NLP: Transformers, BERT and GLUE

Gayle Tan

NLP, Transformers in our daily lives

- Predictions on smartphone keyboards
- Gmail autocomplete
- Chatbots
- Language Translation

Some Terminologies

Semantics

- The meaning of a word or sentence

Syntax

- Rules of a language/grammar, how words can be combined to form sentences

Language Model

- Takes input and predicts the next word or sentence

Transformers

Why use transformers?

Transformers were proposed to solve the issue of RNNs

- RNNs (LSTM, GRU) could remember only parts of the sentence if there are too many tokens.
- It was sequential, so parallelisation/optimisation was not possible.

In a transformer:

All words are processed simultaneously, this allows parallelisation

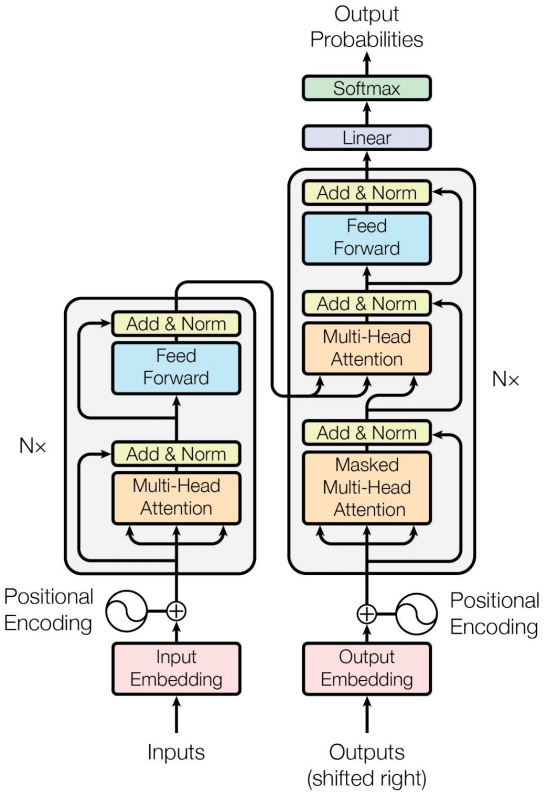
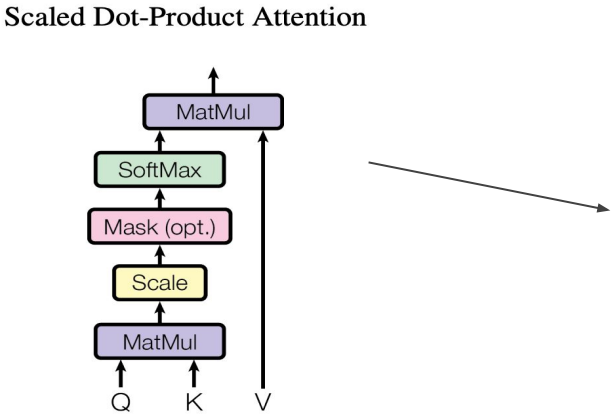
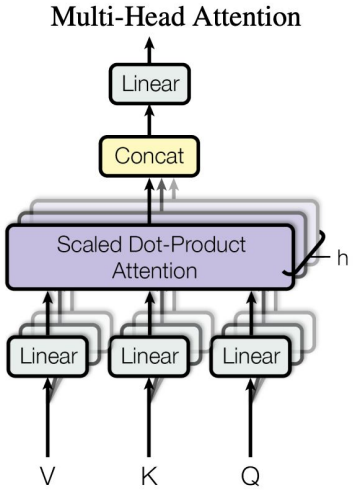


Figure 1: The Transformer - model architecture.

Transformers - Under the Hood





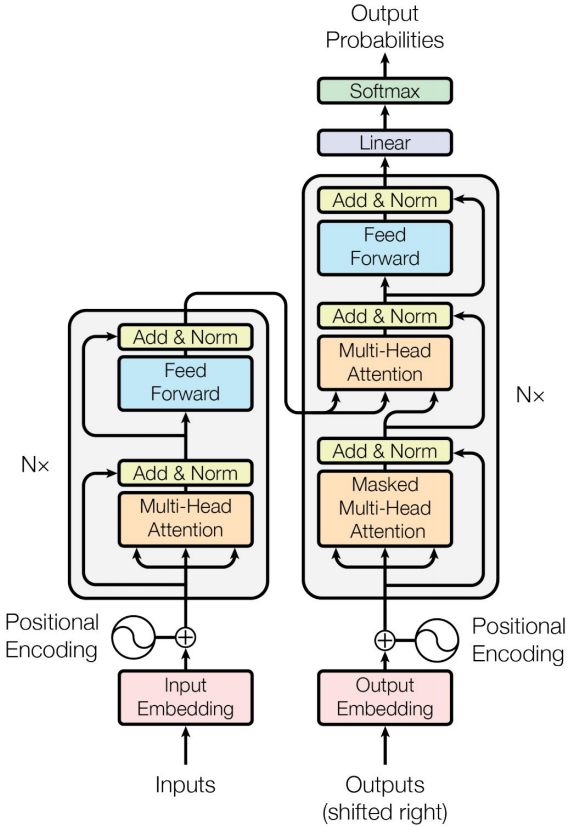


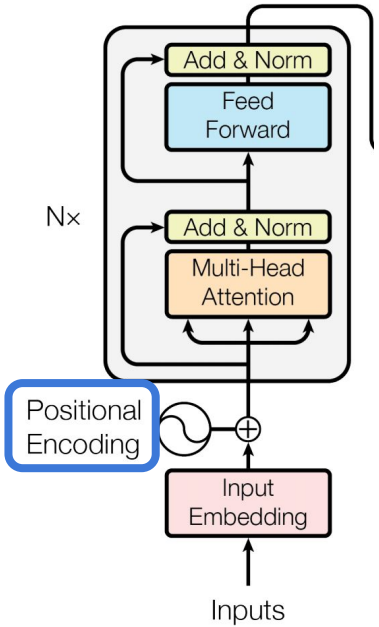
Figure 1: The Transformer - model architecture.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Query: current word and position
Key: all word position vectors
Value: all word position vectors

