Towards Efficient and Robust Machine Reading Comprehension

Minghao Hu

Joint work with Yuxing Peng[1], Furu Wei[2], Nan Yang[2], Zhen Huang[1] and Ming Zhou[2]

Oct 13, 2018

1 

2 Microsoft

Machine Reading Comprehension

Towards the Machine Comprehension of Text: An Essay

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

December 23, 2013

Teaching Machines to Read and Comprehend

Karl Moritz Hermann[†] **Tomáš Kočiský**[†‡] **Edward Grefenstette**[†]
Lasse Espeholt[†] **Will Kay**[†] **Mustafa Suleyman**[†] **Phil Blunsom**[†‡]
[†]Google DeepMind [‡]University of Oxford
{kmh,tkocisky,etg,lespeholt,wkay,mustafasul,pblunsom}@google.com

Machine Reading Comprehension

"A machine **comprehends** a **passage** of text if, for any **question** regarding that text that can be **answered** correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question."

(Burges, 2013)

Machine Reading Comprehension

Cloze-style MRC

Passage (P):

(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 "

Question (Q): characters in " @placeholder" movies have gradually become more diverse.

Answer (A): @entity6

(Hermann et al., 2015)

Machine Reading Comprehension

Extractive MRC

Passage (P):

In 1870, Tesla moved to Karlovac, to attend school at the Higher Real Gymnasium, where he was profoundly influenced by a math teacher Martin Sekulić.:32 The classes were held in German, as it was a school within the Austro-Hungarian Military Frontier. Tesla was able to perform integral calculus in his head, which prompted his teachers to believe that he was cheating. He finished a four-year term in three years, graduating in 1873.:33

Question (Q): Who was Tesla influenced by while in school?

Answer (A): Martin Sekulić

(Rajpurkar et al., 2016)

Dataset

Before 2015:

- MCTest: 2600 questions
- ProcessBank: 500 questions

Dataset

Before 2015:

- MCTest: 2600 questions
- ProcessBank: 500 questions

Between 2015-2017:

- **CNN/Daily Mail**
- Children Book Test
- WikiReading
- LAMBADA
- **SQuAD**
- Who did What
- NewsQA
- **MS MARCO**

…

Dataset

Before 2015:

- MCTest: 2600 questions
- ProcessBank: 500 questions

Between 2015-2017:

- **CNN/Daily Mail**
- Children Book Test
- WikiReading
- LAMBADA
- **SQuAD**
- Who did What
- NewsQA
- **MS MARCO**
- …

After 2017:

- Quasar
- SearchQA
- **TriviaQA**
- RACE
- NarrativeQA
- QAngaroo
- **SQuAD 2.0**
- **CoQA**
- **QuAC**
- **HotpotQA**
- …

Dataset

The Stanford Question Answering Dataset (SQuAD)

SQuAD: 100,000+ Questions for Machine Comprehension of Text

Pranav Rajpurkar and **Jian Zhang** and **Konstantin Lopyrev** and **Percy Liang**
{pranavsr,zjian,klopyrev,pliang}@cs.stanford.edu
Computer Science Department
Stanford University

EMNLP2016 Best Resource Paper

SQuAD1.1 Leaderboard

Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Dataset

Adversarial SQuAD (Jia and Liang, 2017)

Fool the model by appending an adversarial sentence to the passage.

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Model

Before 2015:

- Lexical Matching
- Logistic Regression

Model

Before 2015:

- Lexical Matching
- Logistic Regression

Cloze-style Model (2015-2016):

- Attentive/Impatient Reader
- Attention Sum Reader
- Gated-attention Reader
- Stanford Attentive Reader
- Iterative Attentive Reader
- Attention-over-Attention Reader

...

Model

Before 2015:

- Lexical Matching
- Logistic Regression

Cloze-style Model (2015-2016):

- Attentive/Impatient Reader
- Attention Sum Reader
- Gated-attention Reader
- Stanford Attentive Reader
- Iterative Attentive Reader
- Attention-over-Attention Reader

…

Extractive Model (2016-now):

- Match-LSTM
- BiDAF
- DCN
- FusionNet
- Reinforced M-Reader
- SLQA
- MARS
- R-Net
- QANet
- NLNet

…

Model

Alibaba and Microsoft AI beat human scores on Stanford reading test

Neural networks edged past human scores on the measure of machine reading.

Human-level performances in terms of Exact Match (EM)

SQuAD 1.1	Test EM	Test F1
LR Baseline	40.4	51.0
DCN+ (*ensemble*)	78.8	86.0
FusionNet (*ensemble*)	79.0	86.0
BiDAF + Self Attention + ELMo (*ensemble*)	81.0	87.4
Reinforced Mnemonic Reader (*ensemble*)	**82.3**	88.5
SLQA+ (*ensemble*)	**82.4**	88.6
Hybrid AoA Reader+ (*ensemble*)	**82.5**	89.3
R-Net+ (*ensemble*)	**82.6**	88.5
QANet (*ensemble*)	**82.7**	89.0
Human Performance	82.3	91.2

(Results extracted form https://rajpurkar.github.io/SQuAD-explorer/ on May 9, 2018)

Does the Extractive MRC being Solved?

1. Although *effective*, the ensemble models are *not efficient*.

Inference time is slow; Huge amount of resource is required

SQuAD 1.1	Test EM	Test F1
LR Baseline	40.4	51.0
DCN+ (*ensemble*)	78.8	86.0
FusionNet (*ensemble*)	79.0	86.0
BiDAF + Self Attention + ELMo (*ensemble*)	81.0	87.4
Reinforced Mnemonic Reader (*ensemble*)	**82.3**	88.5
SLQA+ (*ensemble*)	**82.4**	88.6
Hybrid AoA Reader+ (*ensemble*)	**82.5**	89.3
R-Net+ (*ensemble*)	**82.6**	88.5
QANet (*ensemble*)	**82.7**	89.0
Human Performance	82.3	91.2

(Results extracted form https://rajpurkar.github.io/SQuAD-explorer/ on May 9, 2018)

Does the Extractive MRC being Solved?

2. These models are *not robust* as they are vulnerable to adversarial attacks.

Adversarial SQuAD	AddSent F1	AddOneSent F1
BiDAF + Self Attention + ELMo (*single*)	44.4	54.7
SLQA+ (*single*)	52.1	62.7
Reinforced Mnemonic Reader (*single*)	58.5	67.0
FusionNet (*ensemble*)	51.4	60.7
SLQA+ (*ensemble*)	54.8	64.2
Reinforced Mnemonic Reader (*ensemble*)	61.1	68.5
Human Performance	**79.5**	**89.2**

How to Improve Efficiency and Robustness?

