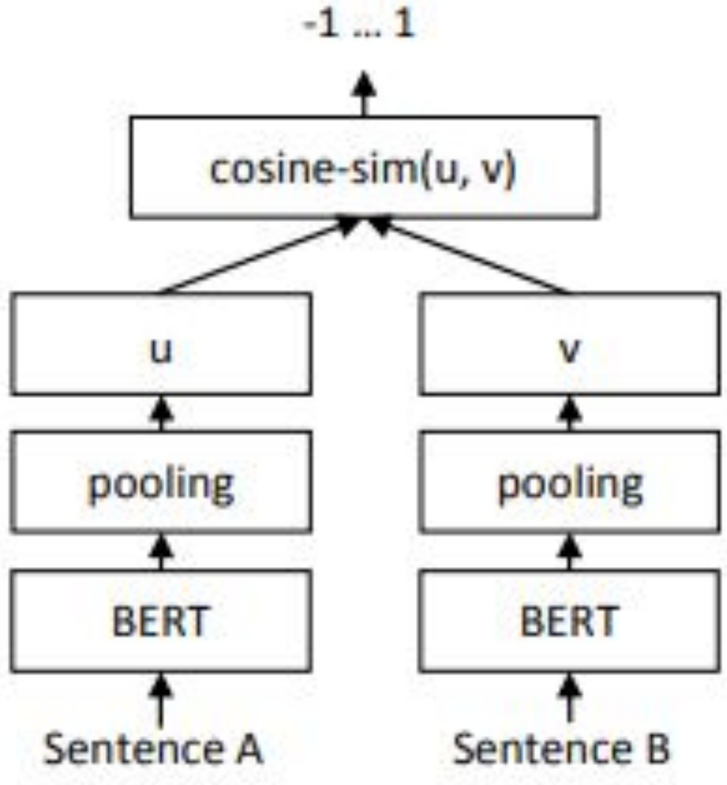
Hyperboloid embeddings and Hypersphere similarity metrics for Sentence-BERT

Garvit Goyal, Gunjan Gupta, Tusheet Goli

Sentence-BERT (SBERT)

- SBERT is a modified BERT network
- Better for deriving semantically meaningful sentence embeddings
- State-of-art sentence embedding model that is computationally efficient to get sentence embeddings
- Uses siamese and triplet network structures
- Uses distance as similarity metrics and contrastive loss

Sentence-BERT (SBERT) Architecture



Our Goal

- Analyze and understand SBERT
- Aim to improve SBERT
    - Changing the pooling layer to go from BERT word embeddings to sentence embeddings
    - Exploring other similarity metrics that might be more suited for the semantic similarity task

Reference Sentence: That is a happy person.

Sentence	Similarity Score
That is a very happy person	0.943
That is a happy dog	0.665
That is a happy girl	0.685

Table 1: Example of semantic similarity between sentences from SBERT model

Datasets

STSb
(Semantic Textual Similarity Benchmark)

- A selection of English datasets that have been primarily used as a benchmark dataset for SemEval between 2012 and 2017
- Includes text from image captions, news headlines, and user forum
- Consists of 8628 sentence pairs from multiple genres
  - Divided into a train, dev, and test splits

sentence1	sentence2	similarity score
A plane is taking off.	An air plane is taking off.	5
Three men are playing chess.	Two men are playing chess.	2.6
A man is smoking.	A man is skating.	0.5
A man pouts oil into a pot.	A man pours wine into a pot.	3.2
A man is playing a guitar.	A girl is playing a guitar.	2.8

