Penalizing Confident Predictions on Largely Perturbed Inputs Does Not Improve Out-of-Distribution Generalization in Question Answering

Kazutoshi Shinoda, Saku Sugawara, Akiko Aizawa

@shino__c kazutoshi.shinoda0516@gmail.com

東京大学 THE UNIVERSITY OF TOKYO

NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

KnowledgeNLP-AAAI'23@Washington, DC

Summary

- QA models are shown to be insensitive to several types of perturbations that remove features that are necessary for human reading from inputs.
- We show that entropy maximization for four types of perturbations is effective to make QA models recognize those features..
- Even though models become sensitive to the above features, the out-of-distribution generalization (other QA datasets / adversarial SQuAD) is not improved.

Background 1: Insensitivity of QA Models

QA models often maintain high accuracy and confidence scores even though features necessary for human reading are removed from inputs; e.g., question word deletion (Feng+ 2018,Sugawara+ 2018), word order shuffling, sentence deletion, and sentence order shuffling (Sugawara+ 2020). These phenomena imply that QA models do not have human-like language understanding abilities.

Background 2: Lack of Out-of-Distribution Generalization

QA models lack out-of-distribution (OOD) generalization. Namely, QA models trained on a certain dataset fail to generalize to datasets from other domains (Talmor and Berant, 2019). They also lack robustness to adversarial attacks that append fake sentences to contexts (Jia and Liang, 2017).

Research Question

If the overconfident predictions of QA models for various types of perturbations are penalized, will the OOD generalization be improved?

Examined Perturbations

We examined four types of perturbations (as shown below) that remove word- and sentence-level semantics and syntactics from inputs. We adopted these perturbations because a QA model is relatively insensitive to them compared to other types of perturbations as found in Sugawara+ (2020)

	Perturbation σ	Description	Intended feature removed by perturbation
σ_1	Del_{func}	Delete all the function words	Function words
σ_2	Del_{que}	Delete the question	Question words
σ_3	$\text{Shuf}_{\text{word}}$	Shuffle the word order in each sentence	Syntactic information
σ_4	$\text{Shuf}_{\text{sent}}$	Shuffle the sentence order in a context	Discourse relations

Example of Original and Largely Perturbed Inputs

Original input

Context	The American Football Conference (AFC) champion **Denver Broncos** defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title.
Question	Which NFL team represented the AFC at Super Bowl 50?

Original Input x

Perturbed with function word deletion

Context	American Football Conference AFC champion **Denver Broncos** defeated National Football Conference NFC champion Carolina Panthers 24 earn third Super Bowl title.
Question	NFL team represented AFC Super Bowl 50?

Perturbed Input x_{σ_1}

Perturbed with word order shuffling

Context	an Carolina the Super 10 American The National third their defeated NFC Conference champion Football to **Denver Broncos** 24 AFC (Panthers (champion.
Question	at represented NFL team the AFC 50 Which Bowl Super?

Perturbed Input x_{σ_3}

Method

To penalize confident predictions on largely perturbed inputs, we employed entropy maximization (Feng+ 2018). The loss function to be minimized is computed as follows:

$$\mathcal{L} = \mathcal{L}_{ce} - \lambda_\sigma H(Y|X_\sigma).$$

where $\mathcal{L}_{ce}$ is the cross entropy loss and H is the entropy term which is given by