Solution: “Attention-Guided Answer Distillation”

Compress an ensemble model into a single model via:

1. Vanilla knowledge distillation
2. Answer distillation
3. Attention distillation

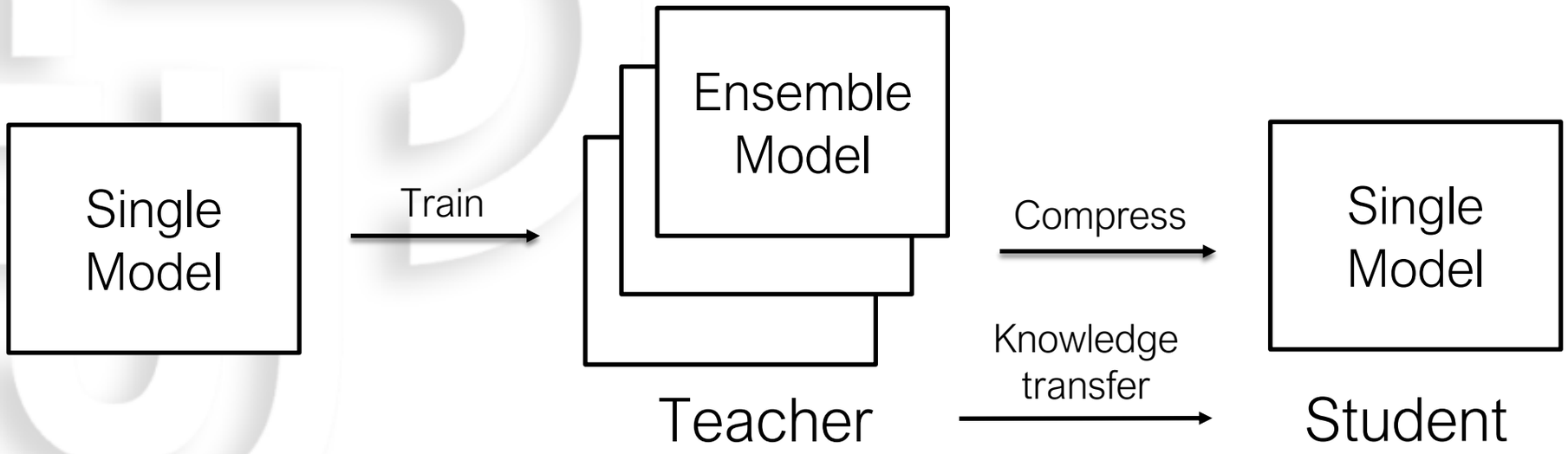
(Hu et al., 2018)

How to Improve Efficiency and Robustness?

Solution: “Attention-Guided Answer Distillation”

Compress an ensemble model into a single model via:

1. Vanilla knowledge distillation
2. Answer distillation
3. Attention distillation



(Hu et al., 2018)

Vanilla Knowledge Distillation

- Train the model with standard cross-entropy loss:

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

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

where p^1 and p^2 are the model distributions of answer boundaries, y^1 and y^2 are ground-truth answer start and end positions.

Vanilla Knowledge Distillation

- Train the student with the supervision from a teacher.
- Replace gold answer spans y with teacher's soft probabilities q.

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

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

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

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

(Hinton et al., 2014)

Answer Distillation

Biased distillation problem

- There exists many *confusing answers* in MRC datasets.
- Once the teacher is fooled by confusing answers, biased knowledge will be distilled to supervise the student.

Answer Distillation

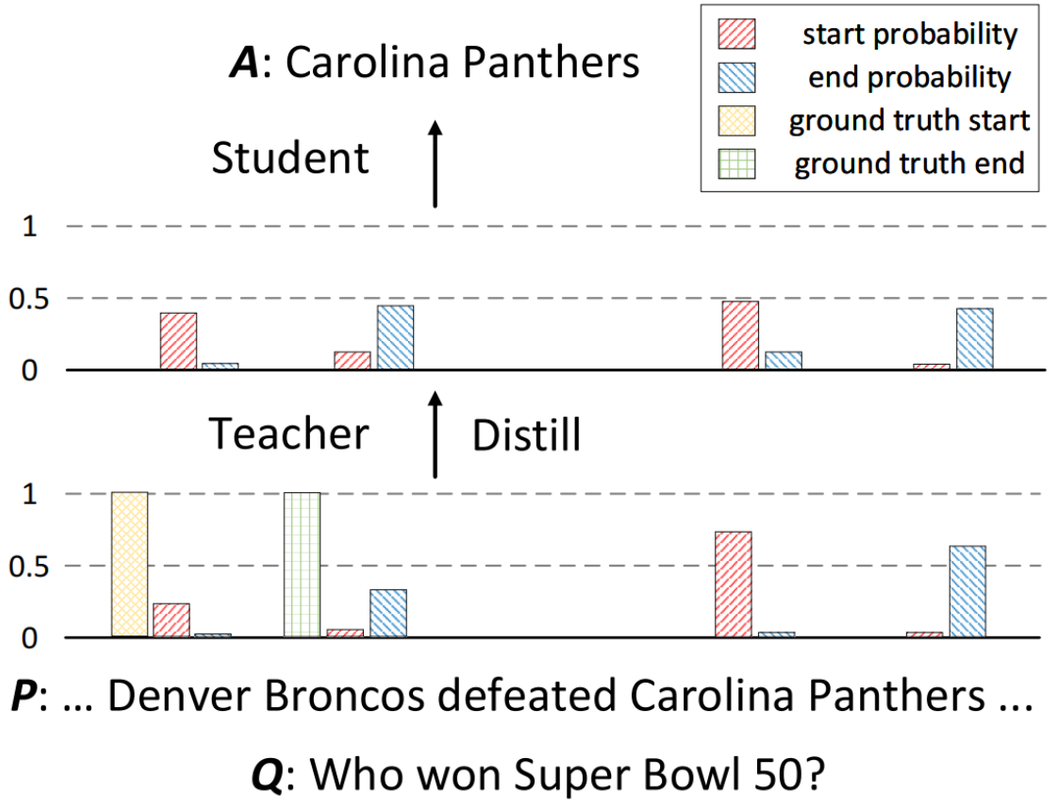
Biased distillation problem

- There exists many *confusing answers* in MRC datasets.
- Once the teacher is fooled by confusing answers, biased knowledge will be distilled to supervise the student.