"Attend to" Weights at different positions

Transformers - Positional Encoding



Transformer Components

Positional Encoder :vector that gives context based on position of word in sentence

$$\begin{bmatrix}0.37\\0.99\\0.01\\0.08\end{bmatrix} + \text{Positional Encoding} \rightarrow \begin{bmatrix}0.42\\0.84\\0.12\\0.81\end{bmatrix}$$

Embedding of "Dog"

Vector Encoding of position in sentence

Embedding of Dog (with context info)

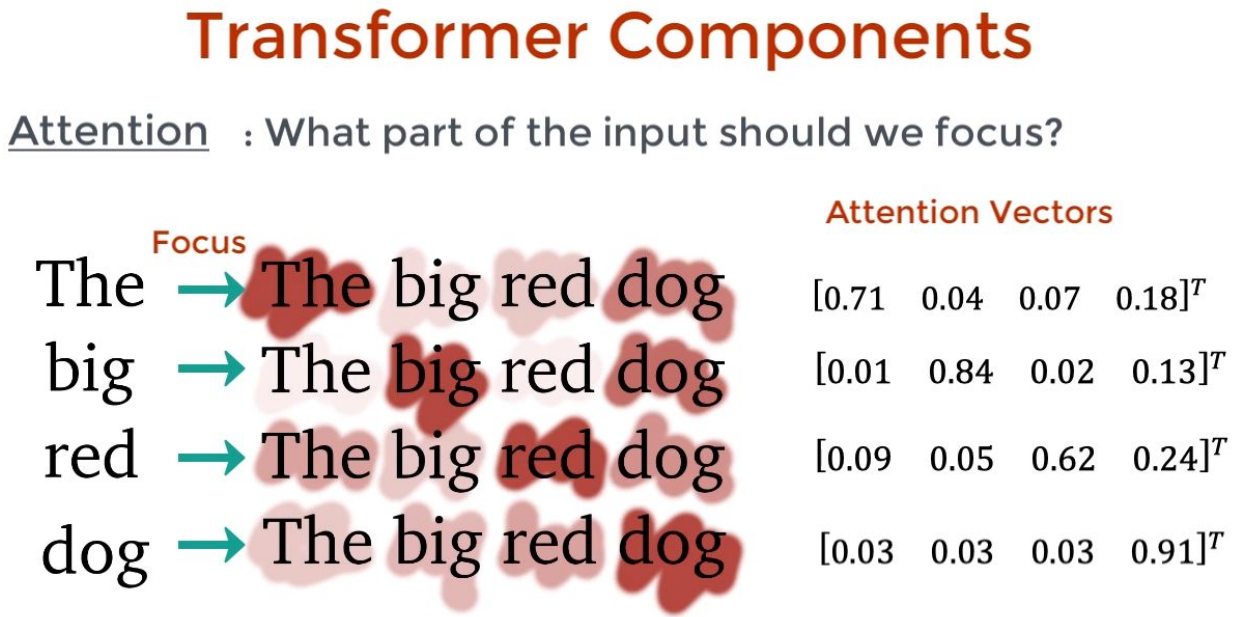
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

