Attention-Guided Answer Distillation for Machine Reading Comprehension

Minghao Hu[†][*], **Yuxing Peng**[†], **Furu Wei**[§],
Zhen Huang[†], **Dongsheng Li**[†], **Nan Yang**[§], **Ming Zhou**[§]
[†] College of Computer, National University of Defense Technology
[§] Microsoft Research Asia
{huminghao09,pengyuxing,huangzhen,dsli}@nudt.edu.cn
{fuwei,nanya,mingzhou}@microsoft.com

Abstract

Despite that current reading comprehension systems have achieved significant advancements, their promising performances are often obtained at the cost of making an ensemble of numerous models. Besides, existing approaches are also vulnerable to adversarial attacks. This paper tackles these problems by leveraging *knowledge distillation*, which aims to transfer knowledge from an ensemble model to a single model. We first demonstrate that vanilla knowledge distillation applied to answer span prediction is effective for reading comprehension systems. We then propose two novel approaches that not only penalize the prediction on confusing answers but also guide the training with alignment information distilled from the ensemble. Experiments show that our best student model has only a slight drop of 0.4% F1 on the SQuAD test set compared to the ensemble teacher, while running 12× faster during inference. It even outperforms the teacher on adversarial SQuAD datasets and NarrativeQA benchmark.

1 Introduction

Machine reading comprehension (MRC), which aims to answer questions about a given passage or document, is a long-term goal of natural language processing. Recent years have witnessed rapid progress from early cloze-style test (Hermann et al., 2015; Hill et al., 2016) to latest answer extraction test (Rajpurkar et al., 2016; Joshi et al., 2017). Several end-to-end neural networks based approaches even outperform the human performance in terms of exact match accuracy on the SQuAD dataset (Wang et al., 2017, 2018; Yu et al., 2018).

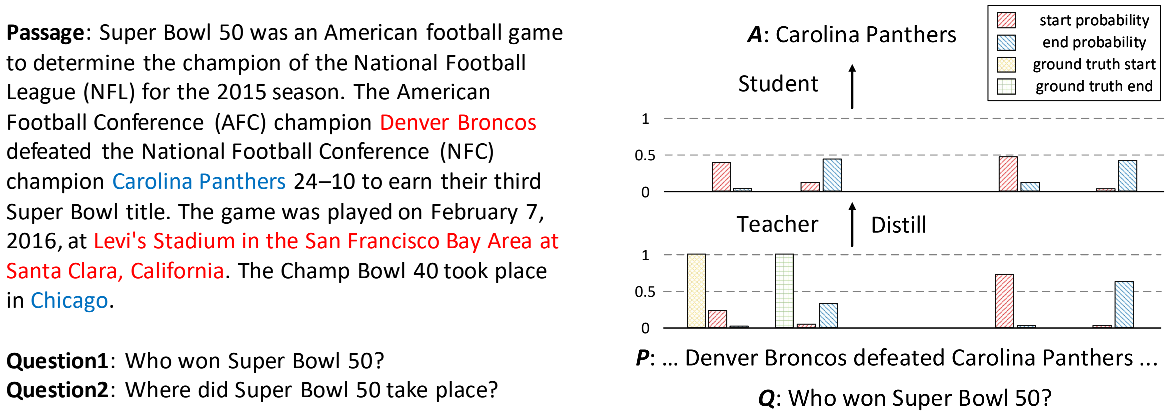
Despite of the advancements, there are still two problems that impedes the deployment of real-world MRC applications. First, although effective, current approaches are *not efficient* because superior performances are usually achieved by ensembling multiple trained models. For example, Seo et al. (2017) submit an ensemble model consisting of 12 training runs and Huang et al. (2018) boost the result with 31 models. The ensemble system, however, has two major drawbacks: the inference time is slow and a huge amount of resource is needed. Second, existing models are *not robust* since they are vulnerable to adversarial attacks. Jia and Liang (2017) show that the models are easily fooled by appending an adversarial sentence into the passage. Such fragility on adversarial examples severely diminishes the practicality of current MRC systems.

One promising direction to address these problems is model compression (Bucilu et al., 2006), which attempts to compress the ensemble model into a single model that has comparable performances. Particularly, the *knowledge distillation* approach (Hinton et al., 2014) has been proposed to train a student model with the supervision of a teacher model. Such idea is further explored to enhance the generalizability and robustness of image recognition systems (Papernot et al., 2016). Some subsequent works attempt to transfer teacher's intermediate representation, such as feature map (Romero et al., 2015; Zagoruyko and Komodakis, 2017), neuron selectivity (Huang and Wang, 2017) and so on, to provide additional supervisions.

In this paper, we present the first work to investigate knowledge distillation in the context of MRC, to improve efficiency and robustness simultaneously. We first apply the standard knowledge distillation to MRC models, by mimicking output distributions of answer boundaries from an ensemble model, and observe consistent improvements upon a strong baseline. We then propose two novel

[*] Contribution during internship at Microsoft Research Asia.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2077–2086
Brussels, Belgium, October 31 - November 4, 2018.