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?

Gold answers (red) versus confusing answers (blue)



Biased distillation from teacher to student

Answer Distillation

To address the biased distillation problem, we

- explicitly inform the student about the boundary of confusing answers
- relatively decrease its confidence compared to the gold spans

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

where β^1 and β^2 are student's pre-softmax logits of answer boundaries.

Answer Distillation

To address the biased distillation problem, we

- explicitly inform the student about the boundary of confusing answers
- relatively decrease its confidence compared to the gold spans

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

where β^1 and β^2 are student's pre-softmax logits of answer boundaries.

In order to obtain the confusing boundary i and j, we

- get top-K candidate answers from the teacher
- compute F1 score between gold answer and each candidate
- choose the one that has the highest confidence but with 0 F1

Attention Distillation

Previous approaches only transfer knowledge through final outputs

We want to distill intermediate knowledge to provide more supervision

Attention Distillation

Previous approaches only transfer knowledge through final outputs

We want to distill intermediate knowledge to provide more supervision

A potential solution: regressing intermediate passage representation

- The dimension of passage representation h x m: low compression efficiency
- No explicit semantics

Attention Distillation

Previous approaches only transfer knowledge through final outputs

We want to distill intermediate knowledge to provide more supervision

A potential solution: regressing intermediate passage representation

- The dimension of passage representation h x m: low compression efficiency
- No explicit semantics

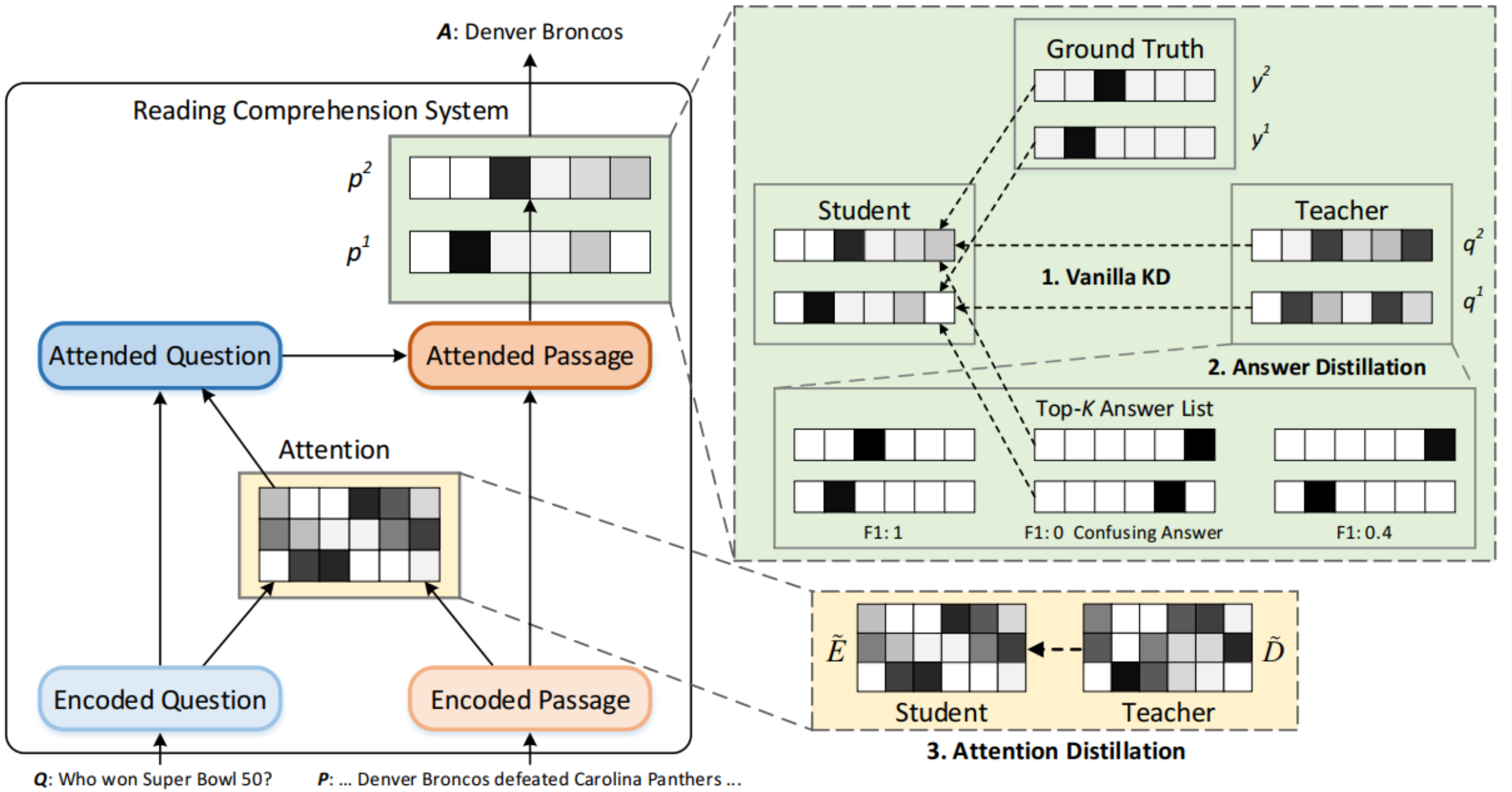
We match the attention distribution between teacher/student instead

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

- The dimension becomes n x m, where n (question length) < h (hidden size)
- Attention distribution reflects semantic similarities

Attention-Guided Answer Distillation (A2D)

Put them together …



Evaluation

- Use Reinforced Mnemonic Reader as the base model

- Experiments on three MRC datasets:

1. SQuAD 1.1
   - extractive MRC benchmark
   - 100,000+ question-answer pairs

2. Adversarial SQuAD
   - two subsets: AddSent and AddOneSent
   - append an adversarial sentence to the passage

3. NarrativeQA
   - story-based MRC benchmark
   - handwritten answer based on a short summary

Evaluation

The student model achieves comparable performances

SQuAD 1.1	Test EM	Test F1
LR Baseline	40.4	51.0
FusionNet (*single*)	76.0	83.9
BiDAF + Self Attention + ELMo (*single*)	78.6	85.8
R-Net+ (*single*)	79.9	86.5
SLQA+ (*single*)	80.4	87.0
Hybrid AoA Reader+ (*single*)	80.0	87.3
QANet (*single*)	80.9	87.8
Reinforced Mnemonic Reader (*single*)	79.5	86.6
Reinforced Mnemonic Reader (*ensemble*)	**82.3**	**88.5**
Reinforced Mnemonic Reader + A2D (*single*)	81.5	88.1

(Results extracted form https://rajpurkar.github.io/SQuAD-explorer/ on May 9, 2018)

Evaluation

The student model outperforms the ensemble teacher model

Better robustness!

Adversarial SQuAD	AddSent (F1)	AddOneSent (F1)
BiDAF + Self Attention + ELMo (*single*)	44.4	54.7
SLQA+ (*single*)	52.1	62.7
FusionNet (*ensemble*)	51.4	60.7
SLQA+ (*ensemble*)	54.8	64.2
Reinforced Mnemonic Reader (*single*)	58.5	67.0
Reinforced Mnemonic Reader (*ensemble*)	61.1	68.5
Reinforced Mnemonic Reader + A2D (*single*)	**61.3**	**69.3**

Evaluation

The student outperforms the ensemble teacher in terms of Bleu-1

NarrativeQA	Bleu-1	Bleu-4	Rouge-L
Seq2Seq (*single*)	15.9	1.3	13.2
AS Reader (*single*)	23.2	6.4	22.3
BiDAF (*single*)	33.7	15.5	36.3
BiAttention (*single*)	36.6	19.8	41.4
Reinforced Mnemonic Reader (*single*)	48.4	24.6	51.5
Reinforced Mnemonic Reader (*ensemble*)	50.1	**27.5**	**53.9**
Reinforced Mnemonic Reader + A2D (*single*)	**50.4**	26.5	53.3

Evaluation

The student is more efficient than the teacher

SQuAD 1.1	Params	Inference Time (minutes)	Speedup
Reinforced Mnemonic Reader (*ensemble*)	6.9m x 12	118.2	-
Reinforced Mnemonic Reader + A2D (*single*)	6.9m	**9.6**	**12.3 x**

Ablation study of different KD approaches

Configuration	SQuAD Dev F1	AddSent F1
Reinforced Mnemonic Reader + A2D	87.5	61.3
- Vanilla KD	86.8	60.1
- Attention Distillation	87.0	60.4
- Answer Distillation	87.1	59.8

Another Perspective of Robustness

Can current models abstain from answering when no answer can be inferred?

Another Perspective of Robustness

Can current models abstain from answering when no answer can be inferred?

Know What You Don't Know: Unanswerable Questions for SQuAD

Pranav Rajpurkar* **Robin Jia*** **Percy Liang**
Computer Science Department, Stanford University
{pranavsr,robinjia,pliang}@cs.stanford.edu

Article: Endangered Species Act
Paragraph: " . . . *Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

Two unanswerable questions along with plausible but incorrect answers (red). Relevant keywords are shown in blue.

ACL2018 Best Short Paper

MRC with Unanswerable Questions

Current Solution: “No-Answer Option”

Additionally predict a special “no-answer” probability in addition to answer span probabilities:

$$\mathcal{L}_{joint} = -\log\left(\frac{(1-\delta)e^{z} + \delta e^{\alpha_a \beta_b}}{e^{z} + \sum_{i=1}^{l_p}\sum_{j=1}^{l_p} e^{\alpha_i \beta_j}}\right)$$

where z is the *no-answer score*, α and β are *answer span scores*, a and b are gold answer boundaries, and δ is 1 if the question is answerable and 0 otherwise.

(Levy et al., 2017; Clark and Gardner, 2018)

MRC with Unanswerable Questions

However, these approaches fail to validate the *answerability* of the question by verifying the *legitimacy* of the predicted answer.

- *Answerability*: whether the question has an answer
- *Legitimacy*: whether the extracted text can be supported by the passage and the question

MRC with Unanswerable Questions

However, these approaches fail to validate the *answerability* of the question by verifying the *legitimacy* of the predicted answer.

- *Answerability*: whether the question has an answer
- *Legitimacy*: whether the extracted text can be supported by the passage and the question

To address the above issue, we propose a *read-then-verify* system that aims to be robust to unanswerable questions.

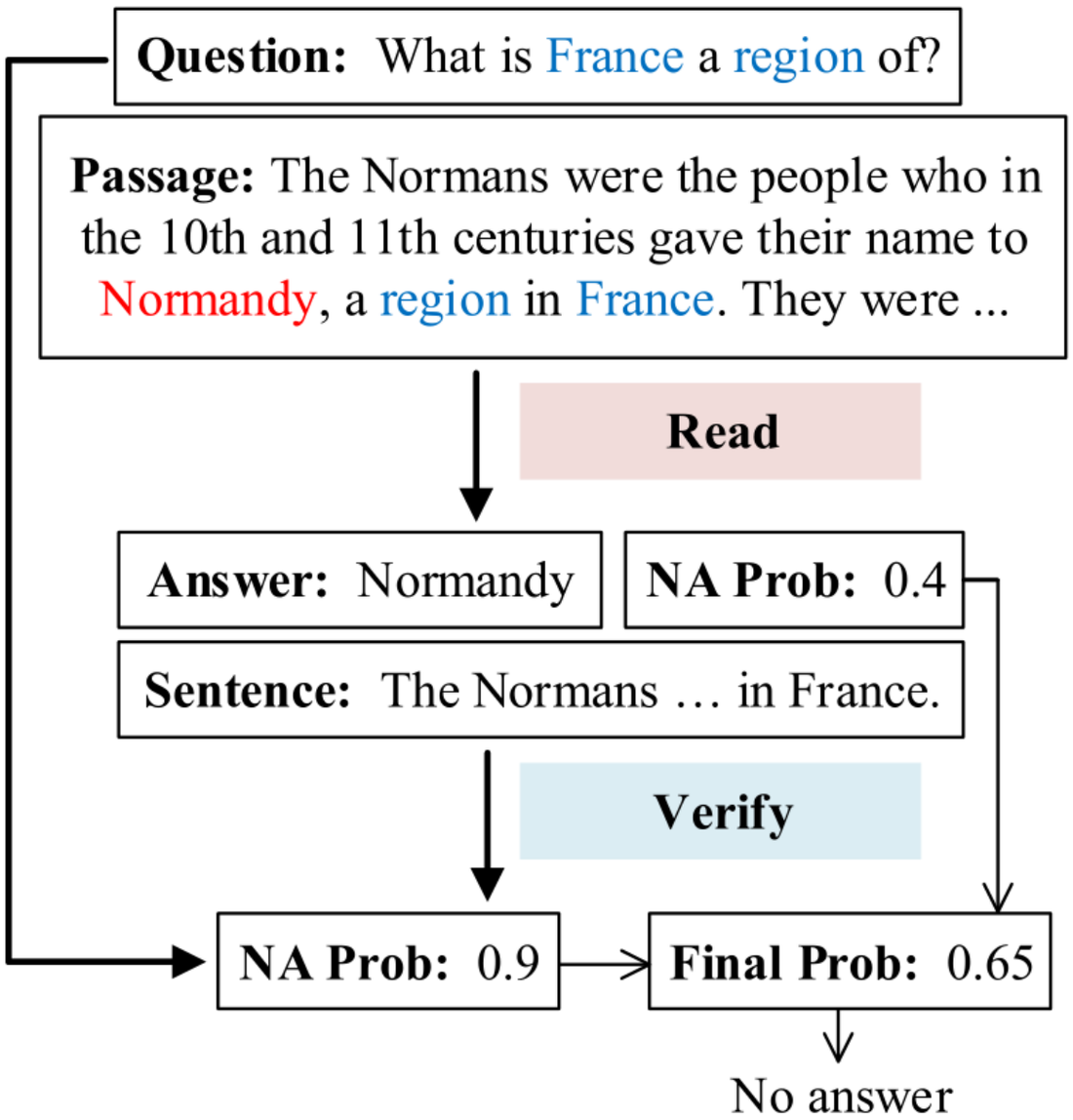
(Hu et al., 2018)

Read-then-Verify System

1. A *reader with auxiliary losses* for extracting candidate answers and detecting unanswerable questions
2. An *answer verifier* for deciding whether or not the extracted candidate is legitimate

Read-then-Verify System

1. A *reader with auxiliary losses* for extracting candidate answers and detecting unanswerable questions
2. An *answer verifier* for deciding whether or not the extracted candidate is legitimate



Reader with Auxiliary Losses

Problems of previous no-answer option approaches:

- The readers are *not trained* to find candidate answers for unanswerable questions, resulting in inaccurate answer extraction.

Reader with Auxiliary Losses

Problems of previous no-answer option approaches:

- The readers are *not trained* to find candidate answers for unanswerable questions, resulting in inaccurate answer extraction.
- Since a shared normalization has been used between two *scores*, inaccurate span probability could lead to imprecise no-answer probabilities, yielding inferior no-answer detection.

Reader with Auxiliary Losses

Therefore, we propose two auxiliary losses to enhance each task:

- *Independent Span Loss*: concentrate on answer extraction by including unanswerable questions as positive examples, and optimize towards their *plausible answers*

Reader with Auxiliary Losses

Therefore, we propose two auxiliary losses to enhance each task:

- *Independent Span Loss*: concentrate on answer extraction by including unanswerable questions as positive examples, and optimize towards their *plausible answers*

$$\mathcal{L}_{indep-I} = -\log\left(\frac{e^{\tilde{\alpha}_{\tilde{a}}\tilde{\beta}_{\tilde{b}}}}{\sum_{i=1}^{l_p}\sum_{j=1}^{l_p} e^{\tilde{\alpha}_i\tilde{\beta}_j}}\right)$$

where $\tilde{\alpha}$ and $\tilde{\beta}$ are another pair of *span scores* generated by a multi-head pointer network, $\tilde{a}$ and $\tilde{b}$ are augmented ground-truth answer boundaries.

Reader with Auxiliary Losses

Therefore, we propose two auxiliary losses to enhance each task:

- *Independent Span Loss*: concentrate on answer extraction by including unanswerable questions as positive examples, and optimize towards their *plausible answers*

$$\mathcal{L}_{indep-I} = -\log\left(\frac{e^{\tilde{\alpha}_{\tilde{a}}\tilde{\beta}_{\tilde{b}}}}{\sum_{i=1}^{l_p}\sum_{j=1}^{l_p} e^{\tilde{\alpha}_i\tilde{\beta}_j}}\right)$$

where $\tilde{\alpha}$ and $\tilde{\beta}$ are another pair of *span scores* generated by a multi-head pointer network, $\tilde{a}$ and $\tilde{b}$ are augmented ground-truth answer boundaries

- *Independent No-Answer Loss*: focus on no-answer detection by exclusively optimizing the no-answer score

$$\mathcal{L}_{indep-II} = -(1-\delta)\log\sigma(z) - \delta\log(1-\sigma(z))$$

Answer Verifier

Recognize local textual entailment by comparing the answer sentence and the question, with three different architectures:

Answer Verifier

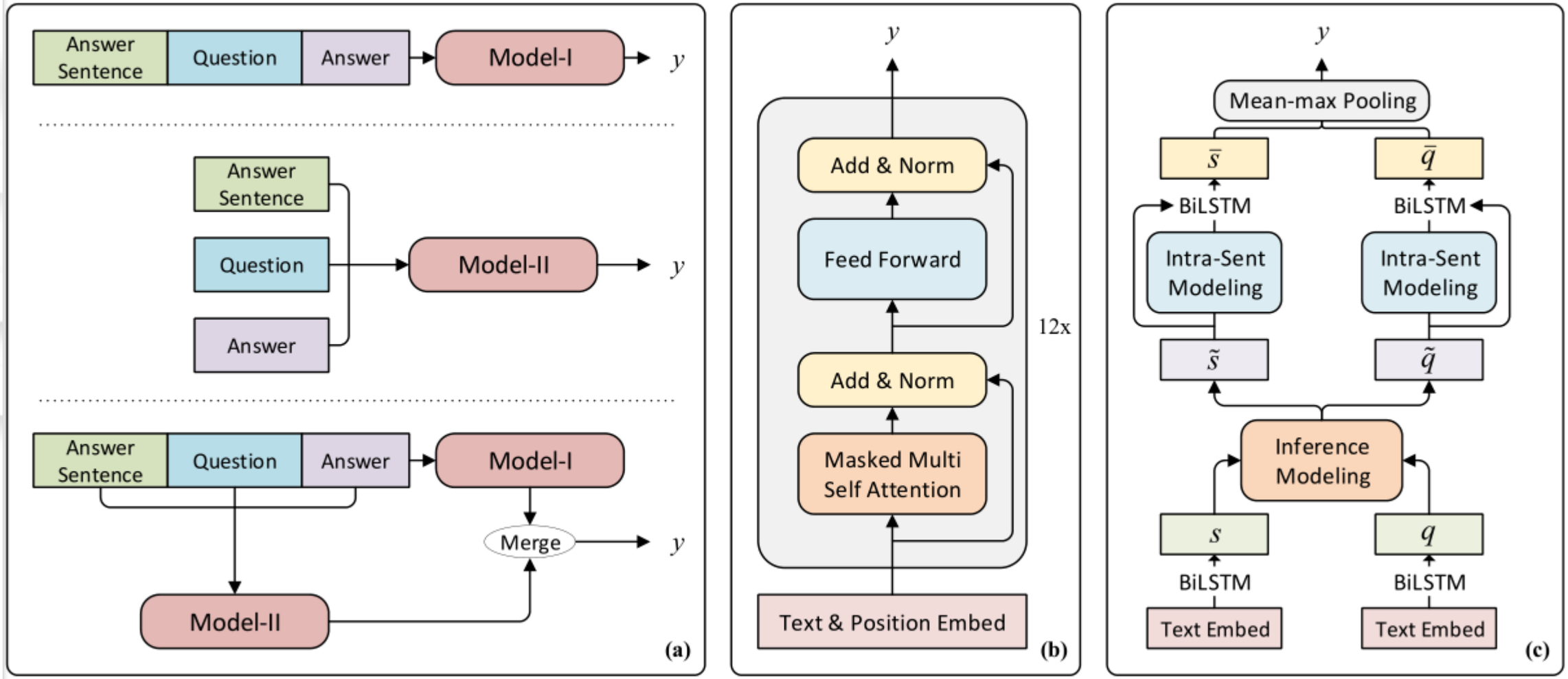
Recognize local textual entailment by comparing the answer sentence and the question, with three different architectures:

1. a *sequential model* that takes the inputs as a long sequence
2. an *interactive model* that encodes two sentences interdependently
3. a *hybrid model* that takes both of the two approaches into account

Answer Verifier

Recognize local textual entailment by comparing the answer sentence and the question, with three different architectures:

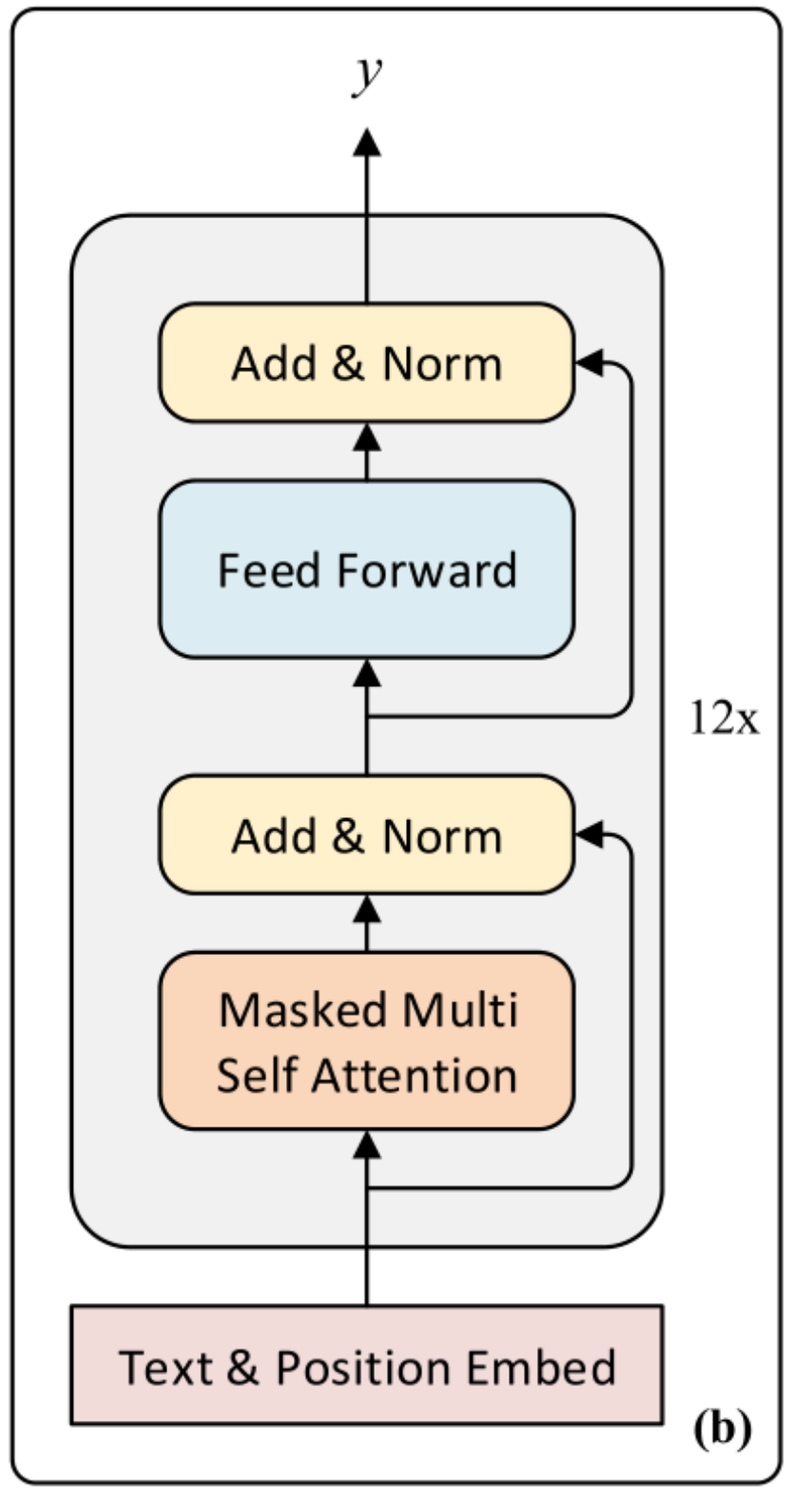
1. a *sequential model* that takes the inputs as a long sequence
2. an *interactive model* that encodes two sentences interdependently
3. a *hybrid model* that takes both of the two approaches into account



Sequential Model

The Finetuned Transformer model that

- first trained with a LM objective
- then finetuned on the target task
- consists of 12 transformer blocks



(Radford et al., 2018)

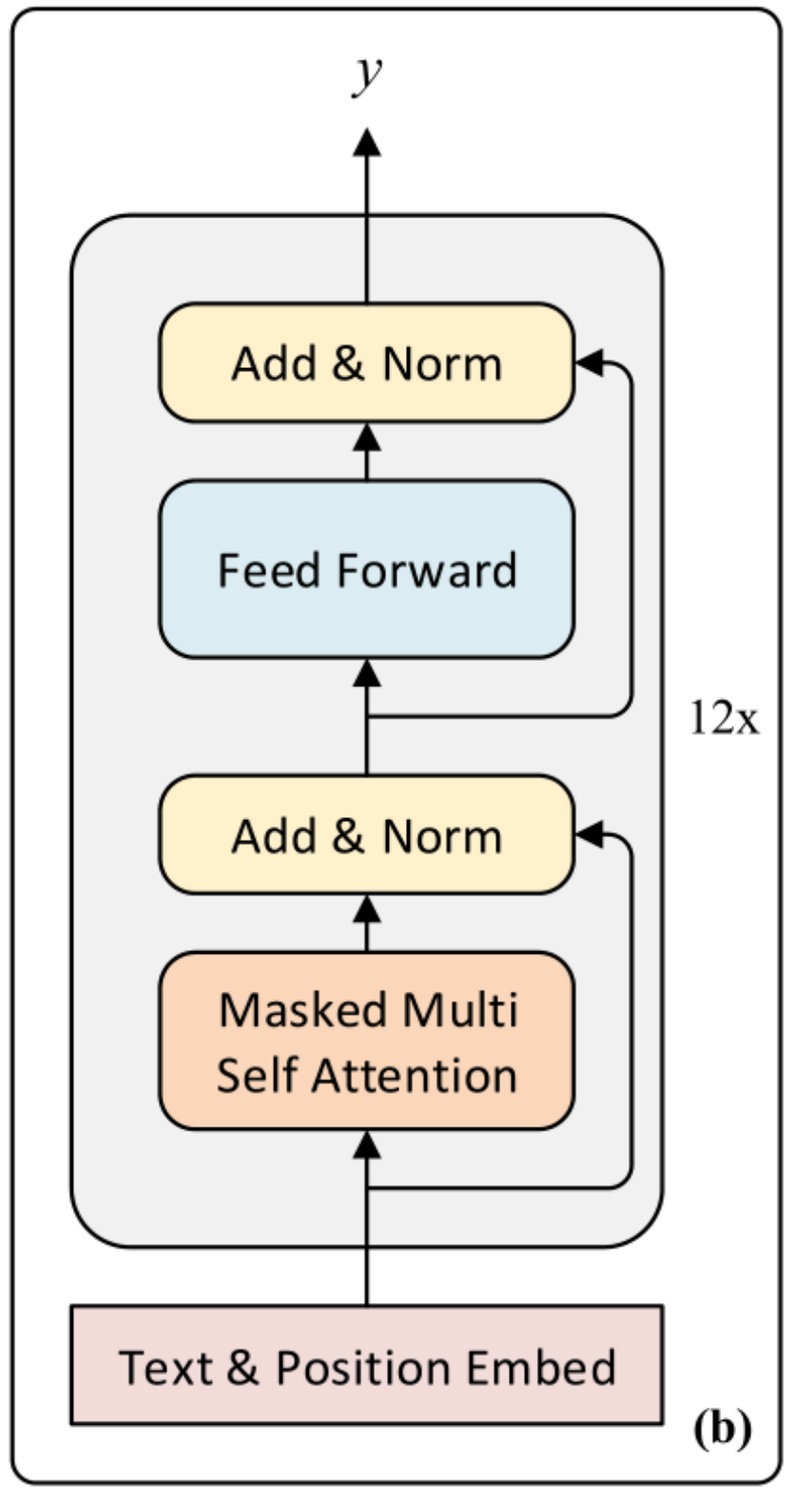
Sequential Model

The Finetuned Transformer model that

- first trained with a LM objective
- then finetuned on the target task
- consists of 12 transformer blocks

The last token's activation h is used to produce the no-answer probability y:

$$p(y|X) = \text{softmax}(h_n^{l_m} W_y)$$



(Radford et al., 2018)

Sequential Model

The Finetuned Transformer model that

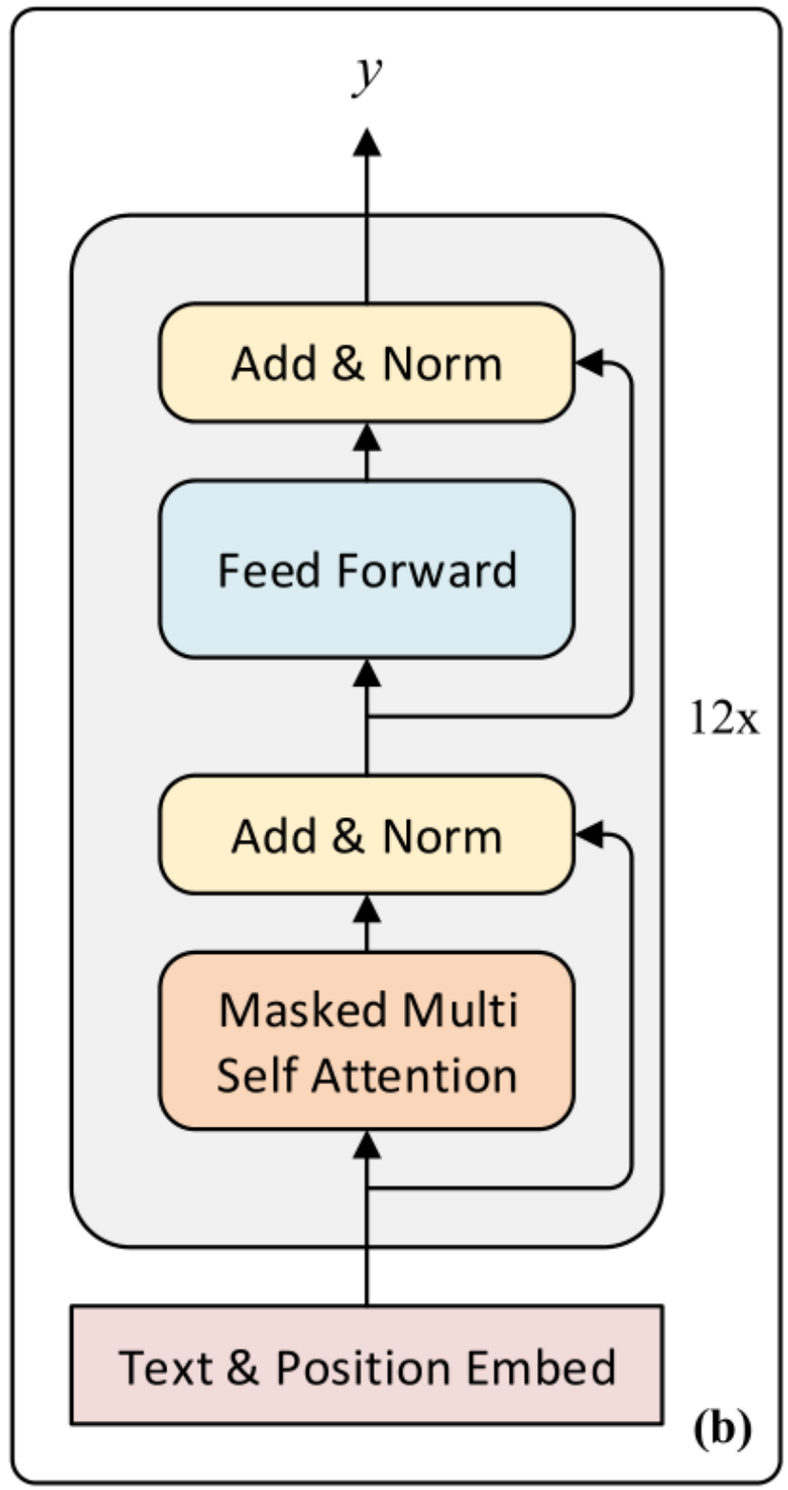
- first trained with a LM objective
- then finetuned on the target task
- consists of 12 transformer blocks

