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

Kazutoshi Shinoda[1,2], **Saku Sugawara**[2], **Akiko Aizawa**[1,2]
[1]The University of Tokyo
[2]National Institute of Informatics
shinoda@is.s.u-tokyo.ac.jp, {saku, aizawa}@nii.ac.jp

Abstract

Question answering (QA) models are shown to be insensitive to large perturbations to inputs; that is, they make correct and confident predictions even when given largely perturbed inputs from which humans can not correctly derive answers. In addition, QA models fail to generalize to other domains and adversarial test sets, while humans maintain high accuracy. Based on these observations, we assume that QA models do not use intended features necessary for human reading but rely on spurious features, causing the lack of generalization ability. Therefore, we attempt to answer the question: If the overconfident predictions of QA models for various types of perturbations are penalized, will the out-of-distribution (OOD) generalization be improved? To prevent models from making confident predictions on perturbed inputs, we first follow existing studies and maximize the entropy of the output probability for perturbed inputs. However, we find that QA models trained to be sensitive to a certain perturbation type are often insensitive to unseen types of perturbations. Thus, we simultaneously maximize the entropy for the four perturbation types (i.e., word- and sentence-level shuffling and deletion) to further close the gap between models and humans. Contrary to our expectations, although models become sensitive to the four types of perturbations, we find that the OOD generalization is not improved. Moreover, the OOD generalization is sometimes degraded after entropy maximization. Making unconfident predictions on largely perturbed inputs per se may be beneficial to gaining human trust. However, our negative results suggest that researchers should pay attention to the side effect of entropy maximization.

1 Introduction

Pretrained language models (Devlin et al. 2019; Liu et al. 2019; Lewis et al. 2020; Radford et al. 2019; Brown et al. 2020) have achieved human-level performance on natural language understanding (NLU) tasks, such as question answering (QA) (Rajpurkar et al. 2016), and natural language inference (NLI) (Williams, Nangia, and Bowman 2018). Additionally, recent studies have shown that pretrained language models capture linguistic features based on an analysis with probing classifiers (Tenney et al. 2019; Hewitt and Manning 2019; Hewitt and Liang 2019; Belinkov 2022).

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?
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 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?

Table 1: Examples of largely perturbed inputs taken from SQuAD. In word order shuffling, we ensure that the answer spans indicated by **bold** remain as they are.

However, whether language models fine-tuned on QA tasks have human-like language understanding abilities remains debatable. QA models often maintain high accuracy and confidence scores even when the inputs are transformed by large perturbations (e.g., word deletion (Feng et al. 2018; Sugawara et al. 2018), word order shuffling, sentence deletion, and sentence order shuffling (Sugawara et al. 2020)). See Table 1 for the examples of largely perturbed inputs in extractive QA. Similar phenomena have been observed in NLI (Sinha et al. 2021), and other NLU tasks (Gupta, Kvernadze, and Srikumar 2021). Meanwhile, our human evaluation (§3.2) indicates that humans cannot correctly derive answers from such invalid inputs. We argue that such phenomena imply that models do not adequately use semantic and syntactic features removed by perturbations to make predictions, while these features are indispensable for humans to

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

Table 2: Four types of perturbations σ studied in this work. Different perturbations remove different types of intended features necessary for human reading from the inputs.

understand language.

In addition, QA models are shown to lack out-of-distribution generalization. QA models trained on a certain dataset fail to generalize to datasets from other domains (Yogatama et al. 2019; Talmor and Berant 2019; Sen and Saffari 2020). They also lack robustness to adversarial attacks that append fake sentences to contexts (Jia and Liang 2017).

Given the two characteristics of QA models (i.e., the insensitivity to large perturbations and the lack of generalization ability), we aim to answer the following question: *If the overconfident predictions of QA models for various types of perturbations are penalized, will the out-of-distribution (OOD) generalization be improved?* Previous studies have shown that maximizing the entropy (Shannon 1948) of the output probability for perturbed inputs can successfully reduce model confidence for such perturbed inputs (Feng et al. 2018; Sinha et al. 2021; Gupta, Kvernadze, and Srikumar 2021). We adopt this method to make QA models sensitive to the perturbations listed in Table 2.

