BERT vs DistilBERT

How well do they generalize from video game reviews?

Inspiration

- DistilBERT has 97% of BERTs accuracy
- DistilBERT is 40% smaller than BERT
- DistilBERT is 60% faster than BERT
- What price does it pay?
- Main hypothesis: DistilBERT generalizes worse than BERT

Performance: DistilBERT, a distilled version of BERT, Sanh et Al

Method

- Fine-tune on balanced binary Steam dataset
  - 300,000 samples
  - Standard Huggingface hyperparameters
- Test on several other balanced binary datasets
  - ~50,000 samples
- Test on non-binary dataset
  - 50,000 samples
- Visualize embeddings
- Visualize attention

Balance: Adapt or Get Left Behind, Rietzler et Al

Binary 👍 👎 test results

Dataset	Steam	IMDb	SST-2
BERT accuracy	0.861	0.867	0.791
DistilBERT accuracy	0.881	0.877	0.797
Difference	-0.020	-0.010	-0.006

Multi-class on Yelp dataset

Idea:

Measure generalization ability
via extension to a multi-class task

Using non discretized probabilities: Effect of Using Regression on Class Confidence Scores in Sentiment Analysis of Twitter Data, Onal et al.

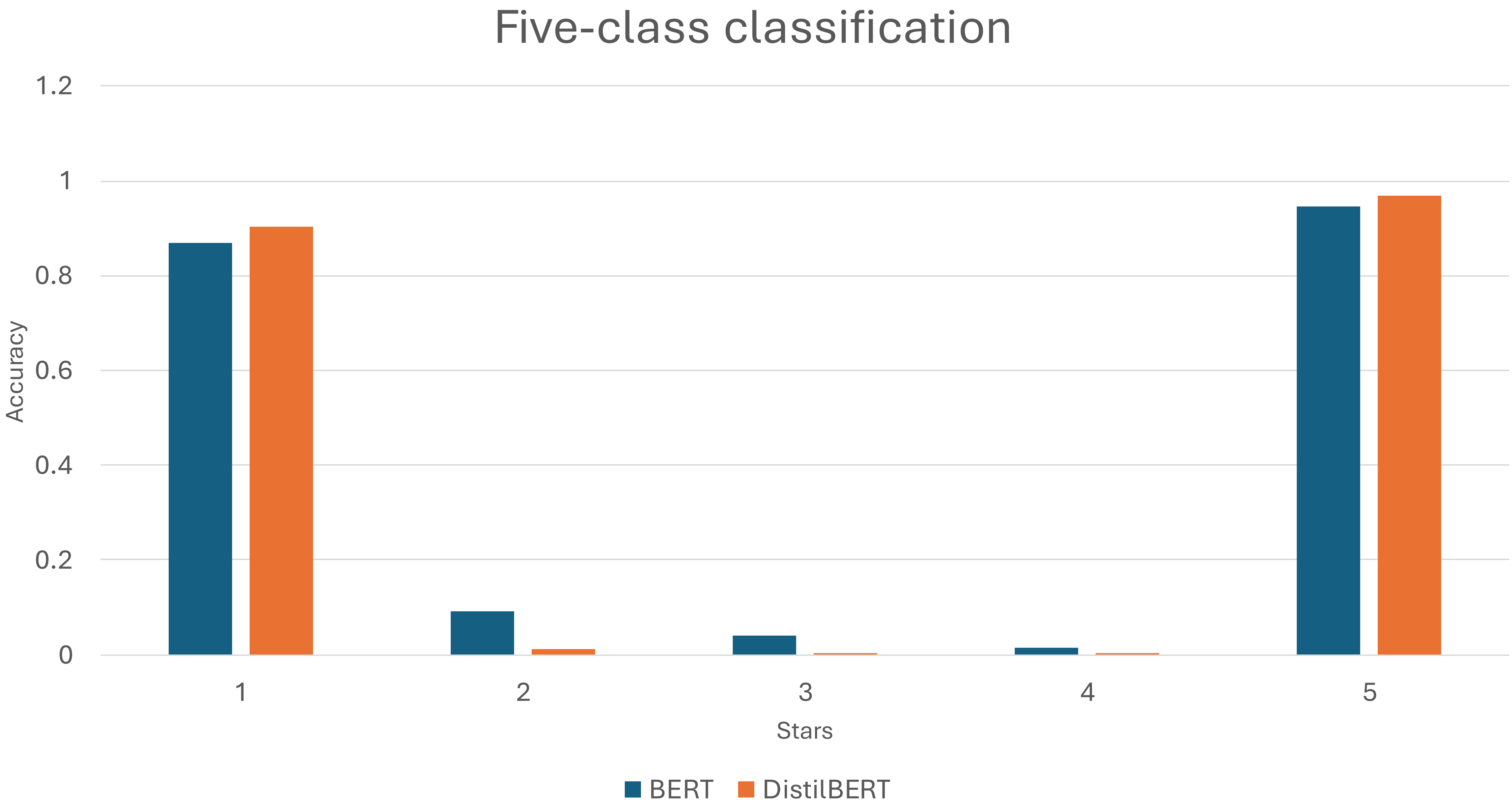
Binary accuracy:

