Semantic Similarity Search

Josh Hedtke and Sergei Petrov
{jhedtke, spetrov}@stanford.edu

Stanford

Motivation

- How do we search for similar, potentially lengthy, phrases in a database of N=**100b+ phrases**?
- We used transfer learning to fine tune the **BERT** model [4] using a fully connected single layer trained on labeled paraphrase pairs from to get fixed length representations of sentences
- These fixed length vectors are then used in approximate nearest neighbor searches based on **cosine similarity**

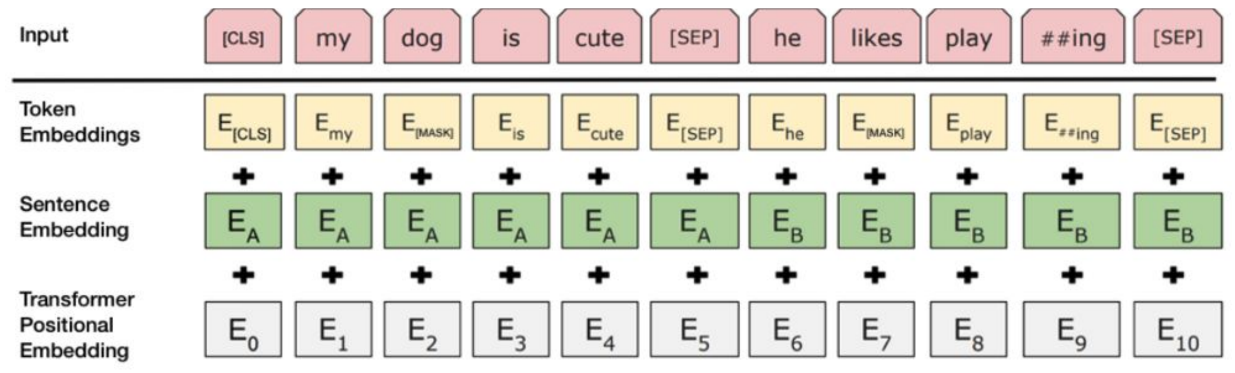
Datasets

- The validation dataset is 89 pairs of sentences with similarity 1:
  - they said before i had a visa, they told me I had a visa earlier
- Another dataset was constructed pairing the first sentences from the original validation dataset with unrelated sentences a set of pairs with similarity 0:
  - they said before i had a visa , In this section we demonstrate the workflow by training and testing a model for salt body interpretation
- Two finetuned models were trained on public datasets MRPC and STS-B. Both datasets have sentence pairs and similarity labels

  MRPC:
  - 0 Rudder was most recently senior vice president for the Developer & Platform Evangelism Business. Senior Vice President Eric Rudder, formerly head of the Developer and Platform Evangelism unit, will lead the new entity.

  STS-B (scores were converted into binary labels):
  - 3.800 A man is playing a large flute. A man is playing a flute.

Features

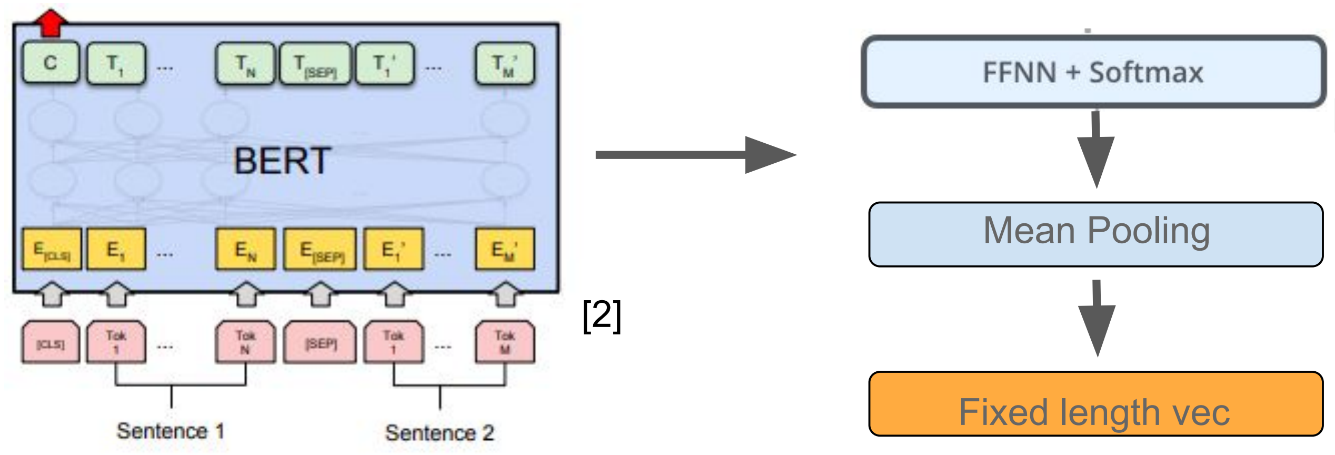


[1]