However, we observe that entropy maximization for a certain perturbation type can transfer to the seen perturbation type but often fails to transfer to unseen perturbation types. For example, after maximizing the entropy for question deletion, models are not sensitive to function word deletion. To mitigate this lack of transferability, we propose to simultaneously maximize the entropy for the predefined perturbation types. We show that this approach is effective to make models recognize all the predefined perturbations while maintaining in-domain accuracy.

Contrary to our expectations, even though models become sensitive to the four types of perturbations, we find that the generalization to other domains or adversarial robustness is not improved. As discussed in Hase, Xie, and Bansal (2021), intentionally perturbed inputs become unnatural and are unlikely to appear in a dataset. Therefore, making models sensitive to largely perturbed inputs may a have negative impact on out-of-distribution generalization. While becoming sensitive to unnatural inputs with entropy maximization like humans can gain trust from humans, our results suggest that researchers should pay attention to the side effect of entropy maximization.

Our main contributions are as follows:

- We find that entropy maximization can mitigate the insensitivity to seen perturbation types, but fail to transfer to unseen perturbation types in QA.
- We show that simply maximizing the entropy for the four perturbation types, including word- and sentence-level ones, can mitigate this issue.
- We show that even though QA models become sensitive to the four types of perturbations, the generalization to other domains or adversarial robustness is not improved but rather sometimes degraded.

2 Method

2.1 Perturbation Types

We list the examined perturbations in Table 2. We adopt two word-level perturbations, function word deletion ($\mathrm{Del}_{\mathrm{func}}$) and word order shuffling ($\mathrm{Shuf}_{\mathrm{word}}$), and two sentence-level perturbations, question deletion ($\mathrm{Del}_{\mathrm{que}}$) and sentence order shuffling ($\mathrm{Shuf}_{\mathrm{sent}}$) to comprehensively assess the sensitivity of QA models to the intended features necessary for humans to understand language, which cover the surface structure and the textbase of the construction–integration model comprehensively. We adopted these perturbations because a QA model is relatively insensitive to them compared to other types of perturbations (e.g., contend word deletion and vocabulary anonymization) as found in Sugawara et al. (2020). We expect that entropy maximization with these perturbations make models learn to recognize the intended features as shown in Table 2. The detailed motivation of the expectation is described in §2.5.

2.2 Entropy Maximization

To penalize the confident predictions of models on perturbed inputs, we adopt entropy maximization used by Feng et al. (2018); Gupta, Kvernadze, and Srikumar (2021). Namely, we minimize the cross-entropy loss while maximizing the entropy of the output probabilities given perturbed inputs.

When the dataset $\mathcal{D}$ consists of pairs of input x and output y, and the model parameters are θ, the cross-entropy loss is given by

$$\mathcal{L}_{ce} = -\frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \log p_\theta(y|x). \quad (1)$$

When a perturbed input x_σ is obtained from x by applying a perturbation σ, the entropy of the model output given perturbed input is given by[1]:

$$H(Y|X_\sigma) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} -p_\theta(y|x_\sigma) \log p_\theta(y|x_\sigma). \quad (2)$$

The loss function to be minimized is computed as follows:

$$\mathcal{L} = \mathcal{L}_{ce} - \lambda_\sigma H(Y|X_\sigma). \quad (3)$$

where the entropy term is scaled by the factor λ_σ (> 0).

[1]Uppercase letters (e.g., X) represent random variables and lowercase letters (e.g., x) represent actual values.

2.3 Conditional Independence Assumption for Extractive QA

In extractive QA tasks, such as SQuAD (Rajpurkar et al. 2016), the models need to specify the start and end positions of the predicted answer span in the context for the given question. When computing the conditional probability $p_\theta(y|x)$ in Equations 1 and 2, as most existing studies implicitly did during training (Seo et al. 2017; Devlin et al. 2019), we assume that the start and end positions of answers, i.e., Y_{start} and Y_{end}, are conditionally independent given the context and question for brevity. Namely, we assume that $p(Y|X) = p(Y_{start}|X)p(Y_{end}|X)$. Based on this assumption, the entropy term in Equations 3 and 5 can be computed as follows:

$$H(Y|X_\sigma) = H(Y_{start}|X_\sigma) + H(Y_{end}|X_\sigma). \quad (4)$$

We adopt this relaxation because it is costly to raise all the possible answer spans meeting the condition that the start position is lower than or equal to the end position. Our experiments show that this does not degrade the in-distribution accuracy.[2]

2.4 Recognizing Multiple Types of Perturbations

Our experiments in §3.3 show that maximizing entropy for a certain perturbation type does not transfer to unseen perturbation types. We need to mitigate this problem because our aim is to investigate whether making models sensitive to the four types of perturbations in Table 2 improves out-of-distribution generalization.

