

Improving Machine Reading Comprehension via Adversarial Training

Ziqing Yang[†], Yiming Cui^{†‡}, Wanxiang Che[‡], Ting Liu[‡], Shijin Wang^{†§}, Guoping Hu[†]

[†]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China

[‡]Research Center for Social Computing and Information Retrieval (SCIR),
Harbin Institute of Technology, Harbin, China

[§]iFLYTEK AI Research (Hebei), Langfang, China

^{†§}{zqyang5, ymcui, sjwang3, gphu}@iflytek.com

[‡]{ymcui, car, tliu}@ir.hit.edu.cn

Abstract

Adversarial training (AT) as a regularization method has proved its effectiveness in various tasks, such as image classification and text classification. Though there are successful applications of AT in many tasks of natural language processing (NLP), the mechanism behind it is still unclear. In this paper, we aim to apply AT on machine reading comprehension (MRC) and study its effects from multiple perspectives. We experiment with three different kinds of RC tasks: span-based RC, span-based RC with unanswerable questions and multi-choice RC. The experimental results show that the proposed method can improve the performance significantly and universally on SQuAD1.1, SQuAD2.0 and RACE. With virtual adversarial training (VAT), we explore the possibility of improving the RC models with semi-supervised learning and prove that examples from a different task are also beneficial. We also find that AT helps little in defending against artificial adversarial examples, but AT helps the model to learn better on examples that contain more low-frequency words.

1 Introduction

Neural networks have achieved superior performances in many tasks in the fields of computer vision (CV) and natural language processing (NLP). However, they are not robust to certain perturbations. Szegedy et al. (2014) found that in the image classification task, neural network model predicts different labels for the original (“clean”) example and the perturbed example even the difference between the two is tiny. They call perturbed examples *adversarial examples*. Subsequently, Goodfellow et al. (2015) proposed *adversarial training* (AT) as a regularization method to improve the robustness by training on the mixture of original examples and adversarial examples. Later, in the field of NLP, Miyato et al. (2017) successfully applied ad-

versarial training and *virtual adversarial training* (Miyato et al., 2016) — a semi-supervised learning version of adversarial training — on the text classification task.

Though there are some successful applications of adversarial training in NLP tasks (Wu et al., 2017; Yasunaga et al., 2018; Bekoulis et al., 2018), the mechanism behind adversarial training in the context of NLP is still unclear, and more investigations are required to improve our understanding of adversarial training. To take one step towards this goal, we aim to apply adversarial training on machine reading comprehension (MRC) tasks.

MRC is an important and popular task in NLP. In MRC, a machine is asked to read a passage and then answer the questions posed based on that passage. This task is challenging since it requires sophisticated natural language understanding. Many models have been proposed and achieved superior results (Kadlec et al., 2016; Cui et al., 2017; Seo et al., 2017; Devlin et al., 2018). Some authors also investigated the robustness of RC models (Jia and Liang, 2017; Wang and Bansal, 2018).

In this paper our goal is to improve RC models by incorporating adversarial training and analyze its effects from multiple perspectives. First, to verify the generality of adversarial training, we apply it to three different MRC tasks: span-based RC, span-based RC with unanswerable questions, and multi-choice RC; and conduct experiments on the representative datasets: SQuAD1.1, SQuAD2.0 and RACE. We use BERT as our base model and adapt it to each task with task-specific modifications. The experimental results show that adversarial training consistently boosts the performance across multiple datasets. Second, we explore the possibility of semi-supervised training on RC models with virtual adversarial training, and conclude that model can benefit from training on *cross-task* examples that are from other tasks. Furthermore, we inves-

tigate whether adversarial training improves the robustness of RC models on artificial adversarial examples. Lastly, we analyze how the model performance is improved with AT, and find that adversarial training helps the model to learn better on examples that contain more low-frequency words.

2 Related Work

Reading comprehension. The objective of MRC is to let a machine read given passages and ask it to answer the related questions. In recent years, more and more large-scale RC datasets became available. These datasets focus on different types of RC tasks, such as cloze-style RC (Hermann et al., 2015; Hill et al., 2016), span-based RC with or without unanswerable questions (Rajpurkar et al., 2016, 2018) and multi-choice RC (Lai et al., 2017). Some tasks require the model to answer yes/no questions in addition to spans (Reddy et al., 2019). With the help of the large-scale datasets, the RC models evolve rapidly and even outperform humans on some tasks (Cui et al., 2017; Seo et al., 2017; Xiong et al., 2018; Radford, 2018; Hu et al., 2018). However, this does not imply that machine has acquired real intelligence, as the machine can be fooled easily on artificial examples (Jia and Liang, 2017).

BERT (Devlin et al., 2018) as a model of pre-trained deep bidirectional representations has shown excellent performance and set the new state of the arts on various NLP tasks, including MRC. It has become an indispensable part of modern high-performance RC model. In this work, we use BERT as our base model and adapt it for different RC tasks.

Adversarial Training. Szegedy et al. (2014) first discovered the existence of small perturbations to the input images that mislead models to predict wrong labels in the image classification. They called the perturbed inputs *adversarial examples*. Goodfellow et al. (2015) proposed a simple and fast *adversarial training* method to improve the robustness of the model by training on both clean examples and adversarial examples. In the context of NLP, Miyato et al. (2017) applied adversarial training and virtual adversarial training (Miyato et al., 2016) to text classification task by perturbing the word embeddings of input sentences. Some authors further applied the adversarial training to various NLP tasks, such relation extraction (Wu et al., 2017), part-of-speech tagging (Yasunaga

et al., 2018) and jointly extracting entities and relations (Bekoulis et al., 2018). A recent work (Sato et al., 2019) investigates the effects of AT on neural machine translation. Wang et al. (2018) studied the effects of applying AT to different set of variables in MRC tasks. Sato et al. (2018) focuses on improving the interpretability of adversarial examples in the context of NLP.

Different from the previously discussed idea of embedding level perturbations, Jia and Liang (2017) generated adversarial examples for MRC tasks at the word token level. They introduced the *AddSent* algorithm, which generates adversarial examples by appending distracting sentences to the input passages. These sentences resemble the question and do not contradict the correct answer. They focused on the evaluation of the RC systems on these artificial adversarial examples. They showed that even the state-of-the-art RC system could be easily fooled by these adversarial examples. In this work, we focus on improving the generalization performance of the RC system by training on adversarial examples.

3 Methodology

We first give the formal definitions of the tasks and introduce the corresponding model of each RC tasks in question. Then we describe the adversarial training method. Lastly we present the strategies that are useful or worth discussing in applying the AT method.

3.1 Task Definition

We consider three types of reading comprehension tasks: span-based extractive RC (SE-RC), span-based extractive RC with unanswerable questions (SEU-RC) and multi-choice RC (MC-RC). All of these tasks require the machine to answer questions related to the given passages. We denote the tokenized passage as $P = \{p_1, p_2, \dots, p_m\}$ and the tokenized question as $Q = \{q_1, q_2, \dots, q_n\}$. For simplicity, we use the term *token* and *word* interchangeably in the following.

- **SE-RC.** Given P and Q , the answer is a continuous span extracted from the passage :

$$A = \{p_i, \dots, p_j\}, \text{ where } 1 \leq i \leq j \leq m \quad (1)$$

RC models on this task predict the start position i and the end position j of the answer.

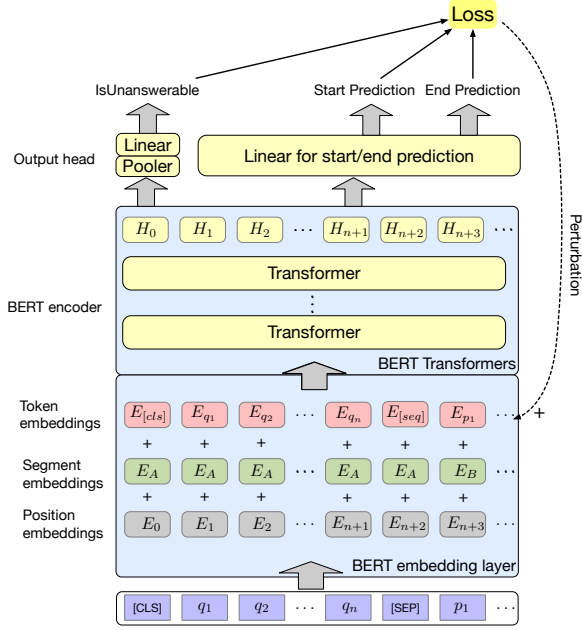


Figure 1: Architecture of the model for SEU-RC. This architecture is also used in the SE-RC by simply ignoring the *IsUnanswerable* prediction head. In the adversarial training phase, The perturbations are added only to the token embeddings of BERT embedding layer.

- **SEU-RC.** Similar to SE-RC, but the passages may contain no answers. The model has to tell if the question is answerable, and if answerable predict the correct answer. Formally, for each P and Q , the answer A is either a valid span as in (1) or empty: $A = \{\}$.
- **MC-RC.** Besides P and Q , additional answer options $\mathcal{O} = \{O^{(1)}, O^{(2)}, \dots, O^{(m)}\}$ are provided, where each option is a sequence of words $O^{(i)} = \{o_1^{(i)}, o_2^{(i)}, \dots\}$. The model is asked to select the correct answer from \mathcal{O} .

3.2 RC Model Architecture

Bidirectional Encoder Representation from Transformers (BERT) has shown great performance and set the new state-of-the-art of various NLP tasks (Devlin et al., 2018). It consists of an embedding layer, followed by multi-layer transformers (Vaswani et al., 2017), and task-specific output heads. For details, the readers may refer to the original paper. We adopt the fine-tuned BERT as our base model and adapt it to the above-listed RC tasks by designing task-specific output heads and loss functions. The model architecture is illustrated in Figure 1.

• SE-RC

Devlin et al. (2018) has shown how to adapt BERT to this task. We briefly recap the input, the outputs and the loss function here. The input is the concatenation of P and Q with special tokens [CLS] and [SEP] as [CLS] Q [SEP] P [SEP]. The outputs are the start/end position probability distributions p_s and p_e . The training objective \mathcal{L}_{span} is the sum of the negative log-likelihood of the correct start and end positions:

$$\mathcal{L}_{span} = \frac{1}{N} \sum_k \mathcal{L}_{span}^{(k)} \quad (2)$$

$$\mathcal{L}_{span}^{(k)} = -y_s^{(k)} \log p_s^{(k)} - y_e^{(k)} \log p_e^{(k)} \quad (3)$$

where the superscript k indicates the k -th example, y_s and y_e are the ground truth start and end positions in one-hot representation, N is the total number of examples.

• SEU-RC

The input representation and the span prediction head are the same as span-based extractive RC. We focus on how to deal with no-answer prediction.

Some previous works introduced a special NoAnswer token and concatenated it to the passage, or use the existed [CLS] as the NoAnswer token in BERT-based model (Sun et al., 2018; Liu et al., 2018). The no-answer prediction could be easily handled by treating NoAnswer as a valid span position. While in this paper, we separate the predictions for no-answer and span since it allows more flexibility and achieves a better performance.

Denote the output from of the last transformer in BERT as $H \in \mathbb{R}^{l \times h}$, where l is the sequence length and h is the hidden dimension. We use the *pooler* of the original BERT to squeeze H into a vector $B \in \mathbb{R}^h$. The no-answer probability is computed as

$$p_{na} = \text{sigmoid}(B \cdot W^{na} + b^{na}) \quad (4)$$

where $W^{na} \in \mathbb{R}^h$, $b^{na} \in \mathbb{R}$. The no-answer loss of the k -th example is

$$\mathcal{L}_{na}^{(k)} = -y_{na}^{(k)} \log p_{na}^{(k)} - (1 - y_{na}^{(k)}) \log(1 - p_{na}^{(k)}) \quad (5)$$

where $y_{na}^{(k)} = 1$ if the k -th example is unanswerable else 0. The total loss is

$$\mathcal{L} = \frac{1}{N} \sum_k (\mathcal{L}_{na}^{(k)} + \mathcal{L}_{span}^{(k)} \cdot y_{na}^{(k)}) \quad (6)$$

In the inference phase, we first find the most-probable valid non-empty span A' and the corresponding span probability $p_{s,i} \cdot p_{e,j}$, and compare the difference between the no-answer probability p_{na} and the total span probability $p_{s,i} \cdot p_{e,j} \cdot (1 - p_{na})^2$ with a threshold (needs to be searched) to judge whether the example is answerable.

- **MC-RC**

Given P, Q and m options $\{O^{(1)}, \dots, O^{(m)}\}$, we construct m input sequences

$$X^{(1)} = [\text{CLS}] P [\text{SEP}] Q [\text{SEP}] O^{(1)} [\text{SEP}]$$

$$\dots$$

$$X^{(m)} = [\text{CLS}] P [\text{SEP}] Q [\text{SEP}] O^{(m)} [\text{SEP}]$$

and add different segmentation embedding before and after (including) Q . We feed the m sequences into BERT and collect the outputs from BERT pooler:

$$\tilde{H} = \{B^{(1)}, \dots, B^{(m)}\} \in \mathbb{R}^{m \times h} \quad (7)$$

The final prediction is obtained by applying a linear transformation followed by softmax over the m options of \tilde{H} :

$$p_o = \text{softmax}(\tilde{H} \cdot W^o + b^o) \in \mathbb{R}^m \quad (8)$$

where $W^o \in \mathbb{R}^{h \times 1}$, $b^o \in \mathbb{R}$. The loss function is the crossentropy loss.

3.3 Adversarial Training Method

Adversarial training (AT) (Goodfellow et al., 2015) as a regularization method improves not only the robustness of the classifier against the perturbations but also the performance on clean inputs. In AT, we first construct adversarial examples by generating worst-case perturbations that maximize the current loss function, then train the model on both of clean examples and adversarial examples.

- **Adversarial Training**

In the context of MRC tasks, the inputs are sequences of words. Following (Miyato et al., 2017), we define the perturbation at the level of word embeddings. Let θ be the trainable parameters of the model. We denote the word embedding vectors of the input sequence X as

$$\mathbf{x} = [\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{l-1}] \in \mathbb{R}^{l \times h} \quad (9)$$

In our model, \mathbf{x} is the token embeddings of BERT's embedding layer (see Figure 1). Let y denote the target. In SE-RC, $y = (y_s, y_e)$; in SEU-RC,

$y = (y_s, y_e, y_{na})$; in MC-RC, y is a single value representing the correct option. The worst-case perturbation \mathbf{r}_{AT} is the one that maximizes the loss with a bounded norm

$$\mathbf{r}_{\text{AT}} = \arg \max_{\mathbf{r}; \|\mathbf{r}\| < \epsilon} \mathcal{L}(\mathbf{x} + \mathbf{r}; y; \hat{\theta}) \quad (10)$$

where $\hat{\theta}$ means treating θ as constant. However, the exact value of \mathbf{r}_{AT} is intractable. We resort to approximating \mathbf{r}_{AT} by linearizing $\mathcal{L}(\mathbf{x} + \mathbf{r}; y; \hat{\theta})$ around \mathbf{x} (Goodfellow et al., 2015) :

$$\mathbf{r}_{\text{AT}} = \epsilon \frac{\mathbf{g}}{\|\mathbf{g}\|}, \text{ where } \mathbf{g} = \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}; y; \hat{\theta}) \quad (11)$$

The adversarial example is constructed as ¹

$$\mathbf{x}_{\text{AT}} = \mathbf{x} + \mathbf{r}_{\text{AT}} = (\hat{\mathbf{x}} + \epsilon \frac{\hat{\mathbf{r}}_{\text{AT}}}{\|\hat{\mathbf{x}}\|}) \|\mathbf{x}\| \quad (12)$$

where $\hat{\mathbf{v}}$ denote the unit vector in the direction of \mathbf{v} . To be more clear, we redefine ϵ to be $\epsilon / \|\mathbf{x}\|$:

$$\mathbf{x}_{\text{AT}} = (\hat{\mathbf{x}} + \epsilon \hat{\mathbf{r}}_{\text{AT}}) \|\mathbf{x}\| \quad (13)$$

The hyper-parameter ϵ under the new definition controls the relative strength of perturbations.

On each batch of inputs, the model is trained on both clean examples and adversarial examples by minimize the original loss \mathcal{L} and the adversarial loss $\mathcal{L}_{\text{AT}}(\mathbf{x}; y; \theta) = \mathcal{L}(\mathbf{x}_{\text{adv}}; y; \theta)$ simultaneously:

$$\hat{\mathcal{L}} = \mathcal{L}(\mathbf{x}; y; \theta) + \mathcal{L}_{\text{AT}}(\mathbf{x}; y; \theta) \quad (14)$$

- **Virtual Adversarial Training**

Virtual adversarial training (VAT) (Miyato et al., 2016) extends AT to semi-supervised training and unlabeled examples. VAT constructs adversarial examples by finding the perturbations that most significantly disturb the predicted distributions:

$$\mathbf{r}_{\text{VAT}} = \arg \max_{\mathbf{r}; \|\mathbf{r}\| < \epsilon} \mathcal{L}_{\text{KL}}(\mathbf{x}, \mathbf{x} + \mathbf{r}, \theta) \quad (15)$$

$$\mathcal{L}_{\text{KL}}(\mathbf{x}, \mathbf{y}, \theta) = \text{KL}[p(\cdot | \mathbf{x}; \hat{\theta}) || p(\cdot | \mathbf{y}; \theta)] \quad (16)$$

where $\text{KL}(\cdot || \cdot)$ is the KL divergence and $p(\cdot | \mathbf{x}; \hat{\theta})$ is the predicted distribution. The exact value of \mathbf{r}_{VAT} is also intractable. An approximated solution is (Miyato et al., 2016)

$$\mathbf{r}_{\text{VAT}} = \epsilon \frac{\mathbf{g}}{\|\mathbf{g}\|}, \mathbf{g} = \nabla_{\xi d} \mathcal{L}_{\text{KL}}(\mathbf{x}, \mathbf{x} + \xi d, \hat{\theta}) \quad (17)$$

¹We do not normalize word embeddings as in (Miyato et al., 2017) and (Yasunaga et al., 2018), since BERT already has a LayerNorm in its embedding layer.

where d is a unit-norm random vector and ξ is a small positive number. With the redefinition of ϵ as above, the loss of VAT is

$$\mathcal{L}_{\text{VAT}}(\mathbf{x}, \theta) = \mathcal{L}_{\text{KL}}(\mathbf{x}, \mathbf{x}_{\text{VAT}}, \theta) \quad (18)$$

$$\mathbf{x}_{\text{VAT}} = (\hat{\mathbf{x}} + \epsilon \hat{\mathbf{r}}_{\text{VAT}}) \|\mathbf{x}\| \quad (19)$$

VAT uses no target y but only \mathbf{x} , which makes it possible to be applied on unlabeled data and perform semi-supervised training.

3.4 Applying AT on MRC

We propose the following three strategies to be used in combination with AT method in MRC tasks. We found they are either helpful, or worth discussing.

- **Negative entropy loss (NEL).** In SEU-RC, the span-prediction head is only trained on the answerable questions. On the unanswerable questions, we expect no spikes in the predicted span probability, which means model are less likely to make mistake. To punish any high-probability span predictions on unanswerable questions, we construct the following negative entropy loss (NEL):

$$\mathcal{L}_{ne} = \frac{1}{N} \sum_k (\mathcal{L}_{ne}^{(k)} \cdot y_{na}^{(k)}) \quad (20)$$

$$\mathcal{L}_{ne}^{(k)} = \sum_i p_{s,i}^{(k)} \log p_{s,i}^{(k)} + p_{e,i}^{(k)} \log p_{e,i}^{(k)} \quad (21)$$

where i is summed over all valid start/end positions. The intuition this is that the uniform distribution has the highest entropy. In practice we find training \mathcal{L}_{ne} only on the adversarial examples gives the best performance.

- **Semi-supervised training.** VAT can be applied on both labeled examples and unlabeled examples. However, there are no unlabeled examples in the datasets we study. We propose to treat answerable questions as labeled examples and unanswerable questions as unlabeled examples. The details will be discussed in Section 5.1.
- **Data augmentation (DA).** We perform data augmentation to generate unanswerable questions. As will be shown in Section 5.2, DA has a great influence on SQuAD2.0. The details of DA are described in Appendix A.

4 Experiments

4.1 Datasets

We evaluate our method on the representative datasets SQuAD1.1 (Rajpurkar et al., 2016), SQuAD2.0 (Rajpurkar et al., 2018) and RACE (Lai et al., 2017).

The passages in SQuAD1.1 are retrieved from Wikipedia articles and the questions are crafted by crowd-workers. The answer to each question is a span in the passage. The sizes of its training, development and test set are roughly 88k/11k/10k.

SQuAD2.0 contains unanswerable questions. About one-third of the examples in the training set, and half of the examples in the development and test set are unanswerable. The sizes of its training, development and test set are roughly 130k/12k/9k.

RACE is a multi-choice RC dataset, which is collected from the English exams for middle and high school Chinese students. RACE-M denotes the middle school exams and RACE-H denotes high school exams. Each example in RACE contains four answer options, and only one of them is correct. The sizes of its training, development and test set are roughly 88k/5k/5k.

4.2 Experimental Setup

- **BERT.** We initialize BERT with the pre-trained weights released by Google². For experiments on SQuAD, we use the *cased* pre-trained weights; for experiments on RACE, we use the *uncased* pre-trained weights.
- **Hyper-parameters.** We set the batch size to 24, learning rate to $5e-5$ for BERT_{BASE} and $3e-5$ for BERT_{LARGE}. The maximum sequence length is set to 416 for SQuAD and 512 for RACE. The number of training epochs is 3 for SQuAD and 5 for RACE. We keep the other hyper-parameters of BERT default. For adversarial training, the hyper-parameter ϵ is set to $1e-2$ for SQuAD and $1e-3$ for RACE. We have found that the optimal value of ϵ of each dataset is rather stable and performs well almost in all experiments. For semi-supervised learning, batch size is 12 for unlabeled samples, and $\xi = 1e-5$.
- **Evaluation.** The test sets of SQuAD1.1 and SQuAD2.0 are hidden. Thus we report the results on development sets, except the model we submitted to the official for online evaluation.

²<https://github.com/google-research/bert>

System	Dev		Test	
	EM	F1	EM	F1
<i>Human Performance</i>	80.3	90.5	82.3	91.2
<i>Ensemble model</i>				
nlnet [†]	-	-	85.4	91.2
BERT (Devlin et al., 2018)	86.2	92.2	87.4	93.2
<i>Single model</i>				
BERT (Devlin et al., 2018)	84.2	91.1	85.1	91.8
KT-NET(Yang et al., 2019a)	-	-	85.9	92.4
BERT+WWM+MT [†]	-	-	88.7	94.4
spanBERT (Joshi et al., 2019)	-	-	88.8	94.6
XLNet (Yang et al., 2019b)	89.0	94.5	89.9	95.1
RoBERTa (Liu et al., 2019)	89.4	94.6	-	-
<i>Our single model</i>				
BERT(ours)	84.9	91.4	-	-
BERT+AT+VAT(2.0)	86.1	92.4	86.9	92.6

Table 1: Results on SQuAD1.1 dev/test set. Best results are in boldface. [†] indicates unpublished works. BERT(ours) is our reimplement of BERT for SQuAD. VAT(2.0) refers to virtual adversarial training with SQuAD2.0 data. See 5.1 for details.

4.3 Overall Results

We have observed universal improvements across all three tasks, which prove the generality of adversarial training. We first show the overall results. The analysis is provided later in Section 5. Notice that the current state-of-the-art models (spanBERT, XLNet, RoBERTa, etc.) use different base models from BERT, which have been pre-trained from scratch on large corpora.

SQuAD1.1. We submitted our best single model on the development set for evaluation. The overall results are shown in Table 1. Our best model BERT+AT+VAT(2.0) archives an EM/F1 score of 86.9/92.6. Compared to our BERT baseline, our model improves 1.2/1.0 on EM/F1 with p -value <0.01 , which means the improvement relative to BERT is significant. Compared to the other results on the leaderboard, BERT+AT+VAT(2.0) is the best one among the BERT-based models that use weights (no whole word masking) released by Google.

SQuAD2.0. The best model on the development set is submitted for evaluation, and the results are shown in Table 2. With the help of NEL and AT, our best model BERT+DA+NEL+AT archives 82.9/86.0 on EM/ F1. On the development set, our best model outperforms our baseline BERT+DA by 1.3/1.4 on EM/F1 respectively with p -value <0.01 .

RACE. Finally, we show the results on the test set of RACE, see Table 3. AT improves the overall

System	Dev		Test	
	EM	F1	EM	F1
<i>Human Performance</i>	86.3	89.0	86.9	89.5
<i>Single model</i>				
BERT (Devlin et al., 2018)	78.7	81.9	80.0	83.1
PAML+BERT [†]	-	-	82.6	85.6
BERT+DAE+AoA [†]	-	-	85.9	88.6
RoBERTa (Liu et al., 2019)	86.5	89.4	86.8	89.8
UPM [†]	-	-	87.2	89.9
XLNet + SV (Zhang et al., 2019b)	-	-	87.2	90.1
<i>Our single model</i>				
BERT + DA	81.5	84.4	-	-
BERT + DA + NEL + AT	82.8	85.8	82.9	86.0

Table 2: Results on SQuAD2.0 dev/test set. Best single model results are in boldface. [†] indicates unpublished works. NEL refers to negative entropy loss. SV refers to SG-Net Verifier++.

accuracy from 66.4% to 68.3% on BERT_{BASE} and from 70.5% to 72.4% on BERT_{LARGE}. AT method achieved significant improvements without sophisticated architecture design.

System	RACE	RACE-M	RACE-H
Amazon Mechanical Turker	73.3	85.1	69.4
<i>Single model</i>			
GPT (Radford, 2018)	59.0	62.9	57.4
OCN (Ran et al., 2019)	71.7	76.7	69.6
DCMN (Zhang et al., 2019a)	72.3	77.6	70.1
BERT + DCMN+ (Zhang et al., 2019a)	75.8	79.3	74.4
XLNet (Yang et al., 2019b)	81.8	85.5	80.2
RoBERTa (Liu et al., 2019)	83.2	86.5	81.8
<i>Our single model</i>			
BERT-base(ours)	66.4	73.5	63.5
BERT-base + AT	68.3	73.8	66.1
BERT-large(ours)	70.5	75.4	68.5
BERT-large + AT	72.4	77.0	70.5

Table 3: Accuracy(%) on the test set of RACE. We also list other competing single models on the leaderboard.

5 Analysis

5.1 Is Semi-supervised Learning Helpful?

There are limited studies on semi-supervised learning on RC tasks (Yang et al., 2017; Dhingra et al., 2018). In this section, we explore this possibility with virtual adversarial training. We conduct the experiments on SQuAD1.1, which only contains answerable questions, and treat the unanswerable questions from SQuAD2.0 training set as the unlabeled examples. We perform experiments with different configurations as shown in Table 4.

Training with AT or VAT solely on SQuAD1.1 training set results in similar improvements no matter on BERT_{BASE} or BERT_{LARGE}. Adding unanswerable questions as unlabeled examples improves performance slightly on BERT_{LARGE} (the

System	Large		Base	
	EM	F1	EM	F1
BERT	84.9	91.4	81.2	88.7
BERT+AT	86.0	92.2	83.6	90.2
BERT+VAT	86.1	92.2	83.5	90.2
BERT+AT+VAT(2.0)	86.1	92.4	83.5	90.2
BERT+VAT(2.0)	85.3	91.9	82.8	89.9

Table 4: Results of different configurations of AT and VAT on SQuAD1.1 development set. +AT and +VAT refers to apply AT or VAT on SQuAD1.1 training set. +VAT(2.0) refers to apply VAT on unlabeled examples from SQuAD2.0 training set.

fourth line in Table 4). So far, we see no significant benefits of training on unlabeled examples with labeled samples. We suppose that in order to further improve the performance with semi-supervised learning, more unlabeled examples are needed since typical semi-supervised learning datasets usually contains far more unlabeled examples than labeled examples (Miyato et al., 2017). However, if we only perform VAT on unlabeled examples which is denoted as BERT+VAT(2.0) in the table, we obtain improvements of 0.4/0.5 and 1.6/1.2 on EM/F1 relative to the baseline on BERT_{LARGE} and BERT_{BASE} respectively. Notice that unanswerable questions are out of the domain of the SQuAD1.1 task, and the models in these experiments are not designed for handling unanswerable questions, but with semi-supervised learning they benefit from these questions. The results prove that even cross-task data could help improving RC models.

5.2 Ablation Study

We do an ablation study to test the effectiveness of different components in our best model BERT + DA + NEL + AT for SEU-RC on SQuAD2.0. We run each experiment three times and report the best performance. To further corroborate the results, we run ablation experiments on both BERT_{LARGE} and BERT_{BASE}. The ablation results are shown in Table 5.

Data augmentation has a critical influence on the performance, as we expected. Adversarial training boosts the performance in any configuration, no matter on BERT_{LARGE} or BERT_{BASE}, with or without data augmentation, which means that adversarial training and data augmentation are two orthogonal methods. Recall that the effective number of training examples are doubled in AT as we

System	Large		Base	
	EM	F1	EM	F1
BERT+ DA + NEL + AT	82.8	85.8	78.8	81.6
-NEL	82.4	85.4	78.6	81.3
-AT, -NEL	81.5	84.4	77.2	80.1
-DA	81.0	83.9	76.5	79.1
-NEL	80.8	83.7	76.2	78.9
-AT, -NEL	78.6	81.9	74.2	76.9

Table 5: Ablation study on SQuAD2.0 development set.

System	AddSent (F1)	AddOneSent (F1)	Test (F1)	Δ_1	Δ_2
<i>Single model</i>					
R.M-Reader (Hu et al., 2018)	58.5	67.0	86.6	28.1	19.6
KAR (Wang and Jiang, 2018)	60.1	72.3	83.5	23.4	11.2
<i>Our single model</i>					
BERT	61.0	71.1	91.4	30.4	20.3
BERT+AT+VAT(2.0)	63.5	72.5	92.4	28.9	19.9
Absolute improvement(%)	2.5	1.4	1.0		
Relative improvement(%)	6.4	4.8	13.2		

Table 6: Model performance on AddSent and AddOneSent. Results on SQuAD1.1 are also provided for comparison. Δ_1 is the difference between Test(F1) and AddSent(F1); Δ_2 is the difference between Test(F1) and AddOneSent(F1).

generate adversarial examples for each input examples, thus AT can be view as a kind of data augmentation in some sense. But we see here neither the artificial data augmentation nor the automatic adversarial examples fully exploit the potential of the model by itself. Model benefits from both of them. With negative entropy loss, the performance is further improved. Though the improvement brought by NEL is not so large as AT and DA, it is stable across different configurations.

5.3 Robustness on artificial adversarial examples

Jia and Liang (2017) constructed two artificial adversarial examples datasets called *AddSent* and *AddOneSent* based on SQuAD1.1 by appending distracting sentences to the passages. Models may be easily fooled on these adversarial examples and predict wrong answers from the distracting sentences because of the high overlap between the distracting sentences and the questions. Although the generation process of these artificial adversarial examples is different from the gradient-based method used in AT, and human annotations are needed during the generation, it is interesting to study how the AT affects the robustness of the model on these human-knowledge-injected adversarial examples.

The results are shown in Table 6. All models are trained on SQuAD1.1 before evaluation. Though the BERT+AT+VAT(2.0) achieves the best results on AddSent and AddOneSent, this is largely due to its high performance on the normal dataset, rather than obtaining additional robustness against AddSent and AddOneSent, since the relative improvements are mediocre. While KAR explicitly utilizes external general knowledge (WordNet), and it has the smallest gap Δ_1 and Δ_2 between F1 on test and F1 on AddSent/AddOneSent. The results show that while AT improves the generalization performance, it is not designed for defending against adversarial examples generated with human knowledge, at least in the reading comprehension tasks. How to bridge the gap between gradient-based and artificial adversarial examples, and how to achieve improvement and robustness on artificial adversarial examples at the same time is still an open question.

5.4 How Does AT Help the Model Learn Better?

AT perturbs the input directly on the embedding vectors. This operation may help to refine the word embeddings, especially the embeddings of low-frequency words (“rare words”) since they are less trained and likely to be under-fitting. The target task may benefit from this refining. We test this hypothesis by studying the performance of the model on different groups of the examples with different number of rare words. We sort all the words by their frequencies of occurrence in the training set and refer the last 10,000 words as rare words. We define the *difficulty*³ of each example as the number of rare words in its passage and question normalized by its total number of words. We categorize all the examples in the development set by their difficulty into several buckets and study the performance on each bucket.

We perform the analysis on SQuAD2.0 dataset for its variety, and train three BERT_{LARGE} baseline models and three BERT_{LARGE} with AT. For each group of models, we average their scores to improve stability. The results on each bucket are shown in Table 7. We plot the relative improvements on each bucket in Figure 2. AT achieves larger improvements on more difficult examples, and the largest improvement is on the examples

³This name is just for simplicity, not necessarily related to the true difficulty of the example. To gain some intuition on the difficulty, we show some examples in Appendix B.

Difficulty range	0~0.01	0.01~0.02	0.02~0.03	0.03~0.05	>0.05
# of total examples	2676	2476	2520	2613	1588
BERT+DA	85.4	85.0	84.1	83.9	82.4
BERT+DA+AT	86.3	86.1	85.6	85.2	84.6
# of HA examples	1356	1237	1271	1283	781
BERT+DA	82.5	80.5	80.8	80.8	82.5
BERT+DA+NEL+AT	83.8	81.6	82.5	81.6	84.4
# of NA examples	1320	1239	1249	1330	807
BERT+DA	88.4	89.5	87.4	86.9	82.3
BERT+DA+AT	88.9	90.7	88.7	88.7	84.9

Table 7: Statistics and performance (F1) of each bucket. HA stands for answerable questions; NA stands for unanswerable questions.

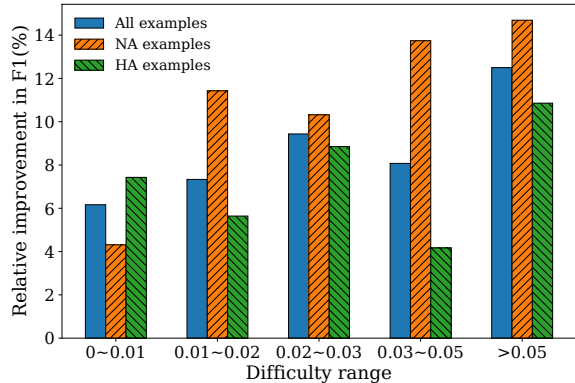


Figure 2: Relative improvement of adversarial training to the baseline model on different bucket.

with difficulty > 0.05. The increase in the relative improvement is more prominent on unanswerable (no-answer, NA) examples than answerable (has-answer, HA) examples. The reason may be that to judge whether a question is unanswerable requires the model to investigate each word in the passage so to make sure that it does not miss any important information, while a span prediction could be made by simply focusing on the context of the matching words, which means predictions on NA examples are more sensitive to the existence of rare words.

6 Conclusion

In this work, we applied adversarial training on MRC tasks and inspect the effects from multiple perspectives. We found that AT improves the performance significantly and consistently across different RC tasks. By the virtue of VAT, we performed semi-supervised learning on MRC tasks. The results show that under semi-supervised learning, the model that is not able to tackle unanswerable questions can benefit from training on unanswerable questions. This inspires us to further explore the possibility of semi-supervised learning on RC in the future. We also found that AT cannot defend against artificial adversarial examples.

Lastly, by a careful analysis of the effect of adversarial training on different sets of examples, we found that AT helps the model to learn better on the examples with more rare words.

References

- Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. [Adversarial training for multi-context joint entity and relation extraction](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836, Brussels, Belgium. Association for Computational Linguistics.
- Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. [Attention-over-attention neural networks for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–602, Vancouver, Canada. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. 2018. [Simple and effective semi-supervised question answering](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 582–587, New Orleans, Louisiana. Association for Computational Linguistics.
- Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. [Explaining and harnessing adversarial examples](#). In *International Conference on Learning Representations*.
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. [Teaching machines to read and comprehend](#). In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.
- Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. [The goldilocks principle: Reading children’s books with explicit memory representations](#). In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. [Reinforced mnemonic reader for machine reading comprehension](#). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4099–4106.
- Robin Jia and Percy Liang. 2017. [Adversarial examples for evaluating reading comprehension systems](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. [Spanbert: Improving pre-training by representing and predicting spans](#). *CoRR*, abs/1907.10529.
- Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. [Text understanding with the attention sum reader network](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 908–918, Berlin, Germany. Association for Computational Linguistics.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale ReAding comprehension dataset from examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- Xiaodong Liu, Wei Li, Yuwei Fang, Aerin Kim, Kevin Duh, and Jianfeng Gao. 2018. [Stochastic answer networks for squad 2.0](#). *CoRR*, abs/1809.09194.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. [Adversarial training methods for semi-supervised text classification](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.
- Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. 2016. [Distributional smoothing by virtual adversarial examples](#). In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- Alec Radford. 2018. Improving language understanding by generative pre-training.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Qiu Ran, Peng Li, Weiwei Hu, and Jie Zhou. 2019. [Option comparison network for multiple-choice reading comprehension](#). *CoRR*, abs/1903.03033.
- Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. [Coqa: A conversational question answering challenge](#). *TACL*, 7:249–266.
- Motoki Sato, Jun Suzuki, and Shun Kiyono. 2019. [Effective adversarial regularization for neural machine translation](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 204–210.
- Motoki Sato, Jun Suzuki, Hiroyuki Shindo, and Yuji Matsumoto. 2018. [Interpretable adversarial perturbation in input embedding space for text](#). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4323–4330.
- Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. [Bidirectional attention flow for machine comprehension](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.
- Fu Sun, Linyang Li, Xipeng Qiu, and Yang Liu. 2018. [U-net: Machine reading comprehension with unanswerable questions](#). *CoRR*, abs/1810.06638.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. [Intriguing properties of neural networks](#). In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Chao Wang and Hui Jiang. 2018. [Exploring machine reading comprehension with explicit knowledge](#). *CoRR*, abs/1809.03449.
- Jiuniu Wang, Xingyu Fu, Guangluan Xu, Yirong Wu, Ziyang Chen, Yang Wei, and Lanyi Jin. 2018. [A3net: Adversarial-and-attention network for machine reading comprehension](#). In *NLPCC*.
- Yicheng Wang and Mohit Bansal. 2018. [Robust machine comprehension models via adversarial training](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 575–581, New Orleans, Louisiana. Association for Computational Linguistics.
- Yi Wu, David Bamman, and Stuart Russell. 2017. [Adversarial training for relation extraction](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783, Copenhagen, Denmark. Association for Computational Linguistics.
- Caiming Xiong, Victor Zhong, and Richard Socher. 2018. [DCN+: mixed objective and deep residual coattention for question answering](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.
- An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019a. [Enhancing pre-trained language representations with rich knowledge for machine reading comprehension](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357, Florence, Italy. Association for Computational Linguistics.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019b. [Xlnet: Generalized autoregressive pretraining for language understanding](#). *CoRR*, abs/1906.08237.
- Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. 2017. [Semi-supervised QA with generative domain-adaptive nets](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1040–1050, Vancouver, Canada. Association for Computational Linguistics.
- Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2018. [Robust multilingual part-of-speech tagging via adversarial training](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 976–986, New Orleans, Louisiana. Association for Computational Linguistics.
- Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019a. [Dual co-matching network for multi-choice reading comprehension](#). *CoRR*, abs/1901.09381.
- Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, and Hai Zhao. 2019b. [Sg-net: Syntax-guided machine reading comprehension](#). *CoRR*, abs/1908.05147.

