Input-specific Attention Subnetworks for Adversarial Detection

Emil Biju, Anirudh Sriram, Pratyush Kumar, Mitesh M. Khapra

Indian Institute of Technology Madras

Abstract

- Self-attention heads are characteristic of Transformer models and have been well studied for interpretability and pruning
- **Problem:** To demonstrate an altogether different utility of attention heads, namely for adversarial detection.
- We propose a method to construct input-specific attention subnetworks (IAS) from which we extract three features to discriminate between authentic and adversarial inputs.
- We demonstrate this utility across 10 NLU datasets and 11 different attack types.

Data Generation

- First, we fine-tune a BERT-based model for each task using its publicly available training set. Then, samples from its test set for which the fine-tuned model makes correct predictions constitute the set of authentic samples for that task.
- Second, we generate adversarial samples by attacking the fine-tuned model using a broad set of 11 hard attack types to comprehensively test AdvNet's performance and its generalizability to diverse perturbations.
- The attacks include **word-level attacks**: deletion , antonyms, synonyms, embeddings), order swap, PWWS, TextFooler and **character-level attacks**: substitution, deletion, insertion, order swap.

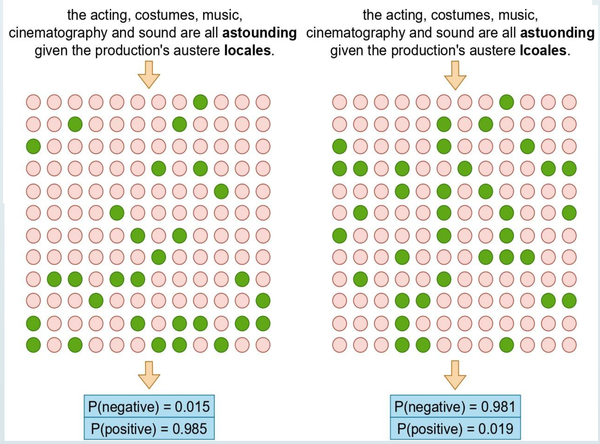
AdvNET Model

- The AdvNET model takes as input a vector F(x) which is the concatenation of **Fmask, Fflip, Flw** and generates a binary output classifying if a given input is authentic or adversarial.
- Since adversarial inputs are slow and computationally expensive to generate, we employ the **CutMix** algorithm for data augmentation
- AdvNet consists of two 1-D convolutional layers with ReLU activation, two fully connected layers with sigmoid activation, and a final classification layer with softmax activation.

Input specific Attention Sub-networks

For each input sample we find the most sparse subset of heads which when retained during pruning will still lead to sub-network to generate the correct output. This subnetwork is referred to as Input specific attention subnetwork(IAS). From these IAS we generate three set of features that is passed as input to the classification AdvNET model.

- **Fmask:** The first feature we extract, Fmask, is just the pre-activation value p for the gating values of each head in the IAS.
- **Fflip:** We flip the gating values of heads in the middle layers of IAS, specifically, the middle (n/3) layers, i.e., we drop heads that were earlier active and include earlier inactive heads.
- **Flw:** Instead of having a single classifier head processing the output of the final layer, we propose to train a classifier head at the output of each layer and use the classes predicted by them as features in adversarial detection

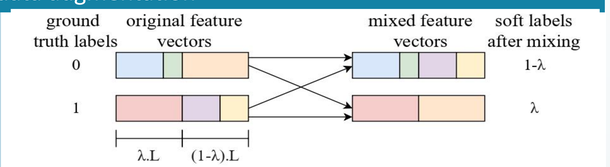


Experiments

- We then perform a comparative study with state-of-the-art adversarial detection methods.
- We perform ablation studies to understand the effect of task, model size, feature combinations and training set attacks on the performance of AdvNet.
- We also analyse the IAS and the constituent features of AdvNet.

CutMix Algorithm for data augmentation

In CutMix, we slice out patches from feature vectors of multiple inputs in the training set, each of which could be authentic or adversarial, and combine them to generate new feature vectors. Their respective ground truth labels are mixed in proportion to the length contributed by each patch.Using soft labels by mixing ground truth labels also offers better generalization and learning speed



Results

- We analyzed our model's performance on both across 10 different tasks, namely SST-2, Yelp polarity, IMDb, AG News, MRPC, RTE, MNLI, SNLI, QQP1 and QNLI, and 11 different attack types.The resultant detector significantly improves (**by over 7.5%**) the state-of-the-art adversarial detection accuracy for the BERT encoder.
- We report an improvement of **6.53%** for the 3 sentiment analysis datasets (SST-2, Yelp, IMDb), **8.05%** for the 4 NLI datasets (RTE, SNLI, MNLI, QNLI) and **6.98%** for the 2 paraphrase detection datasets (MRPC, QQP) over the respective best methods
- **Effect of Model size:** We observe that, across datasets, AdvNet performs better in detecting adversarial inputs fed to the larger BERT-Base model (108M parameters) as opposed to the smaller BERT-Small model (25M parameters). The increase in accuracy averaged across tasks is a significant 10.76%.
- **Effect of training set size:** We observe that AdvNet performs well even when it uses only a fraction of the training set. Specifically, even at 40% of the training examples used, AdvNet outperforms the results obtained with existing state-ofthe-art models on most tasks. This suggests that the CutMix data augmentation is effective and the AdvNet model is sample-efficient.
- **Effect of feature combinations:** We observe that Fmask performs better than Fflip and Flw when used separately as only features. When we used the boolean attention mask Fbmask instead of the realvalued vector Fmask along with Fflip and Flw, we got a decrease in the model performance.The lower accuracy indicates that the real values are more informative.
- **Effect of Cutmix:** we test the model performance when CutMix is not used and conclude that augmenting the training set using CutMix provides higher accuracy.
- **Defense Transferability Analysis**: We perform a study to understand how well the model can perform on unseen attack types. For this purpose, we train AdvNet with samples from only x% of the 11 attack types and report results both on test samples from the remaining attack types and the complete test set for x $\in$ {25, 50, 75}. We observe that even when AdvNet is trained with only 75% of the attack types, the test results on new attacks outperform existing approaches for most datasets, thus showing that our model can generalize to unseen attack methods.

Conclusion

In this paper, we present an altogether new utility of attention heads in Transformer networks - to detect adversarial attacks. We demonstrated that our approach significantly improves the state-of-the-art accuracy across datasets and attack types.In future work, we would like to extend this study to tasks beyond NLU, including vision and speech-related tasks.