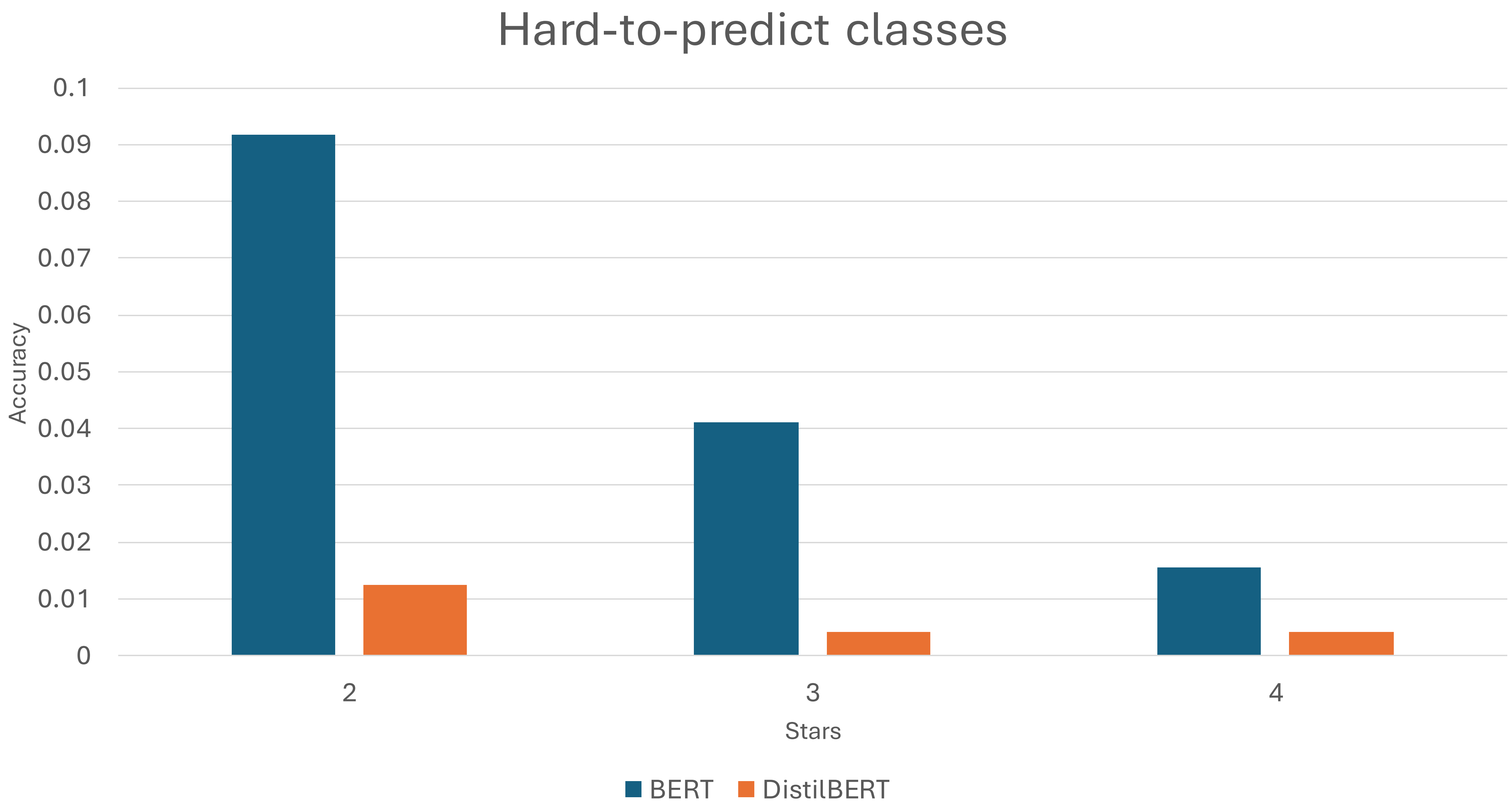
BERT: 0.9082
DistilBERT: 0.9035

Five-class accuracy:

BERT: 0.3927
DistilBERT: 0.3790

Five-class to binary: Sentiment Analysis on Large Scale Amazon Product Reviews, Haque et al.


Five-class classification
1.2
1
0.8
0.6
0.4
0.2
0
Accuracy
1
2
3
4
5
Stars
■ BERT
■ DistilBERT


Hard-to-predict classes
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Accuracy
2
3
4
Stars
BERT
DistilBERT

Multi-class on Yelp dataset

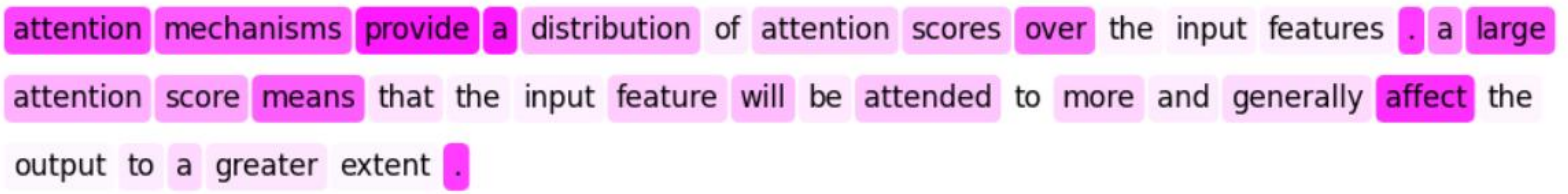
Correct 3-star prediction for both models :

"... It's cheap, delicious & comes with a lot of food...
...However, I ate here for dinner the other night and the wait was ridiculously long!...