To mitigate the lack of transferabiliity, we propose to maximize the entropy term for the four type of perturbations to make models recognize those features as follows:

$$\mathcal{L} = \mathcal{L}_{ce} - \sum_\sigma \lambda_\sigma H(Y|X_\sigma). \quad (5)$$

2.5 Interpretation from the Perspective of Causality

When the maximum predicted probability (confidence score) of model θ for original input x is $p_\theta(\hat{y}|x) = \max p_\theta(y|x)$, the difference in probabilities that the model assigns to $\hat{y}$,

$$d_\theta(x, \hat{y}, \sigma) = p(\hat{y}|x) - p(\hat{y}|x_\sigma), \quad (6)$$

can be regarded as quantifying how much feature s is used by the model for making prediction $\hat{y}$. Quantities of similar definitions have been used as feature importance (Li, Monroe, and Jurafsky 2016; DeYoung et al. 2020; Hase, Xie, and Bansal 2021) to increase the interpretability of the model, or the degree to which a cause affects an outcome in the context of causality (Pearl 2000).

By minimizing the cross entropy while maximizing the entropy in Equation 3, $p(\hat{y}|x)$ is increased while $p(\hat{y}|x_\sigma)$ is decreased. Thus, minimizing $\mathcal{L}$ in Equation 3 is expected to indirectly increase $d_\theta(x, \hat{y}, \sigma)$ in Equation 6. When $d_\theta(x, \hat{y}, \sigma)$ is larger than non-zero values, features in x removed by perturbation σ have causal effects on prediction $\hat{y}$ made by QA model θ. Based on this interpretation, we assume that training with entropy maximization causes QA models to use intended features as listed in Table 2, and have positive impact on out-of-distribution generalization.

However, our experiments show that the results are opposite to the assumption. We will discuss why the out-of-generalization is not improved with the approach in §3.4.

3 Experiment

In this study, we considered a QA task because the inputs of QA datasets consist of questions and contexts, and the contexts often consist of multiple sentences. This enables examination of broader types of perturbations, such as sentence order shuffling, that are absent in other NLU tasks such as NLI and paraphrase identification (Dolan and Brockett 2005), where the inputs are only two sentences.

3.1 Experimental Setups

Model We used BERT-base (Devlin et al. 2019) and RoBERTa-base (Liu et al. 2019) for QA models because they are often adopted in QA.

Dataset We used SQuAD 1.1 (Rajpurkar et al. 2016) for training and evaluation. To evaluate the generalization to other domains, we employed the dev set of NewsQA (Trischler et al. 2017), TriviaQA (Joshi et al. 2017), SearchQA (Dunn et al. 2017), HotpotQA (Yang et al. 2018), and NaturalQuestions (Kwiatkowski et al. 2019) from MRQA 2019 shared task (Fisch et al. 2019). To evaluate adversarial robustness, we used AddSent and AddOneSent from Adversarial SQuAD (Jia and Liang 2017).

Training We used the Adam (Kingma and Ba 2014) optimizer with epsilon 1e-8. The models were trained for two epochs with the learning rate being linearly decreased from 3e-5 to zero. The batch size was set to 32. For other hyperparameters, we generally used the default hyperparameters in the example code provided by Huggingface. We tuned the scaling factor λ_σ in Equation 3 in $\{0.01, 0.1, 1.0, 5.0\}$ for each perturbation σ on the SQuAD dev set based on the F1 scores. The means and standard deviations of the F1 scores over three random seeds are reported.