From youtube video:
https://www.youtube.com/watch?v=TQQIZhbC5ps

Transformers - Attention Vector

Attention Vector =

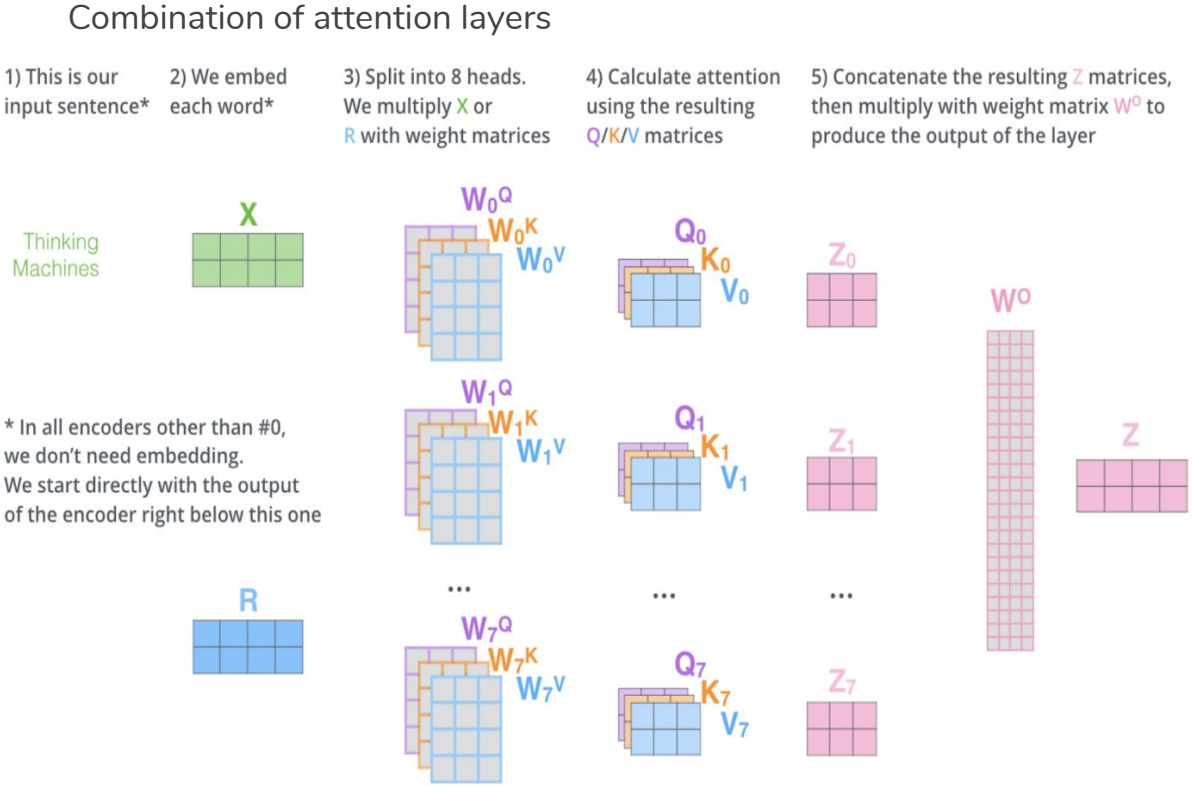
How important is each word within a sentence

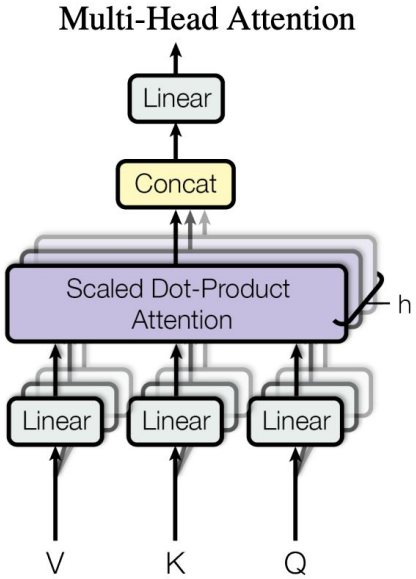


From youtube video:
https://www.youtube.com/watch?v=TQQIZhbC5ps

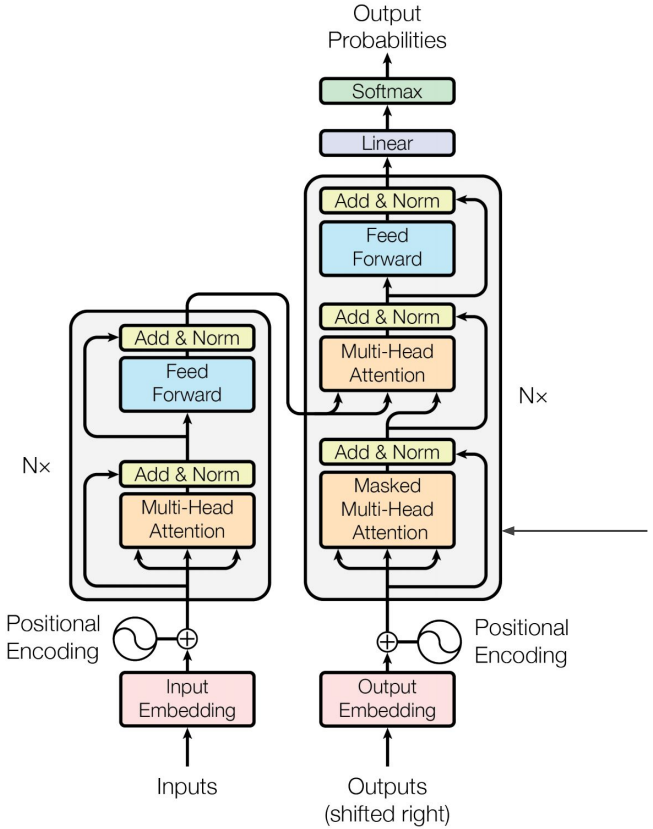
Transformers - Multi-Head Attention

Combination of attention layers





Transformers - Architecture



Masked - so that we do not use future info to predict/translate

Figure 1: The Transformer - model architecture.

BERT - overview (1/3)

Bidirectional Encoder Representations from Transformers

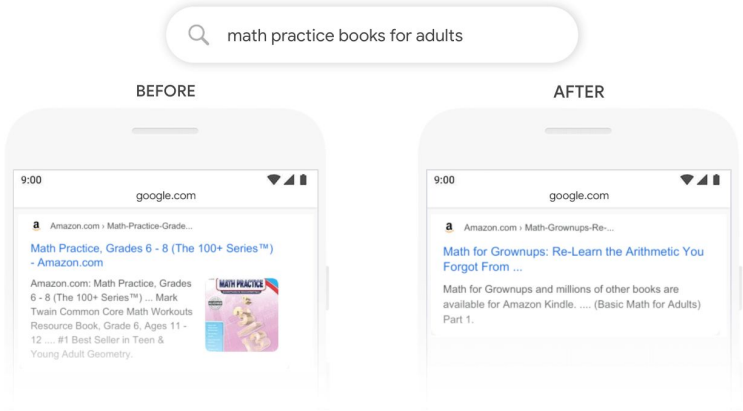
What is it:
A set of pre-trained language models, developed by Google in 2018.