... Lots of food & it was good~ love their short ribs!! It's good, cheap food but you just have to wait a looooong time!"

BERT is slightly more adept at understanding nuance, which arguably entails better understanding/generalization

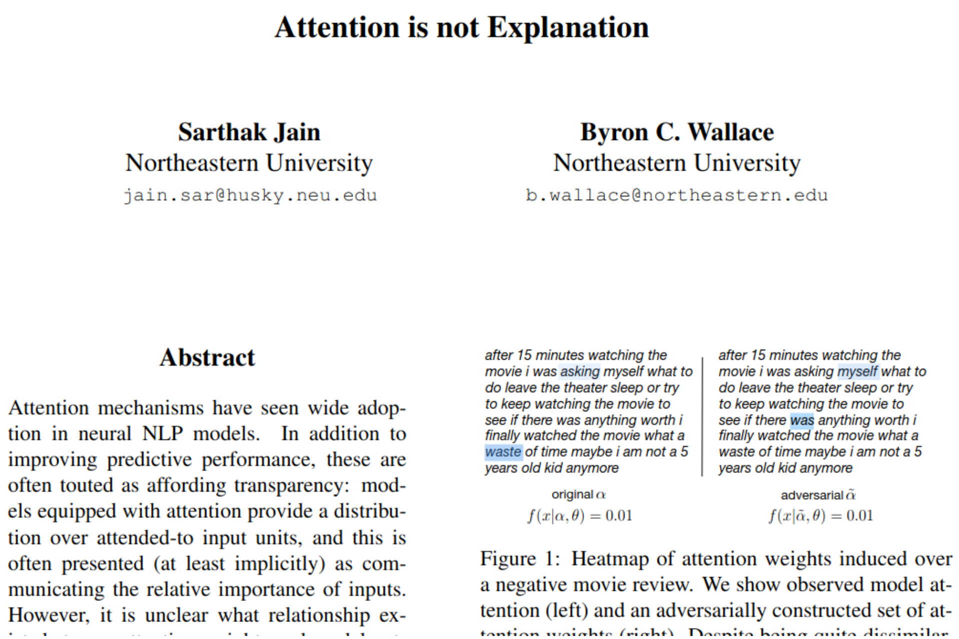
Attention

- Short introduction/refresher: what is attention in NLP?
  - specifically for sequence classification



attention mechanisms provide a distribution of attention scores over the input features . a large attention score means that the input feature will be attended to more and generally affect the output to a greater extent .

- For visualizing the attentions given by a BERT model, the attentions from the 12 attention heads and several layers are averaged



Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Abstract

Attention mechanisms have seen wide adoption in neural NLP models. In addition to improving predictive performance, these are often touted as affording transparency: models equipped with attention provide a distribution over attended-to input units, and this is often presented (at least implicitly) as communicating the relative importance of inputs. However, it is unclear what relationship ex-

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

original α

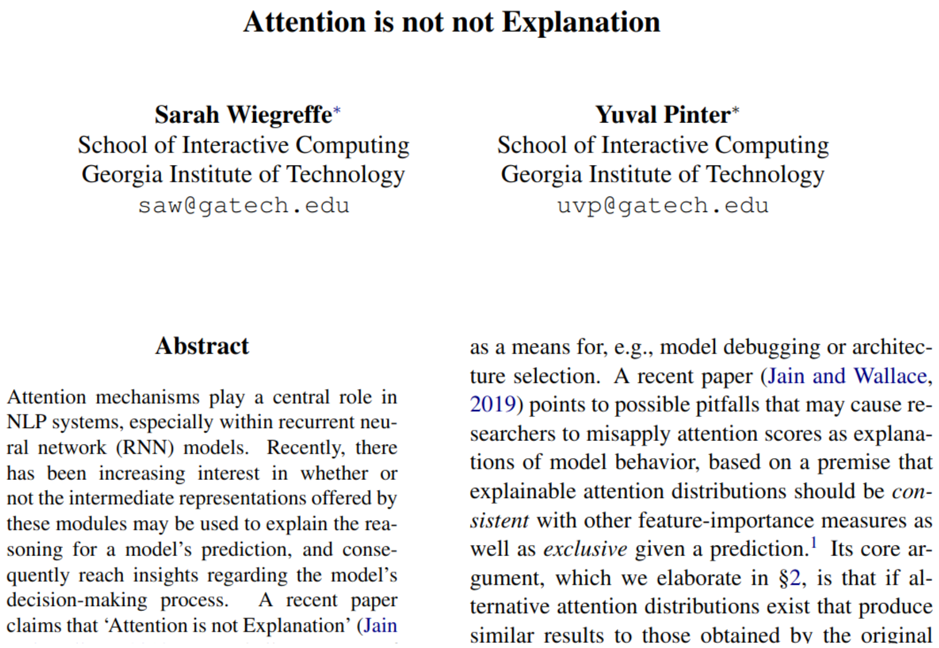
$f(x|\alpha, \theta) = 0.01$

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

adversarial $\tilde{\alpha}$

$f(x|\tilde{\alpha}, \theta) = 0.01$

Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model attention (left) and an adversarially constructed set of attention weights (right). Despite being quite dissimilar



Attention is not not Explanation

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

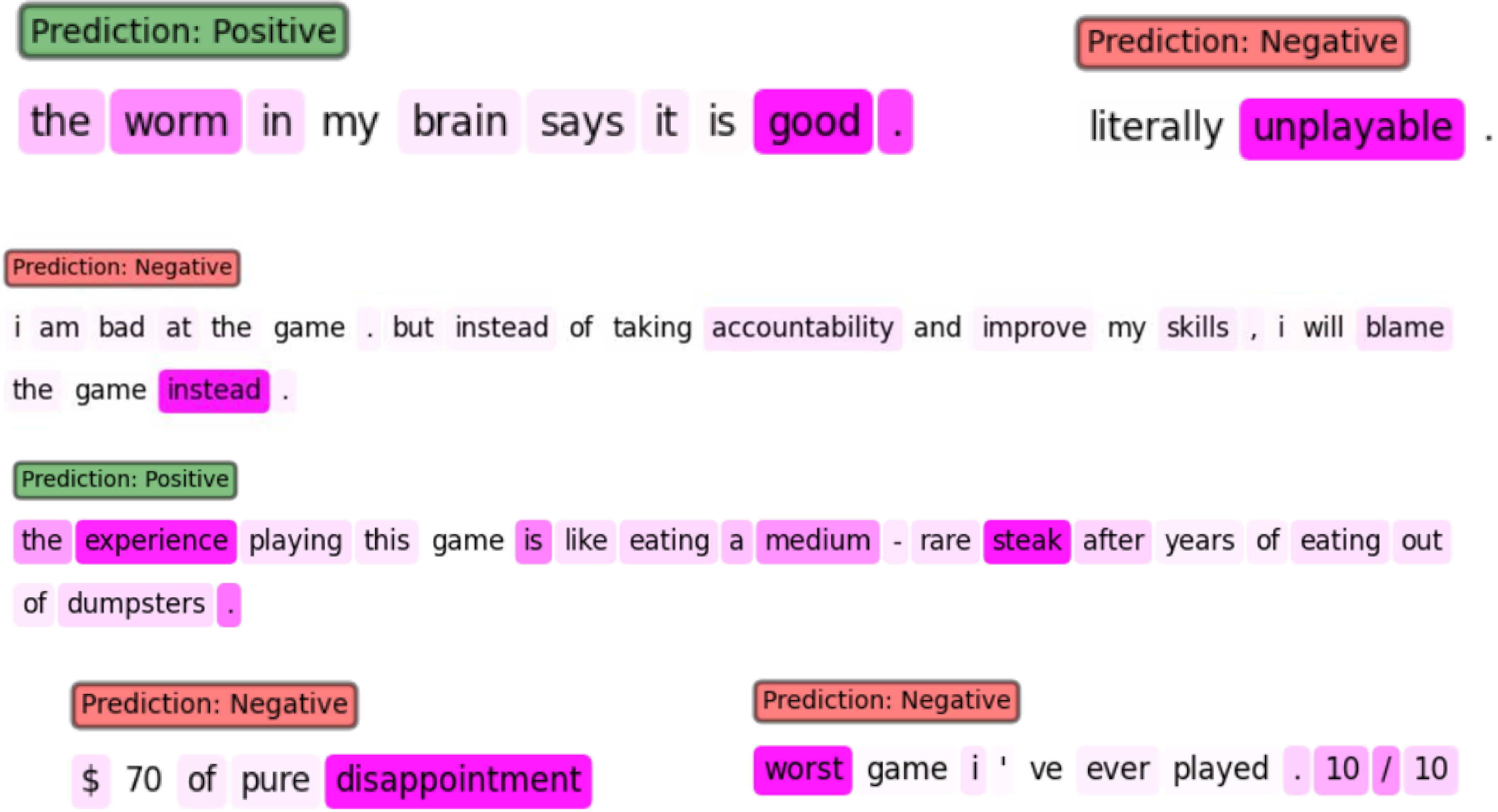
Abstract

Attention mechanisms play a central role in NLP systems, especially within recurrent neural network (RNN) models. Recently, there has been increasing interest in whether or not the intermediate representations offered by these modules may be used to explain the reasoning for a model's prediction, and consequently reach insights regarding the model's decision-making process. A recent paper claims that 'Attention is not Explanation' (Jain

as a means for, e.g., model debugging or architecture selection. A recent paper (Jain and Wallace, 2019) points to possible pitfalls that may cause researchers to misapply attention scores as explanations of model behavior, based on a premise that explainable attention distributions should be *consistent* with other feature-importance measures as well as *exclusive* given a prediction.[1] Its core argument, which we elaborate in §2, is that if alternative attention distributions exist that produce similar results to those obtained by the original

- Attentions distributions don't necessarily give a 'true' explanation
- Completely different attention distributions can produce identical predictions
- Attention scores can still often provide plausible explanations for many tasks

Attention is not Explanation. Jain and Wallace (2019)
Attention is not not Explanation. Wiegreffe and Pinter (2019)



Prediction: Positive

the worm in my brain says it is good .

Prediction: Negative

literally unplayable .

Prediction: Negative

i am bad at the game . but instead of taking accountability and improve my skills , i will blame the game instead .

Prediction: Positive

the experience playing this game is like eating a medium - rare steak after years of eating out of dumpsters .

Prediction: Negative

$ 70 of pure disappointment

Prediction: Negative

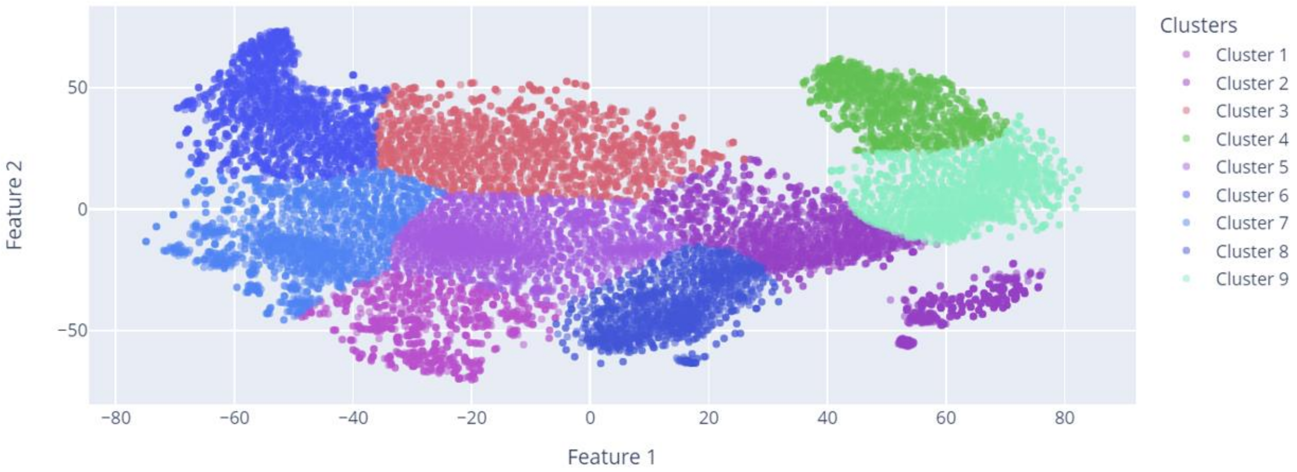
worst game i ' ve ever played . 10 / 10

Embeddings

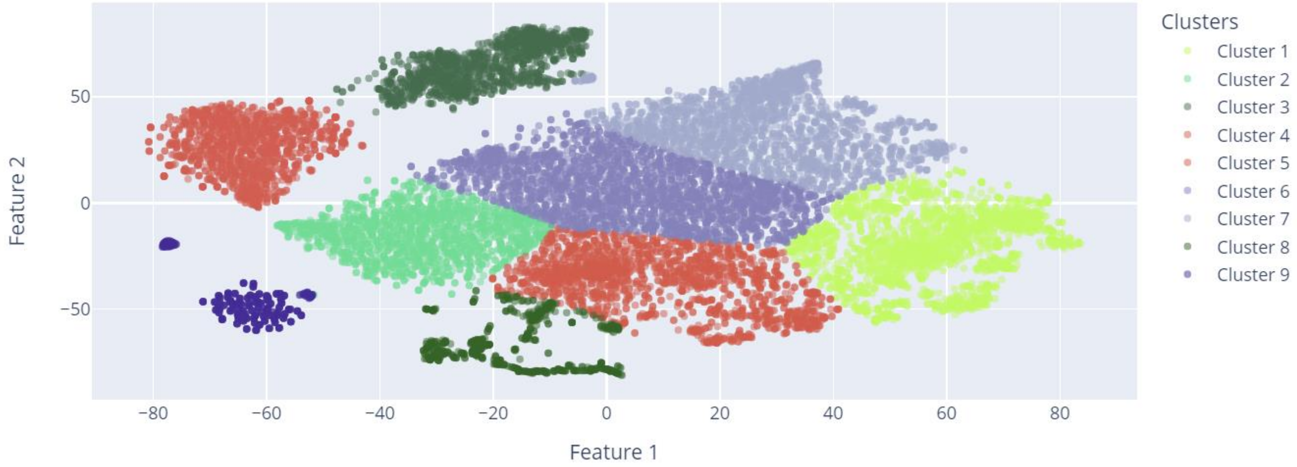
- By comparing embedding layers of different models, we can understand how models are fine-tuned and how they understand languages.
- Use t-SNE(t distributed stochastic neighbor embedding) for dimension reduction
- Then use GMM(Gaussian Mixture Model) to conduct clustering.