Passage: Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. 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. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. The Champ Bowl 40 took place in Chicago.

Question1: Who won Super Bowl 50?
Question2: Where did Super Bowl 50 take place?

(a) Gold answers (red) versus confusing answers (blue) in the SQuAD and adversarial SQuAD datasets.

(b) The knowledge biased towards the confusing answer is distilled from the teacher to the student.

Figure 1: An illustration of confusing answer and biased distillation in machine reading comprehension.

distillation approaches to further transfer knowledge between the teacher and the student.

First, we introduce *answer distillation*, which penalizes the most confusing answer with a margin loss, to deal with the problem that biased knowledge misleads the student into incorrect predictions. We find that in MRC datasets there exists many confusing answers (Figure 1(a)), which match the category of true answers but are semantically-contradicted to the question, and an extreme case is the adversarial example. Once the teacher produces biased probabilities towards these distractors, inaccurate distributions will be distilled and later used to supervise the student. As a result, the student could produce over-confident wrong predictions. We refer to this problem as *biased distillation*, and Figure 1(b) gives an example. To address this problem, our approach distilles the boundary of the strongest distractor from the teacher, and explicitly informs the student to decrease its confidence score. This forces the student to produce comparable and unbiased distributions between candidate answers.

Second, we present *attention distillation* that aims to match the attention distribution between the teacher and the student. We notice that neural attention plays an important role in MRC tasks by enabling the model to capture complex interactions between the question and the passage. Compared to other forms of intermediate representation such as feature map and neuron selectivity, attention distribution is more compressed and more informative in reflecting semantic similarities of input text pairs. Hence, by mimicking the word alignments from an ensemble model, we expect that the student can learn to attend more precisely with better compression efficiencies.

We evaluate our approach on the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016), the adversarial SQuAD dataset (Jia and Liang, 2017) and the NarrativeQA benchmark (Kočiskỳ et al., 2017). Compared to the ensemble teacher, the single student model trained with our approach has a slight drop of 0.4% F1 on the SQuAD test set, while running 12× faster during inference. The student even outperforms the teacher on the adversarial SQuAD dataset, and surpasses the teacher in terms of Bleu-1 score on the NarrativeQA benchmark.

2 Background

2.1 Machine Reading Comprehension

In the extractive MRC task, the question and the passage are described as sequences of word tokens, denoted as $Q = \{q_i\}_{i=1}^n$ and $P = \{p_j\}_{j=1}^m$ respectively. The task is to predict an answer A, which is constrained as a segment of text in the passage: $A = \{p_j\}_{j=k}^l$.

To model complicated interactions between the question and the passage, the attention mechanism (Bahdanau et al., 2015) has been widely used. Let $V = \{v_i\}_{i=1}^n$ and $U = \{u_j\}_{j=1}^m$ denote the encoded question/passage representations, where v_i and u_j are both dense vectors. A similarity matrix $E \in \mathbb{R}^{n \times m}$ can be computed as:

$$E_{ij} = f(v_i, u_j)$$

where f is a scalar function producing the similarity between two inputs. Let $\operatorname{softmax}(x)$ denote the softmax function that normalizes the vec-

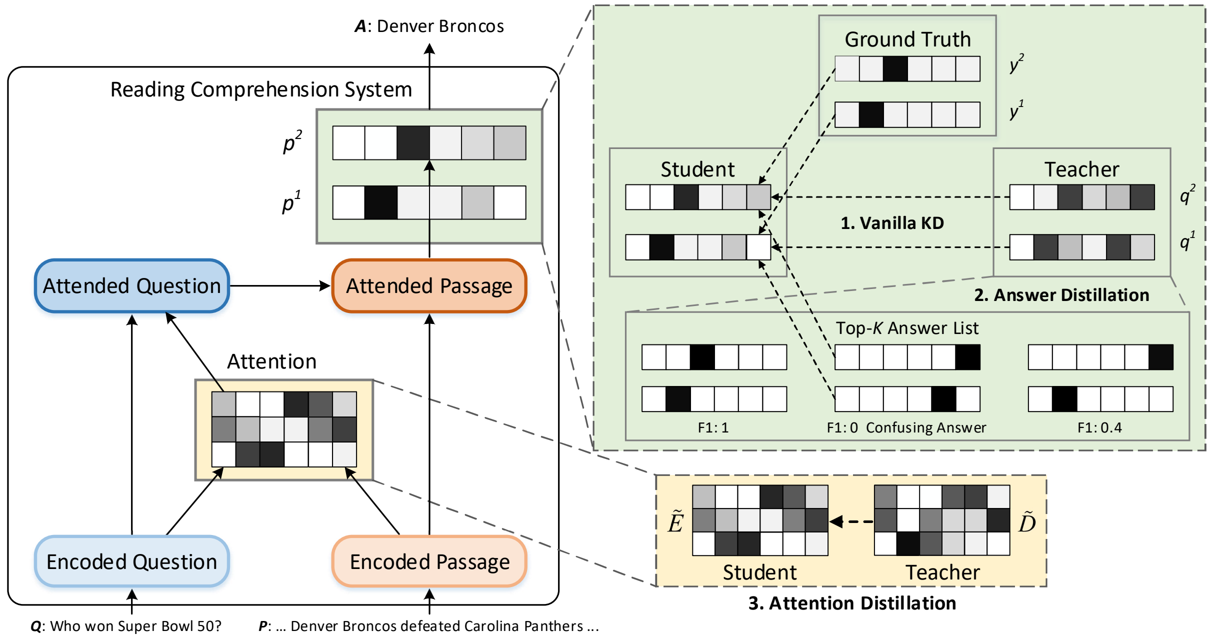


Figure 2: Overview of our approaches. In vanilla knowledge distillation (green), the cross entropy is minimized between the student/teacher distributions of answer positions. In answer distillation (green), the student is trained to penalize the most confusing answer distilled from the teacher. In attention distillation (yellow), mean-squared error is minimized between the student/teacher attention distributions. Darker color denotes higher probability.

tor x. Then the attention distribution across the question for the j-th passage word is computed as $\tilde{E}_j = \text{softmax}(E_{:j})$. This attentive information is commonly used to summarize the entire question as a single vector that is later fused back into the j-th passage word (Seo et al., 2017), resulting in an attended passage representation.

Finally, based on the attended passage, previous works usually apply a pointer network (Vinyals et al., 2015) to produce two probabilities p^1 and p^2 that indicate the answer start and end positions:

$$p(A|Q,P) = p^1(k|Q,P)p^2(l|k,Q,P)$$

where p is the model distribution.

2.2 Knowledge Distillation

The knowledge distillation framework (Hinton et al., 2014) consists of two networks: a *teacher* T, which is a pre-trained large model or an ensemble of multiple models, and a *student* S, which is a smaller network that learns from the teacher. The idea is to supervise the student with not only ground truth labels but also output distributions of the teacher. Concretely, given a data set of examples of the form (X, Y), the cross-entropy loss can be minimized to learn a multi-class classifier:

$$\mathcal{L}_{CE}(\theta) = -\sum_{k=1}^{m} Y_k \log p(k|X;\theta)$$

where p is the model distribution that is parameterized by θ, and m indicates the number of classes. The standard knowledge distillation is to replace ground truth Y with a soft probability distribution q generated by the teacher as:

$$q = \text{softmax}(\frac{\alpha}{\tau}), \quad p = \text{softmax}(\frac{\beta}{\tau})$$

$$\mathcal{L}_{KD}(\theta_S) = -\sum_{k=1}^{m} q(k|X;\theta_T) \log p(k|X;\theta_S)$$

where α and β are pre-softmax logits for teacher and student respectively, and τ is a temperature coefficient that is normally set to 1. A higher τ produces a softer probability distribution, and thus, provides more information about the relative similarity between classes. As Hinton et al. (2014) suggested, the above losses can be jointly optimized as follows:

$$\mathcal{L}(\theta_S) = \mathcal{L}_{CE}(\theta_S) + \lambda \mathcal{L}_{KD}(\theta_S)$$

where λ is usually set as τ^2 since the magnitudes of gradients produced by $\mathcal{L}_{KD}$ scale as $1/\tau^2$.

3 Attention-guided Answer Distillation

Figure 2 gives an overview of our distillation framework, which mainly consists of three approaches including vanilla knowledge distillation, answer distillation and attention distillation. First, we explore standard knowledge distillation in the context of MRC by replacing one-hot labels with soft output distributions extracted from the teacher. We then distill the span of the most confusing answer from the teacher, so that the student can learn to distinguish gold answers from the distractor. Next, we utilize teacher's attention distributions to guide the student's training process for forcing the student to attend more precisely. Finally, the student is jointly trained with the above distilled knowledge.

3.1 Vanilla Knowledge Distillation

The standard training method for MRC models is to minimize the cross entropies on two answer positions (Wang and Jiang, 2017):

$$\mathcal{L}_{CE} = -\sum_{k=1}^{m}\sum_{l=1}^{m} y_k^1 \log p^1(k) + y_l^2 \log p^2(l|k)$$

where y^1 and y^2 are one-hot labels for the answer start and end positions respectively, and m refers to the passage length. We denote $p^1(k|Q,P)$ as $p^1(k)$ and $p^2(l|k,Q,P)$ as $p^2(l|k)$ for abbreviation.

Following the standard procedure of knowledge distillation in Section 2.2, we can replace one-hot labels with output distributions of answer start and end positions predicted by the teacher as:

$$\mathcal{L}_{KD} = -\sum_{k=1}^{m}\sum_{l=1}^{m} q^1(k) \log p^1(k) + \\ q^2(l|k) \log p^2(l|k)$$

Here, we first consider the teacher as a single model, and later extend it to the ensemble scenario in Section 3.4.

3.2 Answer Distillation

Vanilla knowledge distillation allows the student to learn relative similarities between answer candidates at position level. However, the student can suffer from the biased distillation problem once the teacher makes wrong predictions towards confusing answers. As a result, the model may be over-confident on the distractors during inference. Therefore, it is necessary to produce confidence scores that are comparable and unbiased between candidate answers. Since there usually exists one most confusing candidate answer, we hence consider explicitly informing the student about its boundary so as to relatively decrease the corresponding confidence.

Specifically, we perform inference with the teacher network to get a top-K answer list $\mathcal{A}_{\mathcal{K}}$ for each example, and measure the word overlap between the ground truth A^* and each answer candidate A_i from the list. The one that shares no overlap with gold answers and has the highest confidence score is chosen as the *confusing answer*. We use the F1 scoring function $F_1(A^*, A_i)$ as the measurement of word overlap and use the probability $q(A_i)$ as the confidence score. If the F1 scores of all top-K answers are larger than 0, then we argue that the teacher makes a good prediction and therefore do not distill the confusing answer for this example. The above process is performed on the entire training set to annotate all potential confusing answers.

Once the process is finished, we force the student to produce confidence scores of gold answers that are distinguishable from the score of confusing answer, by minimizing a margin ranking loss (Bai et al., 2010) as:

$$\mathcal{L}_{ANS} = \max(0, 1 - \beta_k^1 + \beta_i^1) + \\ \max(0, 1 - \beta_l^2 + \beta_j^2)$$

where β is the student's pre-softmax logit, k and l indicate the boundary of gold span while i and j refer to the confusing boundary. With this loss, we penalize the strongest distractor on which the model is over-confident, and encourage the true answer that is underestimated by the student.

3.3 Attention Distillation

The above distillation approaches allow knowledge to be transferred through teacher's outputs. However, we would like to distill not only the final outputs but also some intermediate representations (Romero et al., 2015), in order to provide additional supervised signals for training the student network. The standard approach is to regress the student's intermediate passage representation to the teacher's corresponding representation as a pre-training step. Nevertheless, this approach is cumbersome in that the dimension of passage representation, denoted as $h \times m$ where h is the hidden size, is quite large, leading to a huge amount of resource consumption as we need to distill all training examples. Since neural attention is more compressed and contains rich informantion about where to attend, we hence propose to match the attention distribution between the teacher and the student instead.

Following the notation in Section 2.1, let $\tilde{D}_j$ and $\tilde{E}_j$ denote the attention distribution across the question for the j-th passage word in teacher and student repectively. We define a mean-squared loss as:

$$\mathcal{L}_{ATT} = \frac{1}{2}\sum_{j=1}^{m}||\tilde{D}_j - \tilde{E}_j||^2$$

Compared to regressing the hidden representation, attention distillation brings in two benefits: 1) the dimension of similarity matrix is $n \times m$, where the question length n is significantly smaller than the hidden size h, resulting in a better compression efficiency; 2) the student can learn to attend more precisely as neural attention directly reflects semantic similarities between input text pairs.

3.4 Joint Training

So far, we define three losses to transfer knowledge between MRC models. Next we consider using a joint objective to train the student network as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda\mathcal{L}_{KD} + \gamma\mathcal{L}_{ANS} + \delta\mathcal{L}_{ATT} \quad (1)$$

where λ is set as τ^2 to scale gradients, γ and δ are two hyper-parameters that control the task-specific weights.

In addition, since the knowledge is distilled from an ensemble model, we need to integrate multiple outputs into an unified label. For vanilla knowledge distillation and attention distillation, we use an arithmetic mean of their individual predicted distributions as the final soft label. As for answer distillation, we choose the confusing answer that has the highest confidence score among all models.

In summary, the entire training procedure is: 1) train an ensemble teacher model; 2) distill knowledge using the teacher on the training set; 3) integrate multiple types of knowledge to build a new training set; 4) train a single student model with Equation 1 on the new dataset.

4 Experiments

4.1 Datasets and Metrics

We conduct experiments on the following three datasets.

SQuAD (Stanford Question Answering Dataset) (Rajpurkar et al., 2016) is a machine comprehension dataset containing $100,000+$ questions that are annotated by crowd workers on 536 Wikipedia articles. The answer to each question is always a span in the corresponding passage.

Adversarial SQuAD (Jia and Liang, 2017) is a dataset aiming to test whether the model truely understands the text by appending an adversarial sentence to the passage. A strong confusing answer is constructed in the adversarial sentence to distract the answer prediction.

NarrativeQA (Kočiský et al., 2017) is a benchmark proposed for story-based reading comprehension. The answers in this dataset are handwritten by human annotators based on a short summary. Following Tay et al. (2018), we compete on the summary setting to compare with reported baselines. We choose the span that achieves the highest Rouge-L score with respect to the gold answers as labels for training.

We use the official metrics to perform the evaluation. Specifically, for both SQuAD and adversarial SQuAD datasets, we report exact match (EM) and F1 scores. As for the NarrativeQA benchmark, the metrics are Blue-1, Bleu-4 and Rouge-L.

4.2 Implementation Details

We implement the Reinforced Mnemonic Reader (RMR) (Hu et al., 2018) as our base model, which contains the standard attention mechanism and the two-step answer prediction described in Section 2.1. Therefore, our distillation approaches can be seamlessly used, and the improvement of our approaches can be directly compared to the baseline. We use slightly different network architectures and hyper-parameters for different datasets. More specifically, we use the original configuration for SQuAD and adversarial SQuAD datasets. As for the NarrativeQA benchmark, we truncate the passage to the first 800 words and use a batch size of 32. Besides, we also remove ELMo embeddings (Peters et al., 2018) and reduce the number of aligning layer to 2 to avoid out-of-memory problem.

We run 12 single models with the identical architecture but different initial parameters to obtain the ensemble model. The student has the same network architecture as the teacher. We reuse the models trained on the SQuAD dataset to test on

Model	Dev		Test	
	EM	F1	EM	F1
LR Baseline[1]	40.0	51.0	40.4	51.0
FusionNet[2]	75.3	83.6	76.0	83.9
BiSAE[3]	77.9	85.6	78.6	85.8
R-Net+[4]	-	-	79.9	86.5
SLQA+[5]	80.0	87.0	80.4	87.0
QANet[6]	-	-	80.9	87.8
BiSAE (E)	79.6	86.6	81.0	87.4
R-Net+ (E)	-	-	82.6	88.5
SLQA+ (E)	82.0	88.4	82.4	88.6
QANet (E)	-	-	82.7	89.0
RMR[7]	78.9	86.3	79.5	86.6
RMR (E)	**81.2**	**87.9**	**82.3**	**88.5**
RMR + A2D	80.3	87.5	81.5	88.1

Table 1: Comparison of different approaches on the SQuAD test set, extracted on May 9, 2018: Rajpurkar et al. (2016)[1], Huang et al. (2018)[2], Peters et al. (2018)[3], Wang et al. (2017)[4], Wang et al. (2018)[5], Yu et al. (2018)[6] and Hu et al. (2018)[7]. BiSAE refers to BiDAF + Self Attention + ELMo. (E: ensemble model)

adversarial SQuAD datasets, but we retrain another pair of teacher and student models for the NarrativeQA benchmark. The temperature τ is tuned among $[1, 2, 3, 5]$ and is set as 2 by default. The weight γ is set to 0.3 and δ is 0.1. K is set to 4 for generating the top-K answer list. All experiments and runtime benchmarks are tested on a single Nvidia Tesla P100 GPU. Our approach is denoted as A2D for abbreviation.

4.3 Main Results

In this section we report main results on three MRC datasets[1]. Since our main purpose is to compress the ensemble model into a single model that possesses better efficiency and robustness, we focus on comparing our approach with the ensemble baseline.

We present the results on the test set of SQuAD in Table 1. We can see that the student network is able to compete with the ensemble model with only a slight drop of 0.4% on F1. Moreover, nearly 80% of the improvement in terms of F1 achieved by the ensemble model is successfully transferred to the distilled model, which indicates the effectiveness of our approach.

To validate the effect of our approach on enhancing robustness, we show the results on two adversarial SQuAD datasets, namely AddOneSent and AddSent, in Table 2. As we can see, the improvement on adversarial data is much higher than the one on original SQuAD dataset: the student network successfully surpasses the teacher. The significant improvement comes from the fact that there exists much more confusing answers in adversarial datasets, and hence the baseline is more likely to be over-confident on distractors. Our approach, however, explicitly decreases distractor's confidence, thus yielding more robust predictions against adversarial examples.

Model	AddSent		AddOneSent	
	EM	F1	EM	F1
LR Baseline	17.0	23.2	22.3	30.4
BiSAE	38.7	44.4	48.0	54.7
SLQA+	-	52.1	-	62.7
DCN+ (MINI)[1]	52.2	59.7	60.1	67.5
FusionNet (E)	46.2	51.4	54.7	60.7
SLQA+ (E)	-	54.8	-	64.2
RMR	53.0	58.5	60.9	67.0
RMR (E)	56.0	61.1	62.7	68.5
RMR + A2D	**56.0**	**61.3**	**63.3**	**69.3**

Table 2: Comparison of different approaches on two adversarial SQuAD datasets: Min et al. (2018)[1]. (E: ensemble model)

Model	Bleu-1	Bleu-4	Rouge-L
Seq2Seq	15.9	1.3	13.2
AS Reader[1]	23.2	6.4	22.3
BiDAF[2]	33.7	15.5	36.3
BiAttention[3]	36.6	19.8	41.4
RMR	48.4	24.6	51.5
RMR (E)	50.1	**27.5**	**53.9**
RMR + A2D	**50.4**	26.5	53.3

Table 3: Comparison of different approaches on the NarrativeQA test set using summaries: Kadlec et al. (2016)[1], Seo et al. (2017)[2] and Tay et al. (2018)[3]. (E: ensemble model)

To verify the generalizability of our approach among various datasets, we further detail the results on the test set of NarrativeQA in Table 3, Despite that both of our single baseline and ensemble model have already achieved top results, the student's performance can still be boosted using our approach, even outperforming the teacher in terms of Bleu-1 score.

4.4 Speedup over Ensemble Model

Next, in order to show the improvement on efficiency brought by our approach, we compare

[1] https://worksheets.codalab.org/worksheets/0xfa2548e70c2540ca8c2d26d7dd402f7f/.

	Params	Time	Speedup
RMR (E)	6.9m×12	118.2	-
RMR + A2D	6.9m	**9.6**	**12.3×**

Table 4: Comparison between the ensemble teacher and the single student on the SQuAD dev set. Time denotes number of minutes needed to perform the entire inference. (E: ensemble model)

the inference speedup of the student against the teacher in Table 4. Since we use 12 single models to make the ensemble, we can easily see that the student is nearly 12× faster than the teacher. Besides, compared to the baseline, the student has nearly the same number of parameters, demonstrating that our approach does not introduce additional computation complexity.

4.5 Ablation Study

To get better insights of our distillation approaches, we conduct in-depth ablation study on both the development set of SQuAD and the AddSent set of adversarial SQuAD. We mainly focus on the F1 score since it is used as the main metric on the SQuAD leaderboard.

Table 5 shows the ablation results. First, we evaluate the vanilla knowledge distillation by removing the KD loss. We find that this loss contributes a lot on both datasets, implying that standard knowledge distillation is helpful for MRC models. Next we ablate the attention loss (ATT), and observe a smaller performance degradation on both datasets, which suggests that matching attention distributions is also beneficial but less effective. We then test the effect of removing answer distillation (ANS), and discover that although the impact on the SQuAD dev set is relatively small, the influence on the AddSent dataset is quite large. We argue that this is because the model trained with answer distillation can better handle the biased distillation problem, which is more severe in the adversarial dataset. Finally, we replace the joint training process with a stage-wise fashion proposed by Romero et al. (2015). Concretely, we first warm up the student by matching the attention distribution as a pre-training step, and then minimize the rest of losses to train the model. The result, however, shows that this strategy does harm to the performance. We think the reason may be that the pre-training step leads the model into a local minima.

	Dev		AddSent	
	F1	ΔF1	F1	ΔF1
RMR + A2D	87.5	-	61.3	-
- Vanilla KD	86.8	-0.7	60.1	-1.2
- ATT	87.0	-0.5	60.4	-0.9
- ANS	87.1	-0.4	59.8	-1.5
- Joint Training	87.3	-0.2	60.8	-0.5

Table 5: Ablation study of different knowledge distillation approaches on SQuAD dev set and AddSent set.

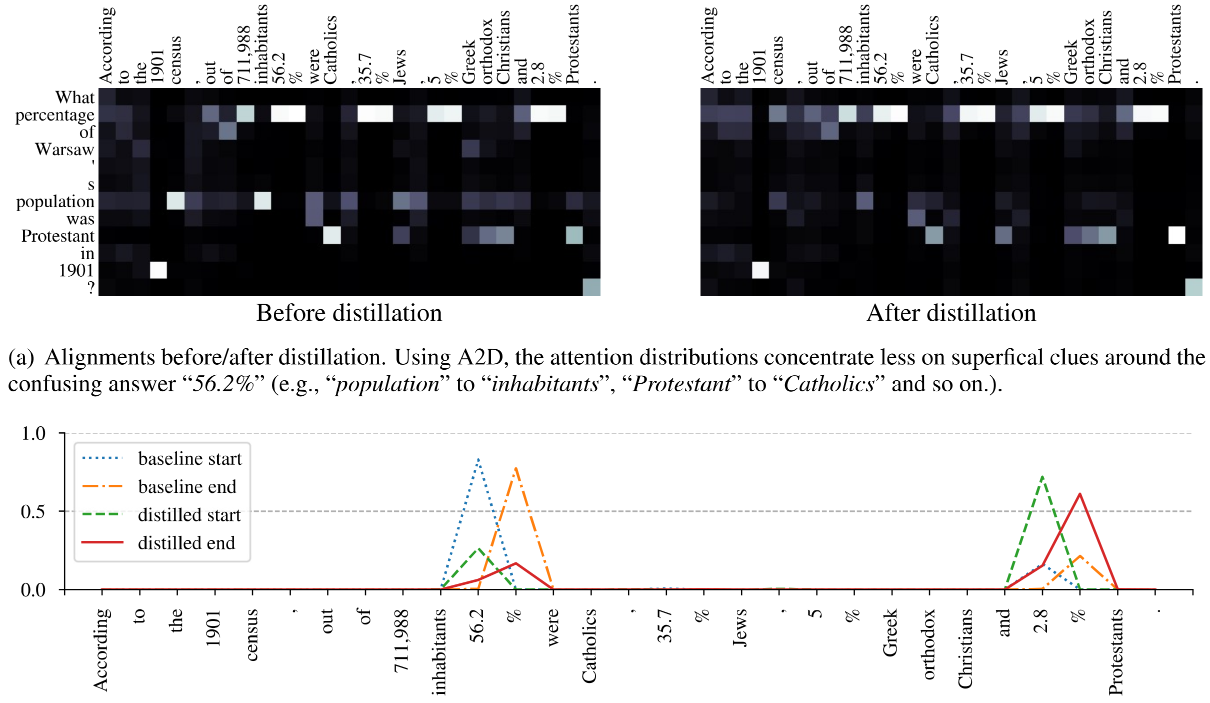
4.6 Analysis and Discussion

We now give a qualitative analysis on our approach by visualizing attention distributions and output probabilities for both of the baseline and the distilled model. From Figure 3(a) we can see that, both models are good at finding candidate answers, such as "*56.2%*" and "*2.8%*", according to the key question word "percentage". Nevertheless, the base model pays more attention on question-passage word pairs around the confusing answer "*56.2%*". As a result, the model is more likely to produce a high confidence score on "*56.2%*", as shown in Figure 3(b). Using our approach, however, leads to less concentrations on superfical clues around the confusing answer. Instead, the distilled model is able to focus more on the critical alignments (e.g., "*Protestant*" to "*Protestants*"), and therefore predicts the correct answer "*2.8%*".

To show on which type of questions our approach is doing better, we report the results in Figure 4. We can see that our approach yields consistent performance gain over the baseline across different question types on various datasets. Particularly, A2D provides an obvious advantage for question types such as "*what*", "*who*", "*which*", "*how*", "*why*" and so on.

5 Related Work

Machine Reading Comprehension. Benefiting from large-scale machine reading comprehension (MRC) datasets (Hermann et al., 2015; Hill et al., 2016; Rajpurkar et al., 2016; Joshi et al., 2017), end-to-end neural networks have achieved promising results (Wang et al., 2018; Yu et al., 2018). Wang and Jiang (2017) combine the match-LSTM with pointer networks to predict the answer boundary. Wang et al. (2017) match the context aginst itself to refine the passage representation. Later, a variety of attention mechanisms have been



(a) Alignments before/after distillation. Using A2D, the attention distributions concentrate less on superfical clues around the confusing answer "*56.2%*" (e.g., "*population*" to "*inhabitants*", "*Protestant*" to "*Catholics*" and so on.).

(b) The base model points to the confusing answer ("*56.2%*"), while the distilled model predicts the correct one ("*2.8%*").

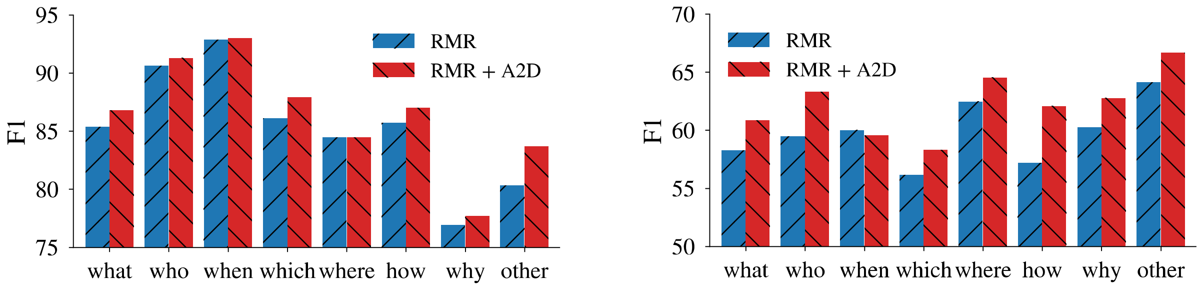
Figure 3: A case study between the base model and the distilled model.

proposed, such as bi-attention (Seo et al., 2017), coattention (Xiong et al., 2018), fully-aware attention (Huang et al., 2018) and reattention (Hu et al., 2018). Among these works, two common traits can be summarized as: 1) compute a similary matrix between the question and the passage; 2) sequentially predict the answer start and end positions. Our proposed approach is a simple and effective adaptation to existing models by taking advantage of these traits, and do not complicate previous works more than necessary.

Efficiency and Robustness in MRC. Improving efficiency and robustness for reading comprehension system has attracted a lot of interest in recent years. For efficiency, previous works mostly concentrate on how to scale passage-level models to large corpora such as a document without increasing computation complexity. Existing approaches (Chen et al., 2017; Clark and Gardner, 2018) usually first retrieve relevant passages with a ranking model and then return an answer with a reading model. As for robustness, Wang and Bansal (2018) train the model with an adversarial data augmentation method. Min et al. (2018) propose to selectively read salient sentences rather than the entire passage, so as to avoid looking at the adversarial sentence. Our approach, however, focuses on improving efficiency and robustness by transferring knowledge from a cumbersome ensemble model to a single model.

Knowledge Distillation. Knowledge distillation is first explored by Bucilu et al. (2006) and Hinton et al. (2014), which attempts to transfer knowledge defined as soft output distributions from a teacher to a student. Later works have been proposed to distill not only the final output but also intermediate representation from the teacher (Romero et al., 2015; Zagoruyko and Komodakis, 2017; Huang and Wang, 2017). Papernot et al. (2016) show that knowledge distillation can be used to prevent the network from adversarial attacks in image recognition. Radosavovic et al. (2017) introduce data distillation that annotates large-scale unlabelled data for omni-supervised learning.

In natural language processing (NLP), Mou et al. (2016) distill task-specific knowledge from word embeddings. Kuncoro et al. (2016) propose to learn a single parser from an ensemble of parsers. Kim and Rush (2016) investigate knowledge distillation for neural machine translation by approximately matching the sequence-level distribution of the teacher. Nakashole and Flauger (2017) propose to learn bilingual mapping functions through a distilled training objective. Xu and Yang (2017) distill discriminative knowledge across languages for cross-lingual text classifica-



(a) Results on the SQuAD dev set.

(b) Results on the AddSent set.

Figure 4: Comparison of different question types between the base model and the distilled model.

tion. Our work shows that the standard knowledge distillation and its novel variants can be successfully applied to the MRC task.

6 Conclusion

In this paper, we investigate knowledge distillation for machine reading comprehension. We first explore vanilla knowledge distillation to transfer knowledge of answer positions, and then propose two variant approaches including answer distillation for penalizing student's predictions on confusing answer spans, and attention distillation for transferring teacher's attentive information. Experiments show that the ensemble model has been successfully compressed into a single model that possesses better efficiency and robustness.

In future work, we will explore new distillation methods that have better compression capabilities for MRC tasks, such as distilling knowledge from a single model instead of the ensemble without lossing performances, adding weights on knowledge based on the distilling quality and so on. We also plan to further study the biased distillation problem and explore the compatibility of our approach in other NLP tasks such as natural language inference (Bowman et al., 2015), answer sentence selection (Yang et al., 2015) and so on.

Acknowledgments

We would like to thank Pranav Rajpurkar for his help with SQuAD submissions. This work is funded by National Key R&D Program of China (No. 0708063216003).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. 2010. Learning to rank with (a lot of) word features. *Information retrieval*, 13(3):291–314.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference systems. In *Proceedings of EMNLP*.

Cristian Bucilu, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of SIGKDD*, pages 535–541. ACM.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of ACL*.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of ACL*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of NIPS*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading childrens books with explicit memory representations. In *Proceedings of ICLR*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. In *Proceedings of NIPS Workshop*.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced mnemonic reader for machine reading comprehension. In *Proceedings of IJCAI*.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2018. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. In *Proceedings of ICLR*.

Zehao Huang and Naiyan Wang. 2017. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of EMNLP*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of ACL*.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of ACL*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of EMNLP*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge. *arXiv preprint arXiv:1712.07040*.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, and Noah A Smith. 2016. Distilling an ensemble of greedy dependency parsers into one mst parser. In *Proceedings of EMNLP*.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. Efficient and robust question answering from minimal context over documents.

Lili Mou, Ran Jia, Yan Xu, Ge Li, Lu Zhang, and Zhi Jin. 2016. Distilling word embeddings: An encoding approach. In *Proceedings of CIKM*, pages 1977–1980. ACM.

Ndapandula Nakashole and Raphael Flauger. 2017. Knowledge distillation for bilingual dictionary induction. In *Proceedings of EMNLP*, pages 2487–2496.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word preprsentations. In *Proceedings of NACCL*.

Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. 2017. Data distillation: Towards omni-supervised learning. *arXiv preprint arXiv:1712.04440*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *Proceedings of ICLR*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hananneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of ICLR*.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Multi-range reasoning for machine comprehension. *arXiv preprint arXiv:1803.09074*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Proceedings of NIPS*.

Shuohang Wang and Jing Jiang. 2017. Machine comprehension using match-lstm and answer pointer. In *Proceedings of ICLR*.

Wei Wang, Ming Yan, and Chen Wu. 2018. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of ACL*.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of ACL*.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. In *Proceedings of NAACL*.

Caiming Xiong, Victor Zhong, and Richard Socher. 2018. Dcn+: Mixed objective and deep residual coattention for question answering. In *Proceedings of ICLR*.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. In *Proceedings of ACL*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of EMNLP*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of ICLR*.

Sergey Zagoruyko and Nikos Komodakis. 2017. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of ICLR*.