Where is it used:
It was rolled out in Google Search in Oct 2019.

Why is it popular:
Can be used for multiple tasks and fine tuning is computationally inexpensive (<1h with a TPU)

Types of Tasks it can help with:

1. Sentence Classification
2. Sentiment Analysis
3. Translation
4. Summarisation



BERT - overview (2/3)

Multi-layer Transformer Encoder with Bi-directionality

Reads entire sentence at once

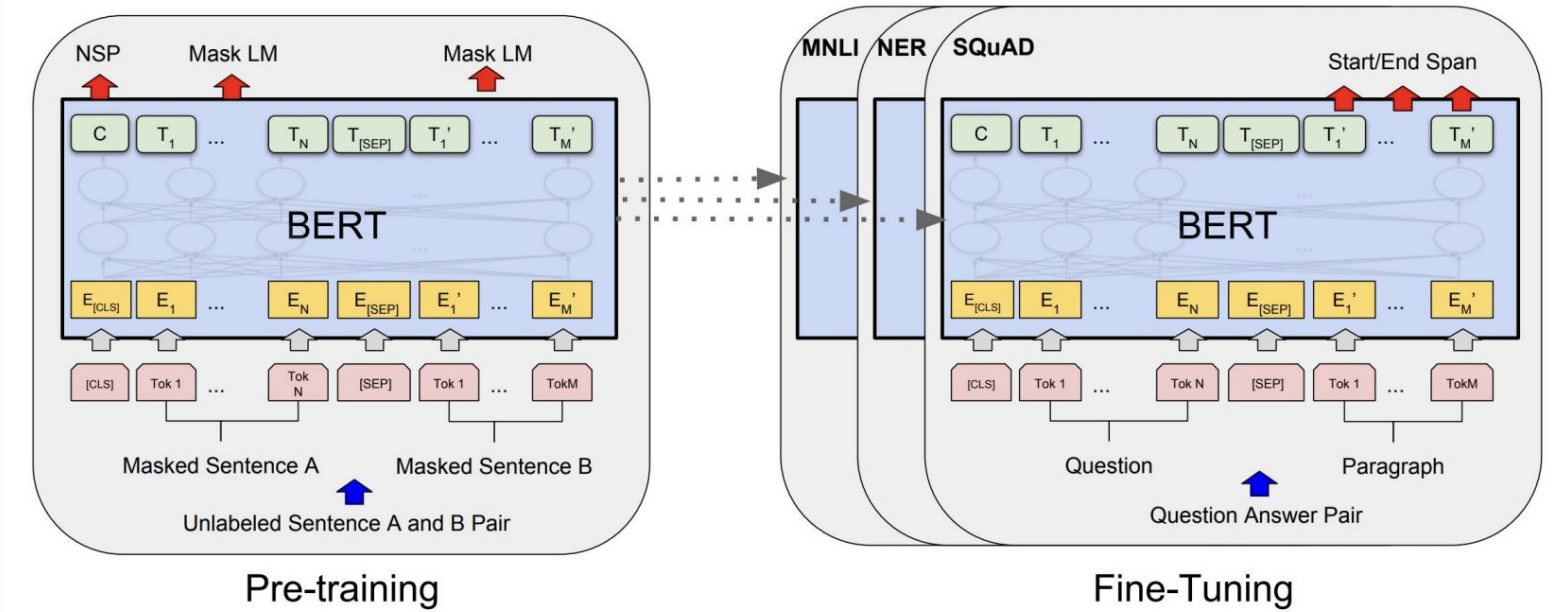
Training Data used (1st model):
1. BooksCorpus (800m words)
2. English Wikipedia (2,500m words)

2 Main Tasks (Pre-training)

1. Masked Language Modelling (MLM) → Filling in blanks in a sentence
2. Next sentence prediction (NSP) → What is the probability that the next sentence follows the current sentence?

BERT - overview (3/3)

2 phases: Pre-training and Fine-tuning



Learn how to understand language

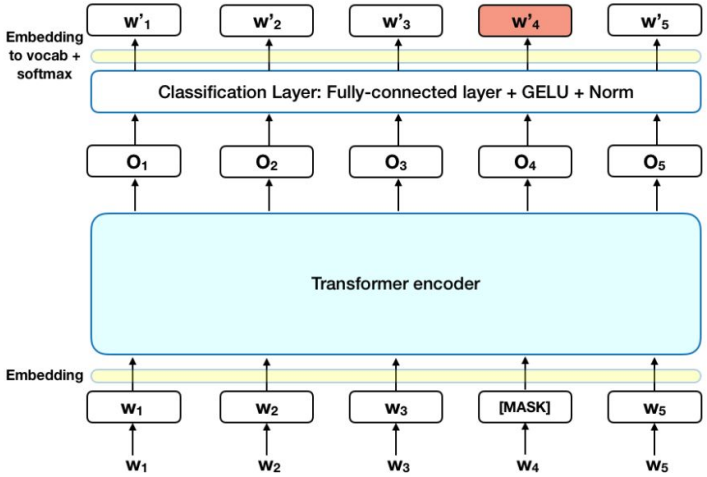
Learn how to do a specific task

BERT - Pre-training phase (1/2)

In order to understand language, masked language is used (fill in the blanks)

Input: The [MASKED WORD 1] brown fox [MASKED WORD 2] over the lazy dog.

Output:
MASKED WORD 1 => quick
MASKED WORD 2 => jumped



They masked 15% of a sentence in this way, the model learns which words could follow another word.

BERT - Pre training phase (2/2)

In order to understand context, Next sentence prediction is used:

Question: Does Sentence A follow Sentence B?

Example 1:

A: The man went to the store

B: He bought a gallon of milk

Label : **IsNextSentence**

Example 2:

A: The man went to the store

B: Penguins are flightless

Label: **NotNextSentence**

BERT -Input Representation

Input: 30,000 words of vocabulary

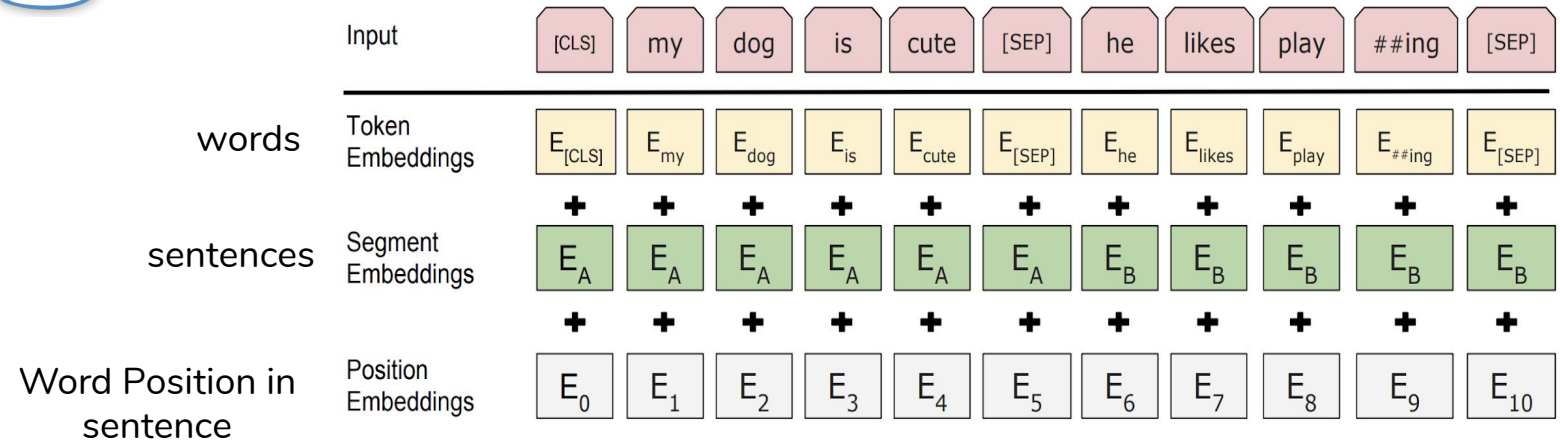


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

BERT - 24 different sizes

There are variants based on size, available at https://github.com/google-research/bert

Released in Mar 2020

You can download all 24 from here, or individually from the table below:

	H=128	H=256	H=512	H=768
L=2	**2/128 (BERT-Tiny)**	2/256	2/512	2/768
L=4	4/128	**4/256 (BERT-Mini)**	**4/512 (BERT-Small)**	4/768
L=6	6/128	6/256	6/512	6/768
L=8	8/128	8/256	**8/512 (BERT-Medium)**	8/768
L=10	10/128	10/256	10/512	10/768
L=12	12/128	12/256	12/512	**12/768 (BERT-Base)**

Note that the BERT-Base model in this release is included for completeness only; it was re-trained under the same regime as the original model.

Here are the corresponding GLUE scores on the test set:

Model	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI(v2)
BERT-Tiny	64.2	0.0	83.2	81.1/71.1	74.3/73.6	62.2/83.4	70.2	70.3	81.5
BERT-Mini	65.8	0.0	85.9	81.1/71.8	75.4/73.3	66.4/86.2	74.8	74.3	84.1
BERT-Small	71.2	27.8	89.7	83.4/76.2	78.8/77.0	68.1/87.0	77.6	77.0	86.4
BERT-Medium	73.5	38.0	89.6	86.6/81.6	80.4/78.4	69.6/87.9	80.0	79.1	87.7

BERT - Fine tuning

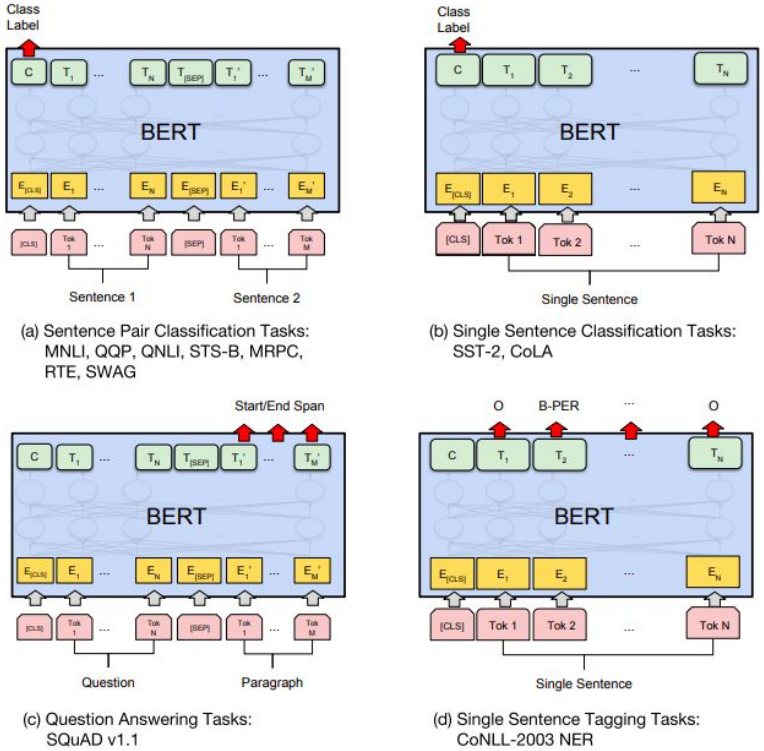
Fine-Tuning Defined:

Take a model that is already trained for a given task, and make it perform a second similar task

How:

We replace the output layer with a specific layer for our chosen task

BERT can be used in different tasks, by adding on layers to suit. (Eg: Classification)



BERT - try out and implement

Try out the notebooks on colab:

BERT End to End (Fine-tuning + Predicting) in 5 minutes with Cloud TPU

Predicting Movie Reviews with BERT on TF Hub Notebook

Sample Process Flow for using BERT in a project:

1. Download a BERT encoder from TensorFlow Hub
2. Load your data
3. Preprocess Data
   - use BERT Tokenizer
   - Encode the sentence
4. Build, Compile your model (eg: add classification layer on top of BERT model)

Variations of BERT

- RoBERTa (Facebook)

  - Much larger model, shows up in the top models for the SuperGLUE benchmarks.

- DistilBERT (HuggingFace)

  - Reduced size by 40%, while retaining 97% of language model

- ALBERT (Google)

  - Much smaller model compared to BERT
  - ALBERT x-large (59m) vs BERT x-large (1.27b) parameters

- ELECTRA (Google)

  - More efficient while matching performance of ROBERTA in the benchmark

Further Info here:
https://www.kdnuggets.com/2019/09/bert-roberta-distilbert-xlnet-one-use.html

BERT Variants and Efficiencies

	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base:** 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

GLUE Benchmarks

General Language Understanding Evaluation - launched in 2018

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI
1	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5
2	ERNIE Team - Baidu	ERNIE		90.4	74.4	97.5	93.5/91.4	93.0/92.6	75.2/90.9	91.4	91.0	96.6
3	Alibaba DAMO NLP	StructBERT		90.3	75.3	97.1	93.9/91.9	93.0/92.5	74.8/91.0	90.9	90.7	96.4
4	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9
5	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2
6	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	91.1	95.8
7	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8
8	Huawei Noah's Ark Lab	NEZHA-Large		89.1	69.9	97.3	93.3/91.0	92.4/91.9	74.2/90.6	91.0	90.7	95.7
9	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6
10	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5
24	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hidden		80.5	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7	85.9	92.7

SuperGLUE benchmark

Launched in Aug 2019

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	2	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
+	3	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
+	4	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84.4/53.4	91.5/91.0	87.9	74.1	91.8	57.6	89.3/75.6
	5	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
+	6	Huawei Noah's Ark Lab	NEZHA-Large		84.8	86.8	94.4/96.0	91.2	82.9/48.8	87.4/86.7	88.5	73.1	90.4	58.0	87.1/74.4
	7	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
+	8	Infosys : DAWN : AI Research	RoBERTa-iCETS		77.4	84.7	88.2/91.6	85.8	78.4/37.5	82.9/82.4	83.8	69.1	65.1	35.2	93.8/68.8
	9	IBM Research AI	BERT-mtl		73.5	84.8	89.6/94.0	73.8	73.2/30.5	74.6/74.0	84.1	66.2	61.0	29.6	97.8/57.3
	10	Ben Mann	GPT-3 few-shot - OpenAI		71.8	76.4	52.0/75.6	92.0	75.4/30.5	91.1/90.2	69.0	49.4	80.1	21.1	90.4/55.3
	11	SuperGLUE Baselines	BERT++		71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	38.0	99.4/51.4
			BERT		69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	23.0	97.8/51.7
			Most Frequent Class		47.1	62.3	21.7/48.4	50.0	61.1/0.3	33.4/32.5	50.3	50.0	65.1	0.0	100.0/50.0
			CBoW		44.5	62.2	49.0/71.2	51.6	0.0/0.5	14.0/13.6	49.7	53.1	65.1	-0.4	100.0/50.0
			Outside Best		-	80.4	-	84.4	70.4/24.5	74.8/73.0	82.7	-	-	-	-
	-	Stanford Hazy Research	Snorkel [SuperGLUE v1.9]		-	-	88.6/93.2	76.2	76.4/36.3	-	78.9	72.1	72.6	47.6	-

Benchmark Tasks

Just a quick comparison of the benchmark tasks.

Only the RTE and Winograd is used for both benchmarks.

GLUE Tasks

Name	Download	More Info	Metric
The Corpus of Linguistic Acceptability			Matthew's Corr
The Stanford Sentiment Treebank			Accuracy
Microsoft Research Paraphrase Corpus			F1 / Accuracy
Semantic Textual Similarity Benchmark			Pearson-Spearman Corr
Quora Question Pairs			F1 / Accuracy
MultiNLI Matched			Accuracy
MultiNLI Mismatched			Accuracy
Question NLI			Accuracy
Recognizing Textual Entailment			Accuracy
Winograd NLI			Accuracy
Diagnostics Main			Matthew's Corr

SuperGLUE Tasks

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy

GLUE Benchmark - tasks

Task Acronym	Task Name	Dataset for task	Expected output	Evaluation criteria
CoLA	Corpus of Linguistic Acceptability	Each example is a sequence of words annotated with whether it is a grammatical English sentence.		Matthews correlation coefficient
SST-2	Stanford Sentiment Treebank	Movie reviews and human annotations of their sentiment. They use only single sentence-level labels	Positive/Negative label for each review	Accuracy.
MRPC	Microsoft Research Paraphrase Corpus	Contains sentence pairs with human annotations for whether the sentences in the pair are semantically equivalent.	Predict if pair of sentences are semantically equivalent.	Accuracy and F1 score.
QQP	Quora Question Pairs	Collection of question pairs.	Predict if question pairs are semantically equivalent.	Accuracy and F1 score.
STS-B	Semantic Textual Similarity Benchmark	Collection of sentence pairs. Each is annotated with a similarity score from 1 to 5.	Predict the similarity scores for each pair.	Pearson and Spearman correlation coefficients.
MNLI (Matched and Mismatched)	Multi-Genre Natural Language Inference Corpus	Collection of sentence pairs with textual entailment annotations. It contains a premise sentence and a hypothesis sentence.	Predict whether premise entails the hypothesis(entailment), contradicts the hypothesis(contradiction), or neither(neutral).	Accuracy.
QNLI	Question-answering NLI, Question paragraph pairs from SQuAD.	Collection of question-paragraph pairs. Converted into a sentence pair classification, each question and sentence in corresponding context. Pairs with low lexical overlap are filtered out. Modifies	Determine whether the context sentence contains answer to the question.	Accuracy.
RTE	Recognizing Textual Entailment	Combines a series of annual textual entailment challenges.	Predict entailment/not_entailment. Datasets having three-class split, neutral and contradiction are collapsed into not_entailment.	Accuracy.
WNLI	Winograd NLI, derived from Winograd Schema Challenge	Reading comprehension task, system must read a sentence with a pronoun and select the referent of that pronoun from a list of choices. It is converted to a sentence pair classification, by replacing the ambiguous pronoun with each possible referent.	Predict if the sentence with the pronoun substituted is entailed by the original sentence.	Accuracy. Each example is evaluated separately.

SuperGLUE Benchmark - tasks

Task Acronym	Task Name	Dataset for task	Expected output	Evaluation criteria
BoolQ	Boolean Questions	A QA task consisting of a short passage and a yes/no question about the same.	Predict the correct answer of the question.	Accuracy.
CB	Commitment Bank	A collection of short texts in which at least one sentence contains an embedded clause. Each example consists of a premise containing an embedded clause and corresponding hypothesis is the extraction of that clause,	Three-class textual entailment.	Accuracy and F1 score.
COPA	Choice of Plausible Alternatives	Causal reasoning task.	A premise sentence is given, system must determine either the cause or the effect of the premise from two possible choices.	Accuracy.
MultiRC	Multi-Sentence Reading Comprehension	A QA task, each example consisting of a context paragraph, a question about that paragraph, and a list of possible answers.	Predict which answers are true and which are false. Each question can have multiple possible correct answers.	F1 over all answer-options (F1a) and exact match of each question's set of answers (EM).
ReCoRD	Reading Comprehension with Commonsense Reasoning Dataset	Multiple-choice QA task. Each example consists of a news article and a Cloze-style question about the article in which one entity is masked out.	Predict the masked out entity from a given list of possible entities in the provided passage, where the same entity may be expressed using multiple different surface forms, all of which are considered correct.	Max (over all mentions) token-level F1 and exact match (EM).
RTE	Recognizing Textual Entailment	Same as GLUE.		Accuracy.
WiC	Word-in-Context	Word sense disambiguation task cast as binary classification of sentence pairs. Given two text snippets and a polysemous word that appears in both sentences.	Determine whether the word is used with the same sense in both sentences.	Accuracy.
WSC	Winograd Schema Challenge	Almost same as GLUE.		Accuracy.

Recognizing Textual Entailment (RTE)

Does sentence 1 allow sentence 2 to be inferred?

Premise:
Claims by a French newspaper that seventime Tour de France winner Lance Armstrong had taken EPO were attacked as unsound and unethical by the director of the Canadian laboratory whose tests saw Olympic drug cheat Ben Johnson hit with a lifetime ban.

Hypothesis:
Lance Armstrong is a Tour de France winner.

Answer: Yes

Winograd Schema Challenge

Test of Reading Comprehension:

Example 1:

The trophy doesn't fit in the brown suitcase because it's too big. What is too big?

Answer 0: the trophy

Answer 1: the suitcase

Example 2:

Joan made sure to thank Susan for all the help she had given. Who had given the help?

Answer 0: Joan

Answer 1: Susan

Conclusion

Topics Covered:

- Transformers
- BERT and its Variants
- GLUE Benchmarks

Resources

The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)

The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - Slides

The Annotated Transformer - Notebook with pytorch implementation

https://towardsml.com/2019/09/17/bert-explained-a-complete-guide-with-theory-and-tutorial/

Medium Articles:

What is a Transformer?

2019 — Year of BERT and Transformer

Research Papers

Attention Is All You Need

Deep contextualized word representations

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Videos:

Attention

Appendix

More detailed examples of the content.