3.2 Human Evaluation

To see whether humans can derive correct answers from inputs with the examined perturbations, we conducted a human evaluation. We asked human annotators to answer a question by extracting an answer span from a given context. The annotators are allowed to submit empty answers when they cannot find plausible answers. The input is transformed by one of the four perturbation types. We randomly chose 200 examples for each perturbation from the SQuAD dev set. Three annotators on Amazon Mechanical Turk were assigned to each example.

[2]The entropy of output probabilities can be also defined in multiple-choice and abstractive QA. Extending our work to other QA formats is future work.

Model	Perturbation train↓ / test→	None	Del_{func}	Del_{que}	$\text{Shuf}_{\text{word}}$	$\text{Shuf}_{\text{sent}}$
BERT-base	None	1.38±0.00	3.43±0.16	7.03±1.10	3.53±0.16	1.43±0.00
	Del_{func}	1.37±0.00	**11.9**±0.00	7.1±0.20	10.48±0.36	1.41±0.01
	Del_{que}	1.37±0.01	3.69±0.28	**11.9**±0.00	4.31±0.21	1.41±0.01
	$\text{Shuf}_{\text{word}}$	1.37±0.00	11.87±0.01	7.43±0.27	**11.9**±0.00	1.41±0.00
	$\text{Shuf}_{\text{sent}}$	1.43±0.01	3.65±0.04	7.94±0.35	3.82±0.23	**1.48**±0.01
	ALL	1.41±0.01	**11.9**±0.00	**11.9**±0.00	**11.9**±0.00	1.47±0.01
RoBERTa-base	None	1.13±0.05	3.17±0.85	8.29±0.39	2.76±0.48	1.45±0.02
	Del_{func}	1.10±0.02	**11.9**±0.0	9.05±0.19	11.89±0.00	1.44±0.04
	Del_{que}	1.12±0.03	2.94±0.47	**11.9**±0.00	2.59±0.32	1.42±0.01
	$\text{Shuf}_{\text{word}}$	1.13±0.03	8.95±2.74	8.51±0.31	**11.9**±0.00	2.38±0.28
	$\text{Shuf}_{\text{sent}}$	1.09±0.01	2.27±0.59	9.08±0.23	**11.9**±0.00	**11.9**±0.00
	ALL	1.14±0.04	**11.9**±0.00	**11.9**±0.00	**11.9**±0.00	**11.9**±0.00

Table 3: Entropy of the model predictions on the original and perturbed SQuAD 1.1 development set. The more confident predictions models make, the lower entropy is.

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

Table 4: F1 scores on the original and perturbed SQuAD dev set. See Table 2 for details of perturbation types. †Copied from the SQuAD 1.1 Leaderboard.

3.3 Cross-Perturbation Evaluation

Previous studies have shown that maximizing entropy for input with certain perturbations, such as word deletion (Feng et al. 2018) and word order shuffling (Sinha et al. 2021), can make models sensitive (i.e., less confident) to the *same* type of perturbations at test time. Intuitively, the word orders and words themselves should convey different types of information. Then, we can ask the question: Can maximizing the entropy for word order shuffling transfer to that for word deletion?

To answer this question, we investigated the transferability of entropy maximization across four types of perturbations. That is, we trained the QA models with entropy maximization for one of the four perturbations and evaluated the models with seen and unseen perturbations. To evaluate the sensitivity of the predictions to the perturbations, we used the entropy of the predicted answers and the F1 score.[3] The entropies and F1 scores for the in-distribution dev set are shown in Tables 3 and 4, respectively.

QA Models Are More Insensitive to Large Perturbations Than Humans First, without entropy maximization, models can somehow correctly answer decent portions of perturbed inputs compared with humans. Notably, QA models trained without entropy maximization were most robust to $\text{Shuf}_{\text{sent}}$ among the four perturbation types. This result is consistent with the findings of Sugawara et al. (2020). While the F1 scores decreased substantially and the entropy is relatively high when the inputs in the dev set are perturbed with

[3] In this experiment, because the maximum length of the context is set to 384, the maximum of the entropy is 11.9 ($= -1/384 * \log(1/384) * 384 * 2$), and the minimum is 0.0 based on Equations 2 and 4.