- BERT tokenizes the input sequence adding a [CLS] token for classification tasks and [SEP] tokens to mark the end of each sentence
- Each token is encoded with its corresponding embedding vector and positional encoding vectors are added
- Sentence embeddings are added to differentiate between the sentences if input comes in pairs
- Attention mask is added to define which tokens are real and which are padding tokens

Model Architecture

Fine-tuned BERT



[2] [2]

Models:

- Baseline: Averaged word2vec
- Bert base uncased (all lower case) [3]
- Bert base uncased finetuned on MRPC dataset
- Bert base uncased finetuned on STS-B dataset

Model	Parameters
Baseline	300 dim word2vec
Bert base uncased	12 layers, 768 hidden units, 12 attention heads, 110m total parameters
Finetuned on MRPC	Seq length = 256 Batch size = 32 Learning rate = 2e-5 N epochs = 3 Adam Kept 11th layer
Finetuned on STS-B	Seq length = 128 Batch size = 32 Learning rate = 3e-5 N epochs = 3 Adam Kept 11th layer

Results

Finetuned BERT Training Results

Model	Train Accuracy	Validation Accuracy
Finetuned on MRPC		0.858
Finetuned STS-B	0.817	0.836

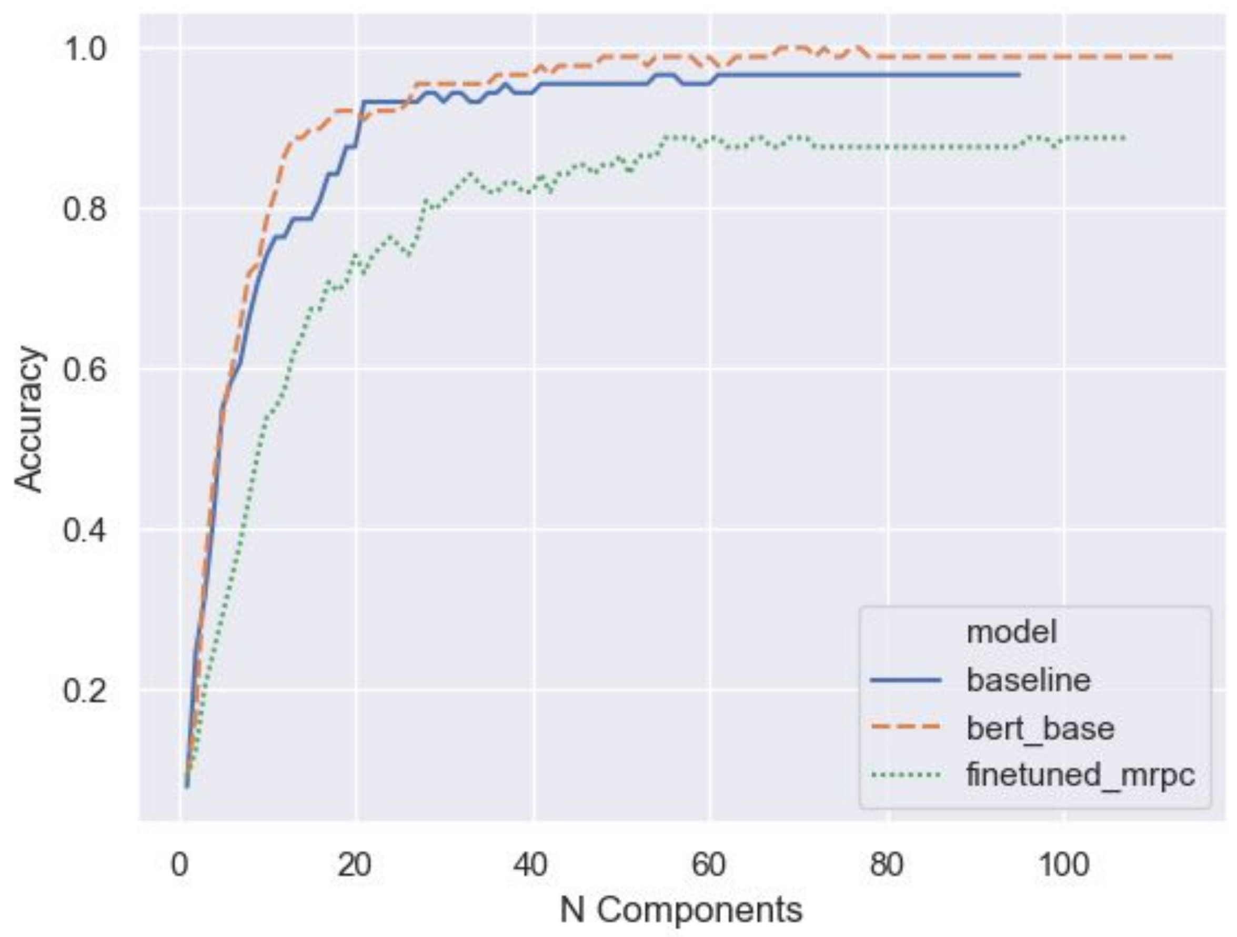
Search Results

Evaluation Metrics

- Top 1: top ranked match by cos similarity is gold match
- Top 5: top 5 ranked matches by cos similarity include gold match

Model	Top 1 Accuracy	Top 5 Accuracy
Baseline	0.831	0.876
Bert base	0.843	0.921
Finetuned on MRPC	0.640	0.809
Finetuned on STS-B	0.719	0.786

Dimension Reduction with PCA



Stopping Criterion:

$$\Delta explainedvariance < \epsilon, \epsilon = 0.001$$

Model	N components	Top 5 Accuracy	Explained Variance
Baseline	95	0.966	0.974
Bert base	113	0.988	0.964
Finetuned on MRPC	107	0.888	0.967

K-means clustering

- The dataset with 0 similarity pairs was used to perform division into 2 clusters. Separation of the data into two semantically different groups was successful only with embeddings of the baseline and bert base algorithms
- The dataset with all pairs similar was separated into 4 clusters. Clusters obtained with bert base embeddings:

1	I can definitely help you and just to make sure i've got the right account what is your full name; What is your income either per year or every month or bimonthly whatever way is easiest; They say they won't take an email but i'm just gonna forward it to them will you send that to me; Hey i need to make sure my insurances are listed on the new address i just changed usaa just notified me to call in
2	Batching doesn't start until midnight so it might not be done processing yet; We could transfer money into my usaa checking account when i need it; I can't hear you let me turn the volume up; Will i see the automatic payments plugged in already when i go online
3	They told me I had a visa earlier; I did it online before; Last time i did it online; How may I help you
4	Um and then once you get it uh let me know if you want to open it and let me know that it looks good for ya uhm so that way you don't have to call back; For my car loan tried to go through today uhm there wasn't enough money in in the account so i'm currently showing negative uhm in the account and the transaction is still pending; [laughter] not a problem sir it happens to me all the time so i know how it goes if you need any help from us sir feel free to give us a call back okay; I like talking to people like you who aren't with usaa so i can explain what'll happen 'cause some people don't get it as readily as you did

Discussion

- The finetuned semantic similarity BERT models surprisingly performed worse on the semantic search task than both the baseline and bert base.
- This may be because the training sets were tight paraphrases whereas the validation set was composed of very loose paraphrases.
- Reduction from 768 to ~100 dim improved performance for the baseline, bert base, and the finetuned on MRPC models, suggesting overfitting.
- Interestingly, performance flattens quickly around 20 principal components.

Future

- Train on a large dataset of the same type of paraphrases as the validation set.
- Finetuning on different layers of the bert base model, which may capture different semantic and structural meaning.

References

[1] Francisco Ingham. Understanding BERT Part 2: BERT Specifics https://medium.com/dissecting-bert/dissecting-bert-part2-335ff2ed9c73

[2] Jay Alammar. The Illustrated BERT, ELMo, and co. https://jalammar.github.io/illustrated-bert/

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.