Transformers - Positional Encoding

Here's the formula they use to calculate the positional encoding:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

In this equation,

- pos is the position of a word in the sentence (e.g. "2" for the second word in the sentence)
- i indexes into the embedding dimension, i.e. it's the position along the positional encoding vector dimension. For a positional encoding vector of length d_model = 512, we'll have i range from 1 to 512.

Why use sine and cosine? To quote the authors, "each dimension of the positional encoding corresponds to a sinusoid. [...] We chose this function because we hypothesized it would allow the model to easily learn to attend by relative positions."

https://glassboxmedicine.com/2019/08/15/the-transformer-attention-is-all-you-need/

BERT - Example of MLM

Masked Language Modelling (MLM)

BERT uses a simple approach for this: We mask out 15% of the words in the input, run the entire sequence through a deep bidirectional Transformer encoder, and then predict only the masked words. For example:

```
Input: the man went to the [MASK1] . he bought a [MASK2] of milk.
Labels: [MASK1] = store; [MASK2] = gallon
```

In order to learn relationships between sentences, we also train on a simple task which can be generated from any monolingual corpus: Given two sentences `A` and `B`, is `B` the actual next sentence that comes after `A`, or just a random sentence

```
Sentence A: the man went to the store .
Sentence B: he bought a gallon of milk .
Label: IsNextSentence
```

```
Sentence A: the man went to the store .
Sentence B: penguins are flightless .
Label: NotNextSentence
```

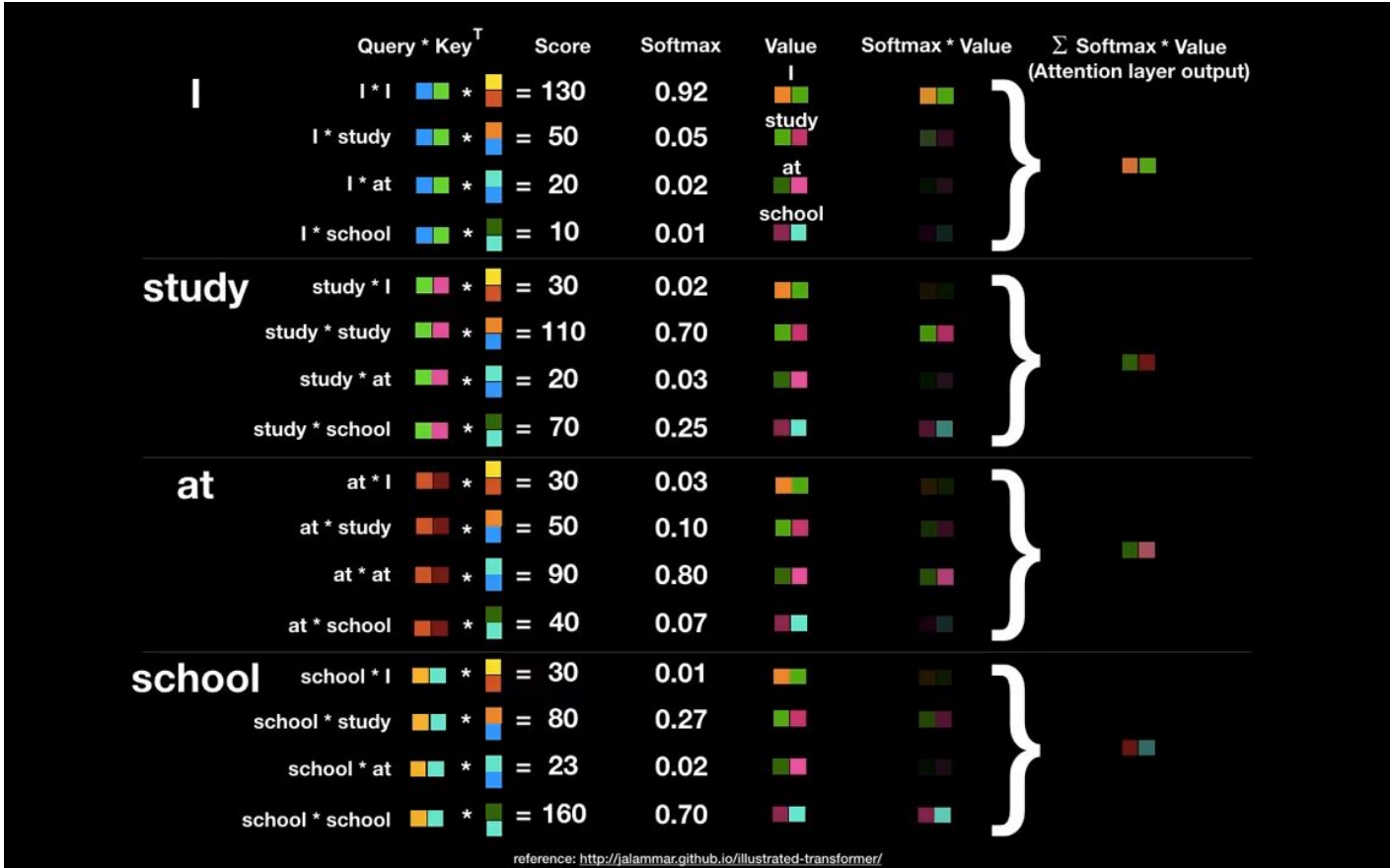
GLUE Benchmark Examples

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

Attention

Attention Layer Output:

Sum of (Softmax * Value)



From youtube video: https://www.youtube.com/watch?v=z1xs9jdZnuY