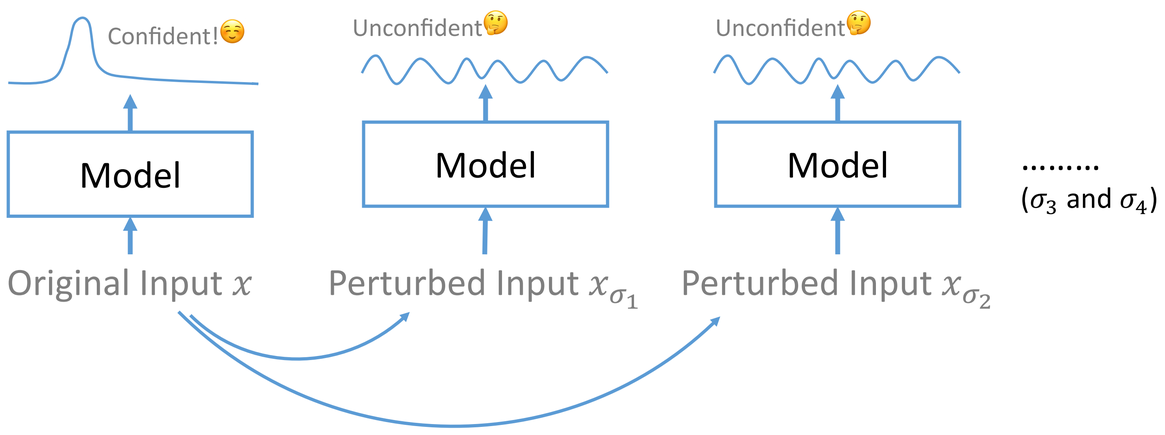
$$H(Y|X_\sigma) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} -p_\theta(y|x_\sigma) \log p_\theta(y|x_\sigma).$$

Our experiments show that maximizing entropy for a certain perturbation type does not transfer to unseen perturbation types. To mitigate the lack of transferability, we propose to maximize the entropy term for the four type of perturbations to make models recognize those features as follows:

$$\mathcal{L} = \mathcal{L}_{ce} - \sum_\sigma \lambda_\sigma H(Y|X_\sigma). \quad (\star)$$

(The method is also motivated by the perspective of causality. See § 2.5 in our paper!)

How Our Method (☆) Changes Model Predictions



Experiments

- Model: BERT-base and RoBERTa-base
- Dataset: SQuAD 1.1 (for train and test), NewsQA, TriviaQA, SearchQA, HotpotQA, Natural Questions, and Adversarial SQuAD (for test)
- Training: The scaling factor λ_σ is tuned in {0.01, 0.1, 1.0, 5.0} for each perturbation σ.

Results:

F1 scores on the Original and Perturbed SQuAD dev set.

Model	Perturbation train↓ / test→	None	Del_{func}	Del_{que}	$\text{Shuf}_{\text{word}}$	$\text{Shuf}_{\text{sent}}$
BERT-base	None	88.0±0.03	54.2±0.06	10.2±0.41	26.5±0.14	83.9±0.06
	Del_{func}	88.1±0.02	**22.2**±3.83	10.2±0.28	24.2±0.73	83.8±0.26
	Del_{que}	88.1±0.12	53.9±0.91	**5.9**±0.74	26.4±0.36	84.1±0.14
	$\text{Shuf}_{\text{word}}$	88.1±0.07	36.4±0.32	10.0±0.37	**16.2**±1.53	83.8±0.25
	$\text{Shuf}_{\text{sent}}$	88.0±0.09	54.3±0.79	9.9±0.53	26.8±0.29	83.9±0.18
	ALL	88.0±0.10	31.1±2.61	7.9±1.81	19.1±0.41	83.9±0.14
RoBERTa-base	None	91.2±0.04	61.0±0.72	11.3±0.33	29.3±0.06	87.3±0.21
	Del_{func}	91.4±0.01	**14.5**±2.21	11.0±0.21	19.2±0.88	87.4±0.12
	Del_{que}	91.2±0.13	60.9±0.53	**7.0**±2.44	28.9±0.41	87.5±0.12
	$\text{Shuf}_{\text{word}}$	91.2±0.17	47.8±4.34	11.2±0.12	12.1±2.05	86.8±0.30
	$\text{Shuf}_{\text{sent}}$	91.3±0.05	59.9±0.50	10.2±0.70	10.0±1.87	**17.0**±5.06
	ALL	91.3±0.08	19.6±3.74	8.9±2.46	**9.7**±1.56	34.8±7.65
Human Score		91.2†	28.1	0.1	10.8	53.2

➡ (1) Standard QA models (None) achieve higher scores on perturbed inputs than humans. (2) Entropy maximization is effective only for penalizing high performance on seen perturbation types. (3) Our method (ALL) suppresses all the perturbation types while maintaining the performance on original inputs.