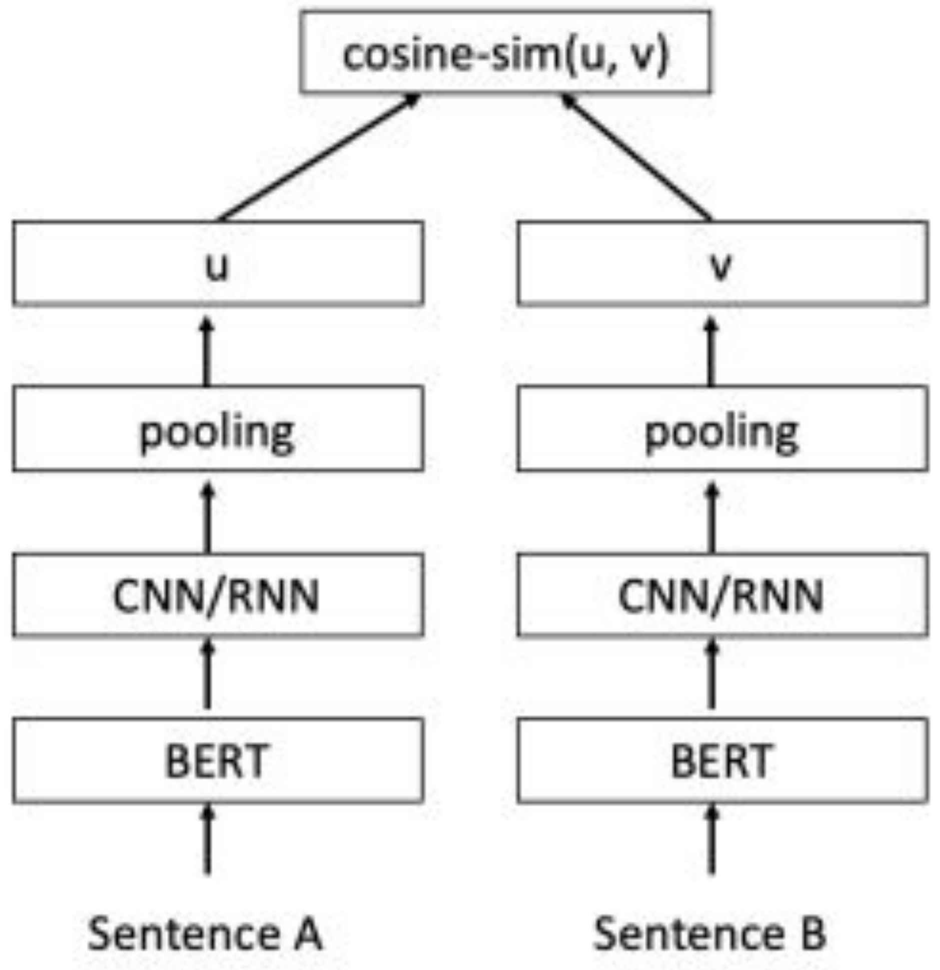
SNLI
(Stanford Natural Language Inference)

- Contains 570,000 sentence pairs labeled as entailment, neutral, or contradiction
- Binary labeled dataset of sentence pairs
- Used for testing our uniformity and alignment as the similarity metrics in our loss function

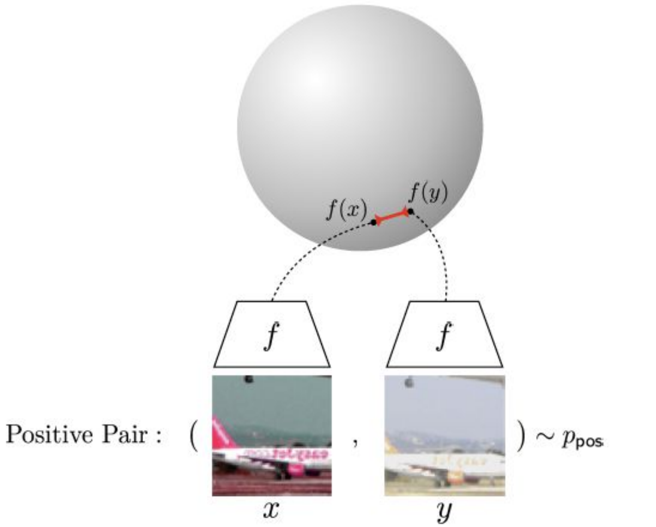
Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

Model

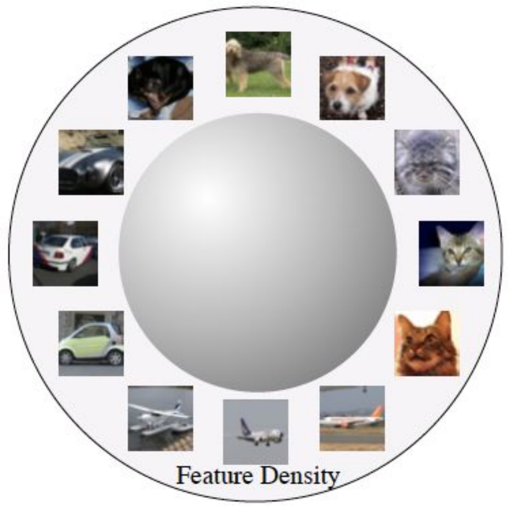
Proposed Architecture



Hyperspherical loss functions



Alignment: Similar samples have similar features.



