_{aux}	**0.911**†	**0.9416**‡	**0.925**‡	**0.908**‡
		SOTA	0.749 (Ben Abacha et al., 2019)			
	QA	TE	0.575	0.287	0.500	0.365
		TE_{conv}	0.718	0.714‡	0.713‡	0.709‡
		TE_{aux}	**0.947**‡	**0.944**‡	**0.947**‡	**0.945**‡
		SOTA	0.783 (Ben Abacha et al., 2019)			
Stack Overflow	DQD	TE	0.919	0.929	0.907	0.918
		TE_{conv}	**0.943**	**0.960**	**0.926**	**0.942**
		TE_{aux}	0.939	0.952	0.924	0.938‡
Average		TE	0.667	0.502	0.592	0.529
		TE_{conv}	0.779	0.743	0.729	0.724
		TE_{aux}	**0.846**	**0.809**	**0.810**	**0.798**

Results on various datasets with BERT and Convolutional BERT variants.

Key findings

- Additional features from the lower layers aids in generalisation and allow the model to better understand syntax, and numerical structure (medical charts) for text pair classification.
- We observe increased gradient propagation to early parts of the network which aids in training.
- Even when the encoder is untrained, leveraging more layer representations aids downstream text pair classification.

Future Work

- Evaluating over more datasets
- Evaluating methodology over other transformer encoders

Acknowledgments

This research is supported by the Australian Research Training Program and the CSIRO Postgraduate Scholarship and CSIRO's Future Science platform for Precision Health.