A Appendices

B Data Augmentation Strategies

We propose two simple strategies to generate unanswerable examples from the SQuAD2.0 training set. We denote the answerable example as (p, q, a) , where p is the passage, q is the question and a is the answer.

B.1 Question-Passage Shuffle

The first strategy replaces the passage p in (p, q, a) with another passage p' that does not contain the answer text. Let \mathcal{P} denotes all the passages that are from the same article as p . We compute the BM25 similarity score between q and each passage in \mathcal{P} . select the highest-score passage p' that does not contain the answer text. We pair p' with q to generate unanswerable example (p', q) .

B.2 Entity Replacement

The second strategy generates unanswerable questions by replacing entities in the questions. Given a passage p , we denote the set of the named entities in p as $\mathcal{T} = \{e_1, e_2, \dots, e_k\}$, the sets of answerable and unanswerable questions related to p as \mathcal{Q}_a and \mathcal{Q}_{na} respectively. For each q in \mathcal{Q}_a , if it contains any entity in \mathcal{T} , we generating a new question q' by replacing that entity with another randomly chosen entity in \mathcal{T} that has the same entity type and does not appear in any question in \mathcal{Q}_{na} . The generated unanswerable example is (p, q') .

These two strategies generate about 70k unanswerable examples in total. We randomly choose 4k examples from question-passage shuffle and 4k examples from entity replacement as our data augmentation set. Though the dataset is small, it is quite effective as shown in the experiments. We did not observe any significant improvements by enlarging the data augmentation set.

C Examples with Different Difficulties

To gain some intuition on the difficulty, we show some examples with different difficulties in Table 8.

Difficulty: 0.0

Passage: After Malaysia ' s independence in 1957 , the government instructed all schools to surrender their properties and be assimilated into the National School system. This caused an uproar among the Chinese and a compromise was achieved in that the schools would instead become “ National Type ” schools . Under such a system , the government is only in charge of the school curriculum and teaching personnel while the lands still belonged to the schools . While Chinese primary schools were allowed to retain Chinese as the medium of instruction , Chinese secondary schools are required to change into English - medium schools . Over 60 schools converted to become National Type schools .

Question:What language is used in Chinese primary schools in Malaysia ?

Difficulty: 0.017

Passage: In the triple form , O₂ molecules are paramagnetic . That is , they impart magnetic character to oxygen when it is in the presence of a magnetic field , because of the spin magnetic moments of the unpaired electrons in the molecule , and the negative exchange energy between neighboring O₂ molecules . Liquid oxygen is attracted to a magnet to a sufficient extent that , in laboratory demonstrations , a bridge of liquid oxygen may be supported against its own weight between the poles of a powerful magnet . [c]

Question:What kind of field is necessary to produce a magnet effect in oxygen molecules ?

Difficulty: 0.024

Passage: According to International Monetary Fund economists , inequality in wealth and income is negatively correlated with the duration of economic growth spells (not the rate of growth) . High levels of inequality prevent not just economic prosperity, but also the quality of a country ' s institutions and high levels of education. According to IMF staff economists , if the income share of the top 20 percent (the rich) increases, then GDP growth actually declines over the medium term , suggesting that the benefits do not trickle down . In contrast, an increase in the income share of the bottom 20 percent (the poor) is associated with higher GDP growth . The poor and the middle class matter the most for growth via a number of interrelated economic , social , and political channels .

Question:What is negatively correlated to the duration of economic growth ?

Difficulty: 0.041

Passage: The neighborhood features restaurants , live theater and nightclub s , as well as several independent shops and bookstore s , currently operating on or near Olive Avenue , and all within a few hundred feet of each other . Since renewal , the Tower District has become an attractive area for restaurant and other local businesses . Today , the Tower District is also known as the center of Fresno ' s LGBT and hipster communities . ; Additionally , Tower District is also known as the center of Fresno ' s local punk / goth / death rock and heavy metal community . [citation needed]

Question:What was Tower District known for before the renewal ?

Difficulty: 0.061

Passage: It has been argued that the term “ civil disobedience ” has always suffered from ambiguity and in modern times , become utterly debased . Marshall Cohen notes , “ It has been used to describe everything from bringing a test - case in the federal courts to taking aim at a federal official . Indeed , for Vice President Al Gore it has become a code - word describing the activities of mugger s , arsonists , draft evaders , campaign hecklers , campus militants , anti-war demonstrators , juvenile delinquents and political assassins . ”

Question:Vice President Al Gore describes Civil disobedience in what activities?

Table 8: Examples with different difficulties from SQuAD2.0 development set. We show the passages and questions after tokenization. Rare words (tokens) are shown in **bold**.