Uniformity: Preserve maximal information.

Source: Wang and Isola (2020)

$$\mathcal{L}_{align}(f;\alpha) \triangleq \mathbb{E}_{(x,y)\sim p_{pos}}[\|f(x) - f(y)\|_2^\alpha]$$

$$\mathcal{L}_{uniform}(f;t) \triangleq \mathbb{E}_{(x,y) \overset{\text{i.i.d.}}{\sim} p_{data}}[\exp(-t\|f(x) - f(y)\|_2^2)]$$

Hyperboloid Embeddings

Lorentzian distance: $d^2_{\mathcal{L}}(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|^2_{\mathcal{L}} = -2\beta - 2\langle \mathbf{a}, \mathbf{b} \rangle_{\mathcal{L}}$

Mapping: $g_\beta : \mathbb{R}^d \to \mathcal{H}^{d,\beta} \; \forall \, \mathbf{x} = (x_1, ..., x_d) \in \mathbb{R}^d$

$$g_\beta(\mathbf{x}) := (\sqrt{\|\mathbf{x}\|^2 + \beta}, x_1, ..., x_d) \in \mathcal{H}^{d,\beta}$$

Implementation Details

Implementation Details

- Baseline: bert-base-uncased followed by mean pooling
- Optimizer: SGD with lr = 0.001 and momentum = 0.9
- The NLI versions were trained on NLI for one epoch, and the STS versions were trained on the STS dataset for 10 epochs.
- Both types of models were validated every 10% of iterations on dev split of STS benchmark, and tested on test split of STS benchmark

Implementation Details

$$\mathcal{L}_{align}(f;\alpha) \triangleq \mathbb{E}_{(x,y)\sim p_{pos}}[\|f(x)-f(y)\|_2^\alpha]$$

$$\mathcal{L}_{uniform}(f;t) \triangleq \mathbb{E}_{(x,y) \overset{\text{i.i.d.}}{\sim} p_{data}}[\exp(-t\|f(x)-f(y)\|_2^2)]$$

$$\mathcal{L}_{total} = w_a\mathcal{L}_{align} + w_u\mathcal{L}_{uniform}$$

$$w_a = 5, w_u = 1, t = 10^{-10}, \alpha = 2$$

$$g_\beta(\mathbf{x}) \coloneqq (\sqrt{\|\mathbf{x}\|^2+\beta}, x_1, ..., x_d) \in \mathcal{H}^{d,\beta}, \beta = 0.1$$

Results

Model	Distance Function	Loss Function	STS Benchmark Test Performance (Spearman coefficient)
BERT-STSB-Baseline	Cosine	Contrastive	0.839
BERT-STSB-CNN + Mean Pooling	Cosine	Contrastive	0.840
BERT-STSB-CNN + Max Pooling	Cosine	Contrastive	0.824
BERT-STSB-hyperboloid	Lorentz	Contrastive	0.591
BERT-NLI-hypersphere	Euclidean	Uniformity - Alignment	0.547
BERT-NLI-hypersphere + hyperboloid	Lorentz	Uniformity - Alignment	0.536

Future Work

- Hyperparameter optimization: more so than usual machine learning models, SBERT's performance is greatly influenced by hyperparameters of the loss function formulation, such as the relative weights of the uniformity and alignment losses and their parameters. However, over the course of experiments, we observed that hyperparameter search is computationally expensive and could be explored further in future works.

- One specific analysis could be to observe the model's performance when optimizing combinations of alignment loss, uniformity loss, and contrastive loss functions. Our work focused on comparing a few combinations of the first two functions against the sole optimization of the third.

- Further, we observed numerical instability when using the logarithm of the expected value of gaussian potential, as proposed by Wang & Isola (2020). Instead, we optimized the expected value of gaussian potential in this work. Future works could try to provide concrete mathematical analyses of this behavior in BERT embeddings since the original work was only tested on CNN and RNN encodings.

References

- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: *Sentence embeddings using siamese bert-networks*.
- Tongzhou Wang and Phillip Isola. 2020. *Understanding contrastive representation learning through alignment and uniformity on the hypersphere*.
- Qian Chen, Zhen-Hua Ling, and Xiaodan Zhu. 2018. *Enhancing sentence embedding with generalized pooling*.
- Maximilian Nickel and Douwe Kiela. 2017a. *Poincaré embeddings for learning hierarchical representations*.

Thank You!