ConTra: A Contrastive Approach to Text Classification using Transformers

Faisal Amin, Krishnavyas Muddineni, Ian Paolo Tagorda

National University of Singapore, School of Computing

Abstract

We developed a self-supervised model for text classification using contrastive learning. Our transformer model called ConTra achieves competitive results against state of the art masked language model DistilBERT while using considerably fewer parameters.

Dataset

-> Distilled Version of AG, which is a collection of over 1 million news articles gathered from other 2000 sources
-> Contains 4 classes: World, Sports, Business and Sci/Tech
-> 40,000 training samples and 1900 testing samples (all classes equally distributed)
-> Is a benchmark dataset with lots of diversity, allowing for a more robust and generalized model solution

Overview of Data Augmentations

-> Explored 7 pre-vectorization data augmentations for text data, example augmentations below are generated on the following sentence: '**The quick brown fox jumps over the lazy dog**'

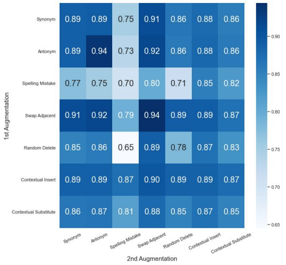
	The **agile** brown fox jumps **terminated** the lazy dog		The quick brown _ jumps over _ _ dog
Synonym	The **prompt** brown **dodger** jumps over the lazy **wiener**	Random Delete	_ Quick _ fox jumps over the lazy _
(Antonym behaves similarly)	The **fast** brown fox **parachute ended** the **otiose** dog		The quick _ fox _ over _ lazy dog
	Tha quikly brown fox jumps over the lazy dog		The **very** quick brown **bengal** fox jumps over the lazy **short** dog
Spelling Mistake	The **quick** brown fox jumps **overt** the **laizy** dog	Contextual Insert	**Even** the quick **young** brown fox jumps over the **eager** lazy dog
	The quick **brow foq** jumps over the lazy **dadg**	(using DistilBERT)	The **surprisingly** quick brown **eyed** fox jumps **up** over the dog
	The quick brown fox jumps **lazy over dog the**		The **dog tracking** fox jumps over the **wild** dog
Swap Adjacent	The **fox quick brown** jumps **the over** lazy dog	Contextual Substitute	**My mischievous red** fox jumps over the lazy dog
	The **brown quick jumps fox** over the lazy dog	(using DistilBERT)	The quick **young man runs to** the lazy dog

Data Augmentation Similarity Study

-> To study the effectiveness of each augmentation, we tried single and multi-chain augmentations
-> Original and Augmented samples are converted to word vectors with GloVe embeddings, then we find the cosine similarity between original and augmented vectors

Figure showcasing difference between original and augmented data after one augmentation

Figure showcasing difference between original and augmented data after 2 successive augmentations



-> Spelling Mistake augmentation causes too large semantic changes, so it was dropped

-> All other 6 augmentations cause a large enough semantic change for the model to find useful distinguishable features between pairs of similar original and augmented data

-> To increase robustness and generalization capabilities of model, every augmentation made is a random one among all remaining 6 ones

-> We tested fully randomized 2-chain, 3-chain, 4-chain, 5-chain and 6-chain augmentations and found fully randomized 4-chain augmentations to be optimal

Contrastive loss

In contrastive learning, a model is trained to identify whether two data points are similar or dissimilar. The model takes two types of inputs - a positive example (a data point that is similar to the input data point) and a negative example (a data point that is dissimilar to the input data point). It is then trained to learn a representation that maps similar data points close together in the representation space while pushing dissimilar data points far apart.

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

We used cosine similarity as the similarity metric between our samples to calculate the loss. The idea is that by minimizing the contrastive loss, the model learns a representation space where similar examples are close together and dissimilar examples are far apart, making it easier to perform similarity comparisons or clustering tasks.

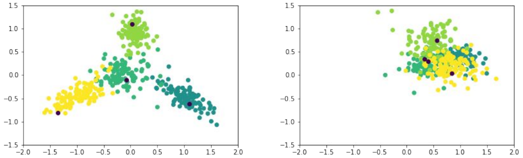
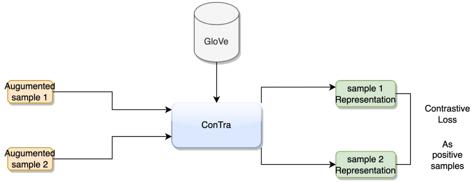


Figure: PCA visualization of sentence embeddings before and after training with contrastive loss.

We can see that after the model transforms the GloVe embeddings, samples of the same class are closer together.

Contrastive Transformer (ConTra)

We have defined a custom transformer(ConTra) model with 6 layers, 12 heads and hidden size of 2048. We utilized GloVe vectors as our embeddings for our data. During the training, we randomly select 2 different augmentations of the same data point as a positive pair. We extracted representations from ConTra model and calculated the contrastive loss with that pair as a positive sample.



t-SNE Visualization

To visualize the data, we first got the the sentence embeddings of each sentence by calculating the mean of the word embeddings. t-SNE was then use visualize the distribution of the sentence embeddings.

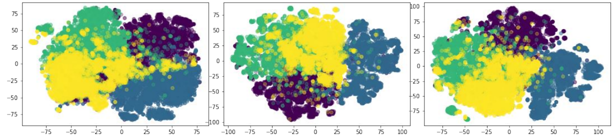


Figure: t-SNE visualizations of each model. (a) shows the GloVe embeddings, (b) shows ConTra embeddings and (c) shows DistilBERT embeddings.

Results

Table: Performance Comparison of different models without fine-tuning

Model	KNN
Glove	80.1842
ConTra	84.5263
DistilBERT	**87.9737**

Table: Performance Comparison of different models after fine-tuning

Model	Number of Parameters	Accuracy
Glove	**1204**	87.6315
ConTra	9562492	89.6973
DistilBERT	66365956	**90.6710**

Discussion and Future Work

Even with a smaller number of parameters, ConTra can achieve performance comparable performance to DistilBERT. It is also good to note that this performance was achieved by ConTra by using only the provided dataset which is extremely smaller compared to that dataset used to train DistilBERT. The next step of the team is to investigate whether DistilBERT's performance will not decrease even if it were trained from scratch like ConTra.

Contact Information

Faisal Amin - faisal.amin@u.nus.edu
Krishnavyas Muddineni - krish-mu@comp.nus.edu.sg
Ian Paolo Tagorda - itagorda@comp.nus.edu.sg