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
	$\mathrm{Del}_{\mathrm{func}}$	27.2±0.98	60.0±0.37	56.2±0.39	55.9±0.37	58.6±0.19
	$\mathrm{Del}_{\mathrm{que}}$	27.4±0.71	60.0±0.21	58.7±0.27	55.5±0.43	58.6±0.46
	$\mathrm{Shuf}_{\mathrm{word}}$	27.8±0.29	60.1±0.02	56.7±0.42	55.9±0.52	58.7±0.16
	$\mathrm{Shuf}_{\mathrm{sent}}$	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
	$\mathrm{Del}_{\mathrm{func}}$	26.8±2.49	66.6±0.25	61.4±0.22	64.5±0.21	62.1±0.58
	$\mathrm{Del}_{\mathrm{que}}$	31.5±0.99	66.2±0.31	62.0±0.45	64.7±0.34	62.8±0.36
	$\mathrm{Shuf}_{\mathrm{word}}$	23.6±1.65	66.7±0.48	57.0±2.39	64.6±0.07	62.2±0.37
	$\mathrm{Shuf}_{\mathrm{sent}}$	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

Table 5: F1 scores on test sets in other domains. The means±standard deviations over three random seeds are reported.

Model	Perturbation	AddSent	AddOneSent
BERT-base	None	50.8±0.40	62.1±0.89
	$\mathrm{Del}_{\mathrm{func}}$	49.6±0.54	61.4±1.26
	$\mathrm{Del}_{\mathrm{que}}$	50.7±0.88	62.2±0.39
	$\mathrm{Shuf}_{\mathrm{word}}$	49.9±0.63	61.8±1.14
	$\mathrm{Shuf}_{\mathrm{sent}}$	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
	$\mathrm{Del}_{\mathrm{func}}$	62.4±1.00	71.6±1.33
	$\mathrm{Del}_{\mathrm{que}}$	61.9±0.94	71.6±0.84
	$\mathrm{Shuf}_{\mathrm{word}}$	61.5±0.37	70.8±0.65
	$\mathrm{Shuf}_{\mathrm{sent}}$	62.2±1.61	71.6±1.10
	ALL	61.9±1.11	71.4±0.41

Table 6: F1 scores on adversarial test sets. The means±standard deviations over three random seeds are reported.

$\mathrm{Del}_{\mathrm{que}}$, the scores of the models are still higher than the human score.

These relatively high F1 scores imply that models can not recognize the intended features removed by the perturbations as humans do. Moreover, the confident predictions of the models on invalid data may harm the reliability of model predictions in real-world applications.

Entropy Maximization Can Penalize Confident and Correct Predictions for Seen Perturbations Entropy maximization make models sensitive to seen perturbation types, as shown by the diagonals of Tables 3 and 4, except for BERT-base with $\mathrm{Shuf}_{\mathrm{sent}}$. The F1 scores decreased and the entropies increased along the diagonal cells. Moreover, the entropies for the seen perturbations almost reached the maximum value of 11.9 without hurting the F1 scores on the original dev set (None).

Entropy Maximization Fails to Transfer to Unseen Perturbations However, maximizing entropy for a certain perturbation type often cannot transfer to unseen perturbation types. For example, maximizing entropy for $\mathrm{Del}_{\mathrm{que}}$ have little impact on the sensitiveness of models to $\mathrm{Del}_{\mathrm{func}}$. To mitigate the lack of cross-perturbation transferability, we simply maximize the entropy terms for all the perturbation types as in Equation 5, which is denoted as ALL. We show that this approach can successfully make models less confident on all the four perturbations except for BERT-base with $\mathrm{Shuf}_{\mathrm{sent}}$. On the other hand, we observed that there are some perturbation types where entropy maximization can transfer to some extent (e.g., from $\mathrm{Del}_{\mathrm{func}}$ to $\mathrm{Shuf}_{\mathrm{word}}$ and from $\mathrm{Shuf}_{\mathrm{sent}}$ to $\mathrm{Shuf}_{\mathrm{word}}$).