F1 scores on test sets in other domains.

Model	Perturbation	SearchQA	HotpotQA	NQ	NewsQA	TriviaQA
BERT-base	None	27.3±0.60	60.6±0.44	59.1±0.50	55.8±0.26	58.5±0.27
	Del_{func}	27.2±0.98	60.0±0.37	56.2±0.39	55.9±0.37	58.6±0.19
	Del_{que}	27.4±0.71	60.0±0.21	58.7±0.27	55.5±0.43	58.6±0.46
	$\text{Shuf}_{\text{word}}$	27.8±0.29	60.1±0.02	56.7±0.42	55.9±0.52	58.7±0.16
	$\text{Shuf}_{\text{sent}}$	27.6±1.83	60.2±0.20	58.8±0.45	56.0±0.15	58.6±0.08
	ALL	28.0±1.08	60.7±0.45	56.9±0.81	55.3±0.44	57.9±0.49
RoBERTa-base	None	30.7±1.90	66.5±0.73	61.8±0.39	64.3±0.11	62.7±0.30
	Del_{func}	26.8±2.49	66.6±0.25	61.4±0.22	64.5±0.21	62.1±0.58
	Del_{que}	31.5±0.99	66.2±0.31	62.0±0.45	64.7±0.34	62.8±0.36
	$\text{Shuf}_{\text{word}}$	23.6±1.65	66.7±0.48	57.0±2.39	64.6±0.07	62.2±0.37
	$\text{Shuf}_{\text{sent}}$	28.6±1.03	66.2±0.42	17.5±3.90	64.7±0.27	61.4±0.43
	ALL	14.4±3.51	66.5±0.81	25.3±4.56	63.6±0.24	60.6±0.38

F1 scores on adversarial test sets.

Model	Perturbation	AddSent	AddOneSent
BERT-base	None	50.8±0.40	62.1±0.89
	Del_{func}	49.6±0.54	61.4±1.26
	Del_{que}	50.7±0.88	62.2±0.39
	$\text{Shuf}_{\text{word}}$	49.9±0.63	61.8±1.14
	$\text{Shuf}_{\text{sent}}$	49.9±0.87	61.6±1.08
	ALL	50.7±0.71	62.2±0.74
RoBERTa-base	None	62.6±0.90	72.0±0.95
	Del_{func}	62.4±1.00	71.6±1.33
	Del_{que}	61.9±0.94	71.6±0.84
	$\text{Shuf}_{\text{word}}$	61.5±0.37	70.8±0.65
	$\text{Shuf}_{\text{sent}}$	62.2±1.61	71.6±1.10
	ALL	61.9±1.11	71.4±0.41

➡ Contrary to our expectations, the OOD generalization is not improved after increasing the sensitivity to the four perturbation types.

Discussion & Conclusion

- Penalizing Confident Predictions on Largely Perturbed Inputs Does Not Improve Out-of-Distribution Generalization in Question Answering.
- The perturbed inputs are not so natural that they are unlikely included in the test sets. This may be why the OOD generalization is not improved by our method.
- To make QA models use intended features like humans, how to design perturbations is worth being studied in future work. For example, eliminating a feature while maintaining the naturalness of inputs may be an approach worth studying.

References

- Feng+ 2018. Pathologies of Neural Models Make Interpretations Difficult. In EMNLP.
- Sugawara+ 2018. What Makes Reading Comprehension Questions Easier? In EMNLP.
- Sugawara+ 2020. Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. In AAAI.
- Talmor and Berant. 2019. MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension. In ACL.
- Jia and Liang 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In EMNLP.

Link

arXiv: