On Model Stability as a Function of Random Seed

Pranava Madhyastha
Department of Computing
Imperial College London
pranava@imperial.ac.uk

Rishabh Jain
Bloomberg
London
rjain213@bloomberg.net

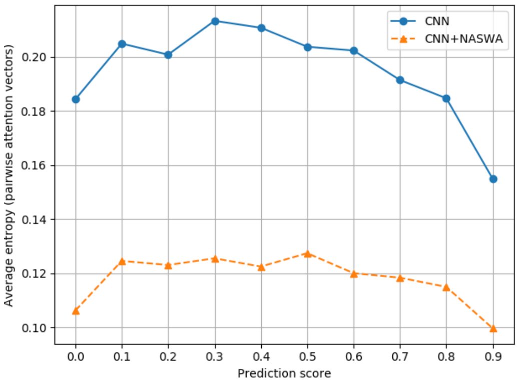


Figure 1: Entropy improvement for tanh Attention based CNN model for the SST dataset using 100 different seeds.

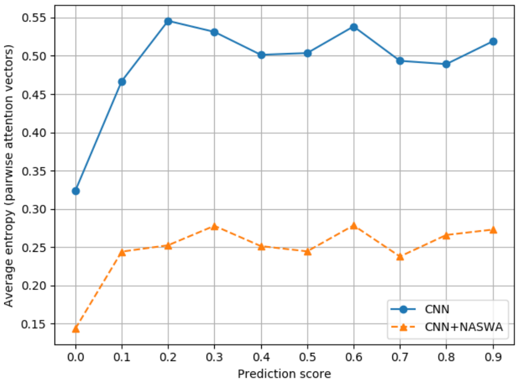


Figure 2: Entropy improvement for tanh Attention based CNN model for the AgNews dataset using 100 different seeds.

A Jaccard Distance Experiments

In Figure **??** and Figure **??** we plot the Jaccard distance plots with CNN models and LSTM models and note that ASWA consistently improves the stability of the interpretations.

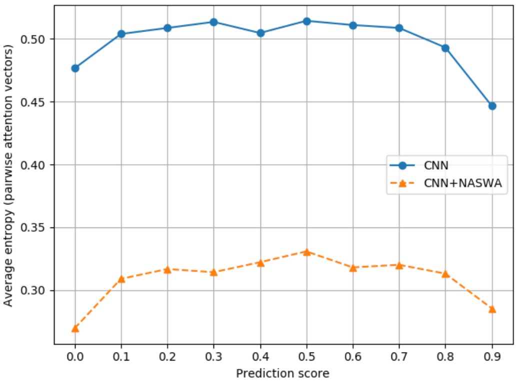


Figure 3: Entropy improvement for tanh Attention based CNN model for the IMDB dataset using 100 different seeds.

B Results with 100 seeds

We perform attention stability based experiments as mentioned in the paper, but now with 100 seeds, i.e, 100 models initialized with different seeds instead of 10. Figures 1, 3, and 2 show the entropy of attention based interpretations for different datasets. Experimenting with 100 seeds helps strengthen our claims about the instability of the model and the effectiveness of our proposed algorithms like ASWA and NASWA.

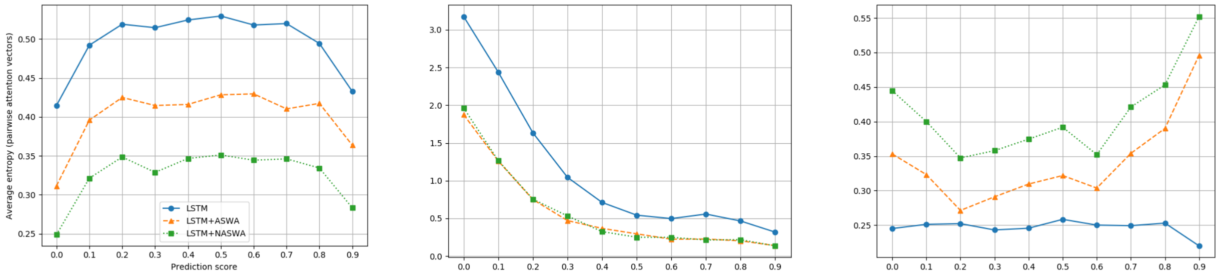
C Hyperparameter Settings

For training purposes, we use the same model settings for the models as mentioned in the paper (Jain and Wallace, 2019) (or the Github implementation[1]), our port of the code is made available at: `https://github.com/rishj97/ModelStability`. Additional hyper-parameters for replication studies are:

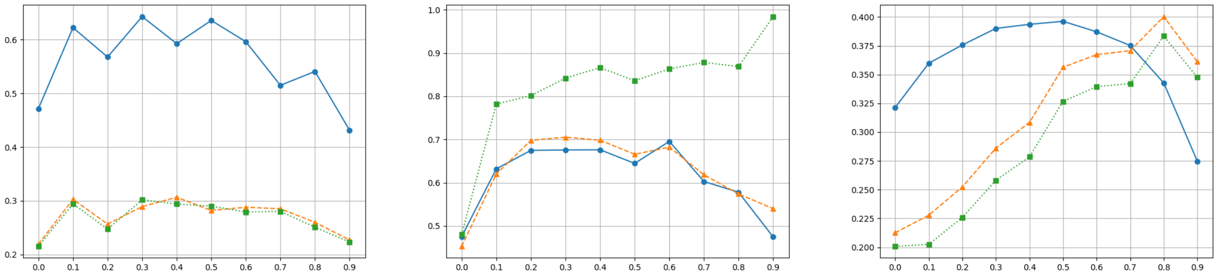
- Number of epochs: 20
- Optimizer: Adam

---

[1] `https://github.com/successar/AttentionExplanation`



(a) IMDB (b) Diabetes (c) SST



(d) Anemia (e) AgNews (f) ADR Tweets

Figure 4: Attention stability improvement from ASWA and NASWA on LSTM based models.

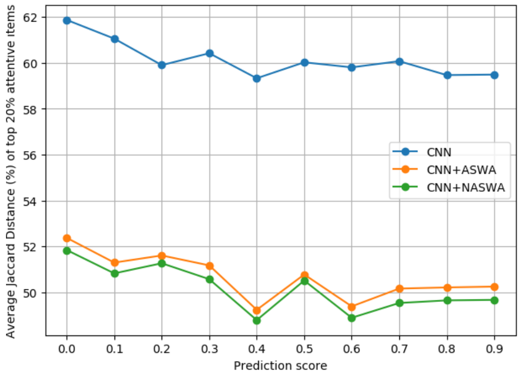


Figure 5: Jaccard Distance improvement for Diabetes.

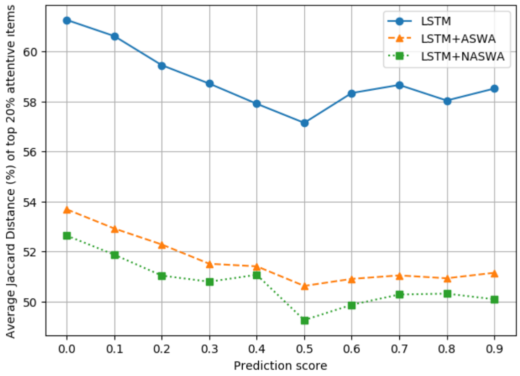


Figure 6: Jaccard Distance improvement for Diabetes.

- Learning rate: 0.001

The exact seeds used for running the experiments can be found in our code repository.

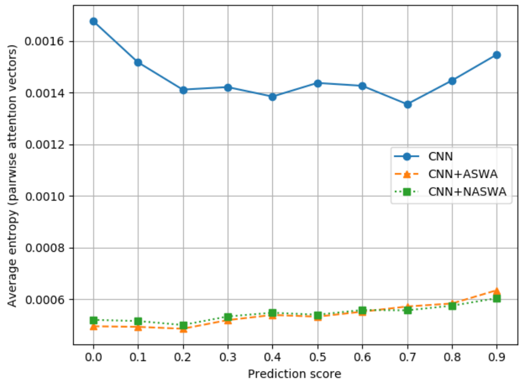


Figure 7: Entropy improvement for dot Attention based CNN model for SST dataset.

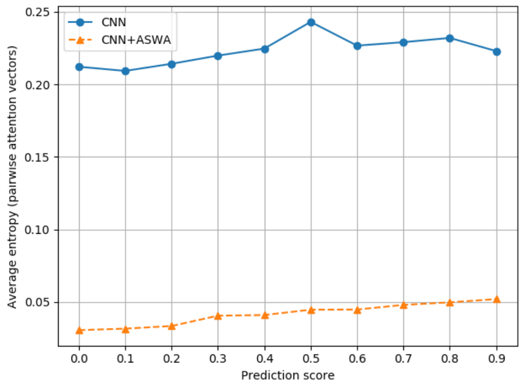


Figure 8: Entropy improvement for dot Attention based CNN model for Diabetes dataset.

D Binary Classification with LSTM based Models

For LSTM based models, we notice (in Figure 4) similar trends as to the CNN models in terms of the instability of the attention based interpretations.

References

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *CoRR*, abs/1902.10186.