Influence of the Scaling Factor Among the perturbation types, BERT-base failed to become less confident on $\mathrm{Shuf}_{\mathrm{sent}}$. To determine the cause, we examined the influence of scaling factors. First, the chosen scaling factor was 0.01, which was insufficient to increase the entropy. Second, we found that the F1 scores of BERT-base on the clean dev set and the dev set perturbed with $\mathrm{Shuf}_{\mathrm{sent}}$ are strongly correlated. This implies that BERT-base cannot distinguish the original sentence order and the shuffled sentence order inherently. Given that this tendency is not consistent with RoBERTa-base, the next sentence prediction task for pretraining BERT-base may cause this difference.

3.4 Out-of-Distribution Generalization

If models can recognize the intended features described in Table 2 after entropy maximization, it is possible that the way models process language becomes closer to the way humans do, and thereby generalize to OOD (Talmor and Berant 2019) and adversarial (Jia and Liang 2017) test sets better than before.

The F1 scores on datasets from other domains and adversarial test sets are shown in Tables 5 and 6, respectively. Based on these results, entropy maximization for any perturbations did not improve the generalization to other domains nor the adversarial robustness, but rather sometimes degraded them.

As discussed in Hase, Xie, and Bansal (2021), intentionally perturbed inputs can be out-of-distribution for models, and do not naturally appear in a dataset. Therefore, regularizing the predictions on largely perturbed inputs may not have a positive effect on the generalization to natural ex-

amples. Making QA models recognize natural changes in inputs using carefully designed perturbations (e.g., masked language modeling (Devlin et al. 2019)) is future work.

4 Related Work

4.1 Insensitivity to Large Perturbations

Recently, there has been a surge of interest in the insensitivity of NLU models to large perturbations (Sugawara et al. 2018, 2020; Hessel and Schofield 2021). These studies showed that NLU models can correctly predict the output even when the inputs is largely perturbed with some transformations such as word deletion and word order shuffling at test time. To penalize the insensitivity of NLU models, entropy maximization has been used (Feng et al. 2018; Gupta, Kvernadze, and Srikumar 2021; Sinha et al. 2021). However, the effect of entropy maximization for large perturbations has been studied in isolation. Given the hypothesis that different perturbation removes different features from input as shown in Table 2, only regularizing entropy for single type of perturbation may not be enough to make models' predictions more human-like.

Making dialog models recognize dialog history with perturbations has been shown to improve the performance of dialog systems (Zhou, Li, and Li 2021). This work is most relevant to our motivation.

4.2 Sensitivity to Small Perturbations

In contrast, deep learning models can misclassify slightly perturbed inputs (Szegedy et al. 2013; Jia and Liang 2017; Mudrakarta et al. 2018), which are called adversarial examples. To mitigate this issue, adversarial training and its variants have been proven to be effective (Goodfellow, Shlens, and Szegedy 2014; Zhu et al. 2020; Jiang et al. 2020; Liu et al. 2020) in maintaining the model prediction when the input is slightly perturbed, which is the exact opposite of entropy maximization used in our work. Similar to our work, the transfer of adversarial training between different perturbation types was studied in computer vision (Kang et al. 2019; Maini, Wong, and Kolter 2020).

5 Conclusion

We first showed that entropy maximization often fails to transfer to unseen perturbation types. Maximizing the entropy terms for various types of perturbations is effective in mitigating this problem. The failure of entropy maximization to improve out-of-distribution generalization may be caused by the unnaturalness of the perturbed inputs. Modifying the perturbation functions to effectively improve out-of-distribution generalization is future work.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. This work was supported by JST SPRING Grant Number JPMJSP2108 and JSPS KAKENHI Grant Numbers 21H03502, 22J13751 and 22K17954. This work was also supported by NEDO SIP-2 "Big-data and AI-enabled Cyberspace Technologies".

References

Belinkov, Y. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1): 207–219.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–4458. Online: Association for Computational Linguistics.

Dolan, W. B.; and Brockett, C. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Dunn, M.; Sagun, L.; Higgins, M.; Guney, V. U.; Cirik, V.; and Cho, K. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Feng, S.; Wallace, E.; Grissom II, A.; Iyyer, M.; Rodriguez, P.; and Boyd-Graber, J. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3719–3728. Brussels, Belgium: Association for Computational Linguistics.

Fisch, A.; Talmor, A.; Jia, R.; Seo, M.; Choi, E.; and Chen, D. 2019. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 1–13. Hong Kong, China: Association for Computational Linguistics.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gupta, A.; Kvernadze, G.; and Srikumar, V. 2021. BERT & Family Eat Word Salad: Experiments with Text Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14): 12946–12954.

Hase, P.; Xie, H.; and Bansal, M. 2021. Search Methods for Sufficient, Socially-Aligned Feature Importance Explanations with In-Distribution Counterfactuals. *arXiv preprint arXiv:2106.00786*.

Hessel, J.; and Schofield, A. 2021. How effective is BERT without word ordering? Implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 204–211. Online: Association for Computational Linguistics.

Hewitt, J.; and Liang, P. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2733–2743. Hong Kong, China: Association for Computational Linguistics.

Hewitt, J.; and Manning, C. D. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138. Minneapolis, Minnesota: Association for Computational Linguistics.

Jia, R.; and Liang, P. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2021–2031. Copenhagen, Denmark: Association for Computational Linguistics.

Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; and Zhao, T. 2020. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2177–2190. Online: Association for Computational Linguistics.

Joshi, M.; Choi, E.; Weld, D.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1601–1611. Vancouver, Canada: Association for Computational Linguistics.

Kang, D.; Sun, Y.; Brown, T.; Hendrycks, D.; and Steinhardt, J. 2019. Transfer of adversarial robustness between perturbation types. *arXiv preprint arXiv:1905.01034*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; Toutanova, K.; Jones, L.; Kelcey, M.; Chang, M.-W.; Dai, A. M.; Uszkoreit, J.; Le, Q.; and Petrov, S. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7: 452–466.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Li, J.; Monroe, W.; and Jurafsky, D. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Liu, X.; Cheng, H.; He, P.; Chen, W.; Wang, Y.; Poon, H.; and Gao, J. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Maini, P.; Wong, E.; and Kolter, Z. 2020. Adversarial Robustness Against the Union of Multiple Perturbation Models. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 6640–6650. PMLR.

Mudrakarta, P. K.; Taly, A.; Sundararajan, M.; and Dhamdhere, K. 2018. Did the Model Understand the Question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1896–1906. Melbourne, Australia: Association for Computational Linguistics.

Pearl, J. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. Austin, Texas: Association for Computational Linguistics.

Sen, P.; and Saffari, A. 2020. What do Models Learn from Question Answering Datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2429–2438. Online: Association for Computational Linguistics.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2017. Bidirectional Attention Flow for Machine Comprehension. In *International Conference on Learning Representations*.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Sinha, K.; Parthasarathi, P.; Pineau, J.; and Williams, A. 2021. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7329–7346. Online: Association for Computational Linguistics.

Sugawara, S.; Inui, K.; Sekine, S.; and Aizawa, A. 2018. What Makes Reading Comprehension Questions Easier? In

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 4208–4219. Brussels, Belgium: Association for Computational Linguistics.

Sugawara, S.; Stenetorp, P.; Inui, K.; and Aizawa, A. 2020. Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 8918–8927.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Talmor, A.; and Berant, J. 2019. MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4911–4921. Florence, Italy: Association for Computational Linguistics.

Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; McCoy, R. T.; Kim, N.; Durme, B. V.; Bowman, S.; Das, D.; and Pavlick, E. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Trischler, A.; Wang, T.; Yuan, X.; Harris, J.; Sordoni, A.; Bachman, P.; and Suleman, K. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 191–200. Vancouver, Canada: Association for Computational Linguistics.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. New Orleans, Louisiana: Association for Computational Linguistics.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380. Brussels, Belgium: Association for Computational Linguistics.

Yogatama, D.; d'Autume, C. d. M.; Connor, J.; Kocisky, T.; Chrzanowski, M.; Kong, L.; Lazaridou, A.; Ling, W.; Yu, L.; Dyer, C.; et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Zhou, W.; Li, Q.; and Li, C. 2021. Learning from Perturbations: Diverse and Informative Dialogue Generation with Inverse Adversarial Training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 694–703. Online: Association for Computational Linguistics.

Zhu, C.; Cheng, Y.; Gan, Z.; Sun, S.; Goldstein, T.; and Liu, J. 2020. FreeLB: Enhanced Adversarial Training for Natural Language Understanding. In *International Conference on Learning Representations*.