Pathologies of Neural Models Make Interpretations Difficult

Shi Feng[1] **Eric Wallace**[1] **Alvin Grissom II**[2] **Mohit Iyyer**[3,4]
Pedro Rodriguez[1] **Jordan Boyd-Graber**[1]
[1]University of Maryland [2]Ursinus College
[3]UMass Amherst [4]Allen Institute for Artificial Intelligence
{shifeng,ewallac2,entilzha,jbg}@umiacs.umd.edu,
agrissom@ursinus.edu, miyyer@cs.umass.edu

1 Introduction

One way to interpret neural model predictions is to highlight the most important input features—for example, a heatmap visualization over the words in an input sentence. In existing interpretation methods for NLP, a word's importance is determined by either input perturbation—measuring the decrease in model confidence when that word is removed—or by the gradient with respect to that word. To understand the limitations of these methods, we use input reduction, which iteratively removes the least important word from the input. This exposes pathological behaviors of neural models: the remaining words appear nonsensical to humans and are not the ones determined as important by interpretation methods. We confirm with human experiments that the reduced examples lack information to support the prediction of any label, but models still make the same predictions with high confidence. To explain these results, we draw connections to adversarial examples and confidence calibration: pathological behaviors reveal difficulties in interpreting neural models trained with maximum likelihood. To mitigate their deficiencies, we fine-tune the models by encouraging high entropy outputs on reduced examples. Fine-tuned models become more interpretable under input reduction without accuracy loss on regular examples. This is work accepted to EMNLP 2018.

2 Input Reduction

Instead of looking at the words with high importance values, we study how the model behaves when the supposedly unimportant words are removed. Intuitively, the important words should remain after the unimportant ones are removed.

We experiment with three popular datasets: SQUAD Rajpurkar et al. [2016] for reading comprehension, SNLI Bowman et al. [2015] for textual entailment, and VQA Antol et al. [2015] for visual question answering. We describe each of these tasks and the model we use below, providing full details in the Supplement.

During the iterative reduction process, we ensure that the prediction does not change (exact same span for SQUAD); consequently, the model accuracy on the reduced examples is identical to the original. The predicted label is used for input reduction and the ground-truth is never revealed. We use the validation set for all three tasks.

Most reduced inputs are nonsensical to humans, as they lack information for *any* reasonable human prediction. However, models make confident predictions, at times even more confident than the original.

With beam search, input reduction finds extremely short reduced examples with little to no decrease in the model's confidence on its original predictions. On SQUAD and SNLI the confidence decreases slightly, and on VQA the confidence even increases.

Preprint. Work in progress.

Dataset	Original	Reduced	vs. Random
SQuAD	80.58	31.72	53.70
SNLI-E	76.40	27.66	42.31
SNLI-N	55.40	52.66	50.64
SNLI-C	76.20	60.60	49.87
VQA	76.11	40.60	61.60

Table 1: Human accuracy on *Reduced* examples drops significantly compared to the *Original* examples, however, model predictions are identical. (*-E*), neutral (*-N*), and contradiction (*-C*).

	Accuracy		Reduced length	
	Before	After	Before	After
SQuAD	77.41	78.03	2.27	4.97
SNLI	85.71	85.72	1.50	2.20
VQA	61.61	61.54	2.30	2.87

Table 2: Model *Accuracy* on regular validation examples remains largely unchanged after fine-tuning.

3 Making Sense of Reduced Inputs

3.1 Model Overconfidence

Neural models are overconfident in their predictions Guo et al. [2017]. One explanation for overconfidence is overfitting: the model overfits the negative log-likelihood loss during training by learning to output low-entropy distributions over classes. They are also overconfident on examples outside the training data distribution. On image classification, samples from pure noise can sometimes trigger highly confident predictions.Goodfellow et al. [2015] These *rubbish examples* are degenerate inputs that a human would trivially classify as not belonging to any class but for which the model predicts with high confidence. The confidence of a neural model is not a robust estimate of its prediction uncertainty.

Our reduced inputs satisfy the definition of rubbish examplesGoodfellow et al. [2015]: humans have a hard time making predictions based on the reduced inputs (Table 1), but models make predictions with high confidence. Starting from a valid example, input reduction transforms it into a rubbish example.

4 Mitigating Model Pathologies

To maximize model uncertainty on reduced examples, we use the entropy of the output distribution as an objective. Given a model f trained on a dataset $(\mathcal{X}, \mathcal{Y})$, we generate reduced examples using input reduction for all training examples $\mathcal{X}$. We collect all of these versions together to form $\tilde{\mathcal{X}}$ as the "negative" example set.

We fine-tune the existing model to simultaneously maximize the log-likelihood on regular examples and the entropy on reduced examples: where hyperparameter λ controls the trade-off between the two terms. Similar entropy regularization is used by Pereyra et al. [2017], but not in combination with input reduction; their entropy term is calculated on regular examples rather than reduced examples.

On regular examples, entropy regularization does no harm to model accuracy, with a slight increase for SQuAD. After entropy regularization, input reduction produces more reasonable reduced inputs

Human accuracy increases across all three tasks. We also repeat the *vs. Random* experiment: we re-generate the random examples to match the lengths of the new reduced examples from input reduction, and find humans now prefer the reduced examples to random ones. The increase in both human performance and preference suggests that the reduced examples are more reasonable; model pathologies have been mitigated.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *International Conference on Computer Vision*, 2015.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of Empirical Methods in Natural Language Processing*, 2015.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations*, 2015.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference of Machine Learning*, 2017.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. In *Proceedings of the International Conference on Learning Representations*, 2017.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100, 000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*, 2016.