DistilBert

SST2: GMM Clustering with 9 Clusters

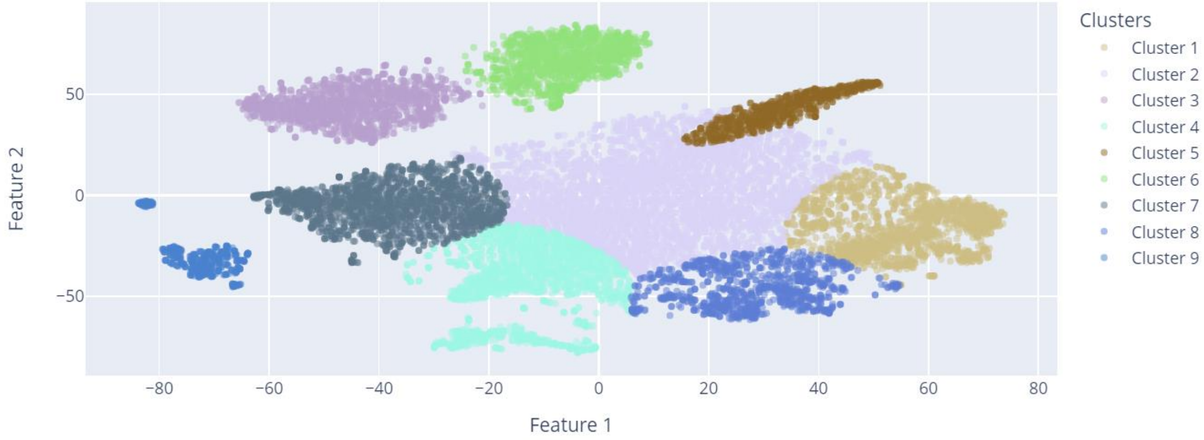


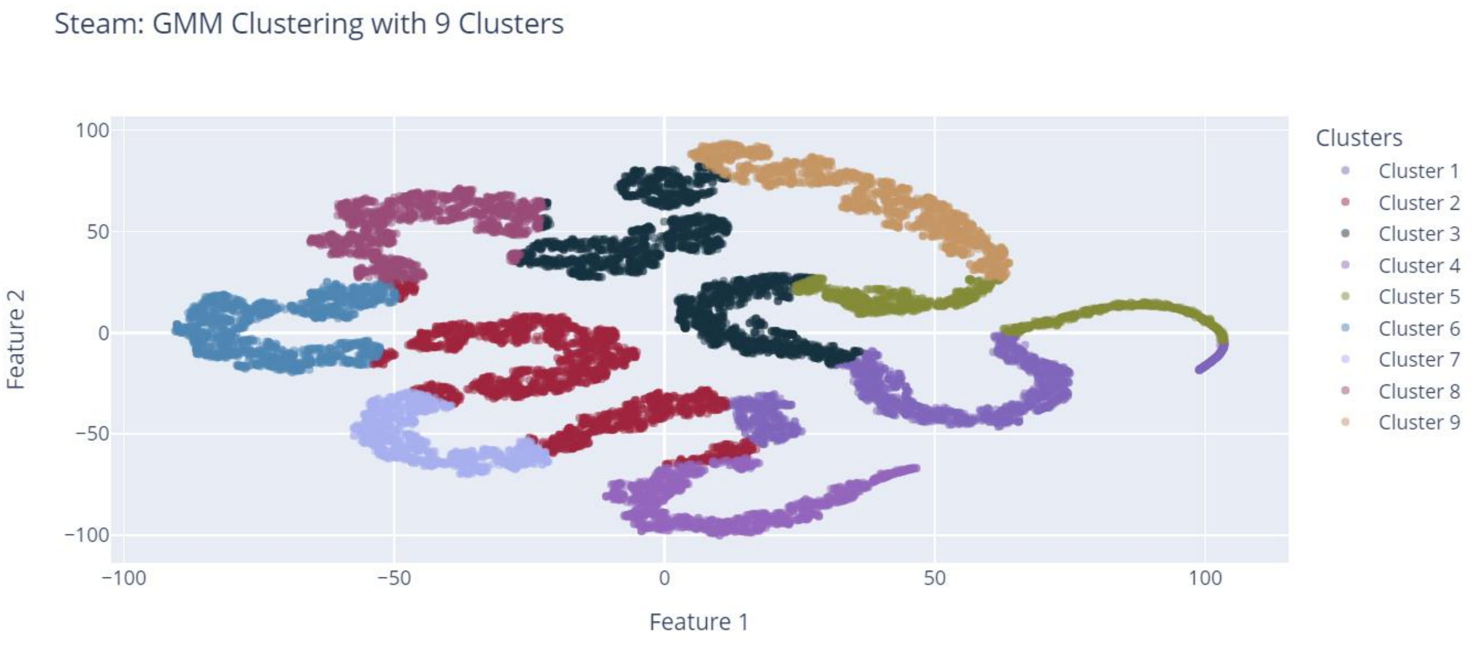
IMDB: GMM Clustering with 9 Clusters

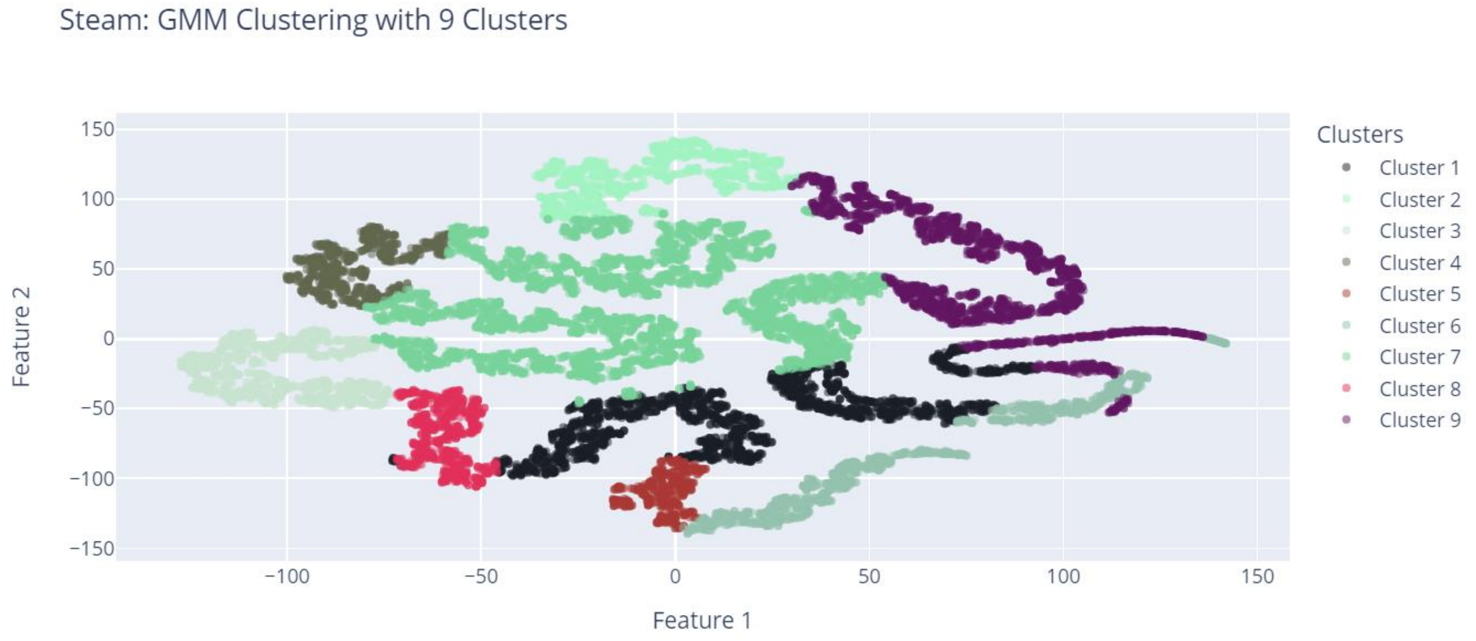


Bert

IMDB: GMM Clustering with 9 Clusters







Visualization on same embedding layer with different hyperparameter of GMM(perplexity)

Conclusion

- DistilBERT performed better
- BERT (arguably) generalized better
- Main hypothesis weakly supported