The last token's activation h is used to produce the no-answer probability y:

$$p(y|X) = \text{softmax}(h_n^{l_m} W_y)$$

The cross-entropy loss is used as

$$\mathcal{L}(\theta) = -\sum_{(X,y)} \log p(y|X)$$



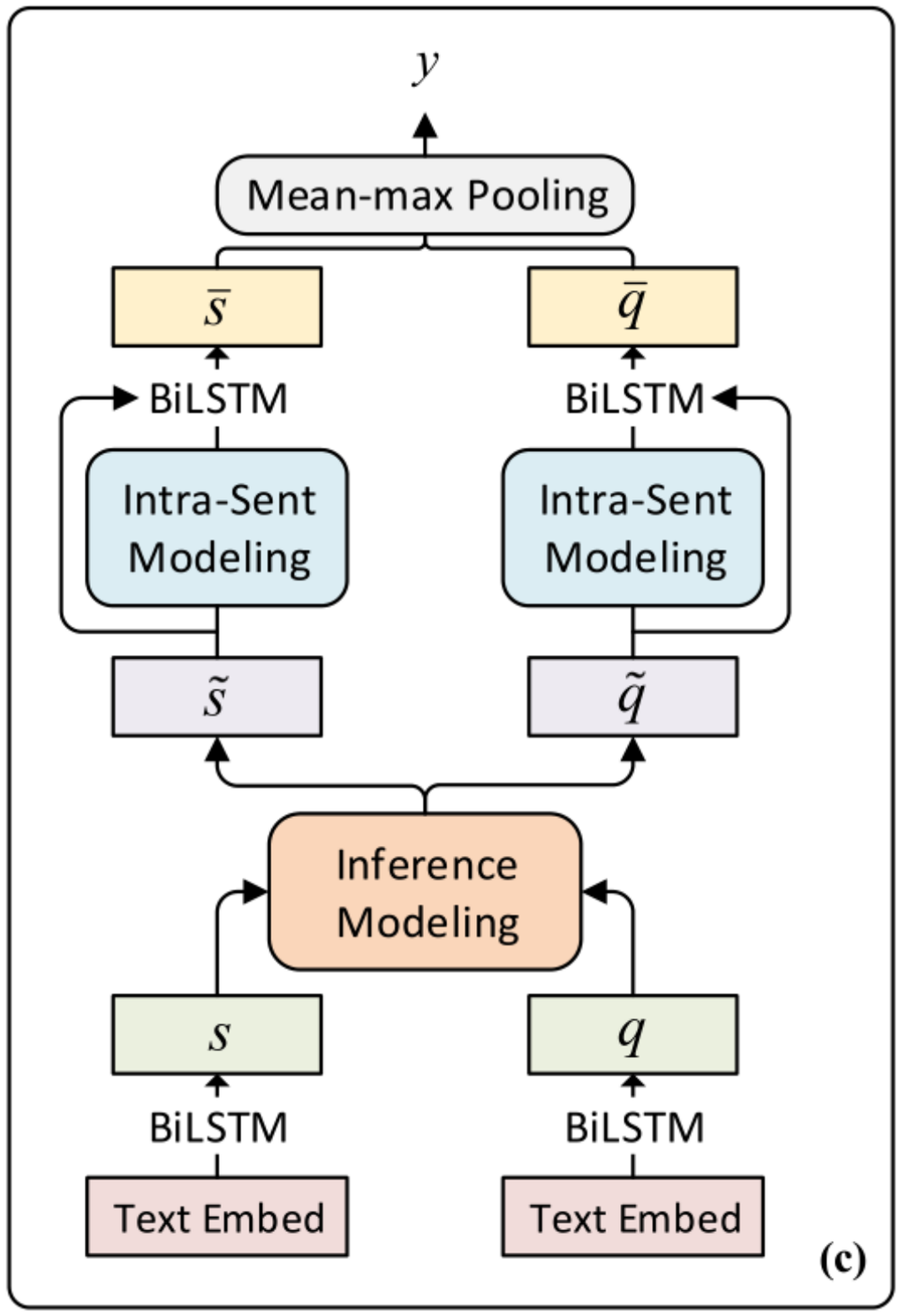
(Radford et al., 2018)

Interactive Model

Modeling inferences with the following layers:

- Encoding

$$s_i = \text{BiLSTM}([\text{word}_i^s; \text{char}_i^s; \text{fea}_i^s]), \forall i \in [1, l_s]$$
$$q_j = \text{BiLSTM}([\text{word}_j^q; \text{char}_j^q; \text{fea}_j^q]), \forall j \in [1, l_q]$$



Interactive Model

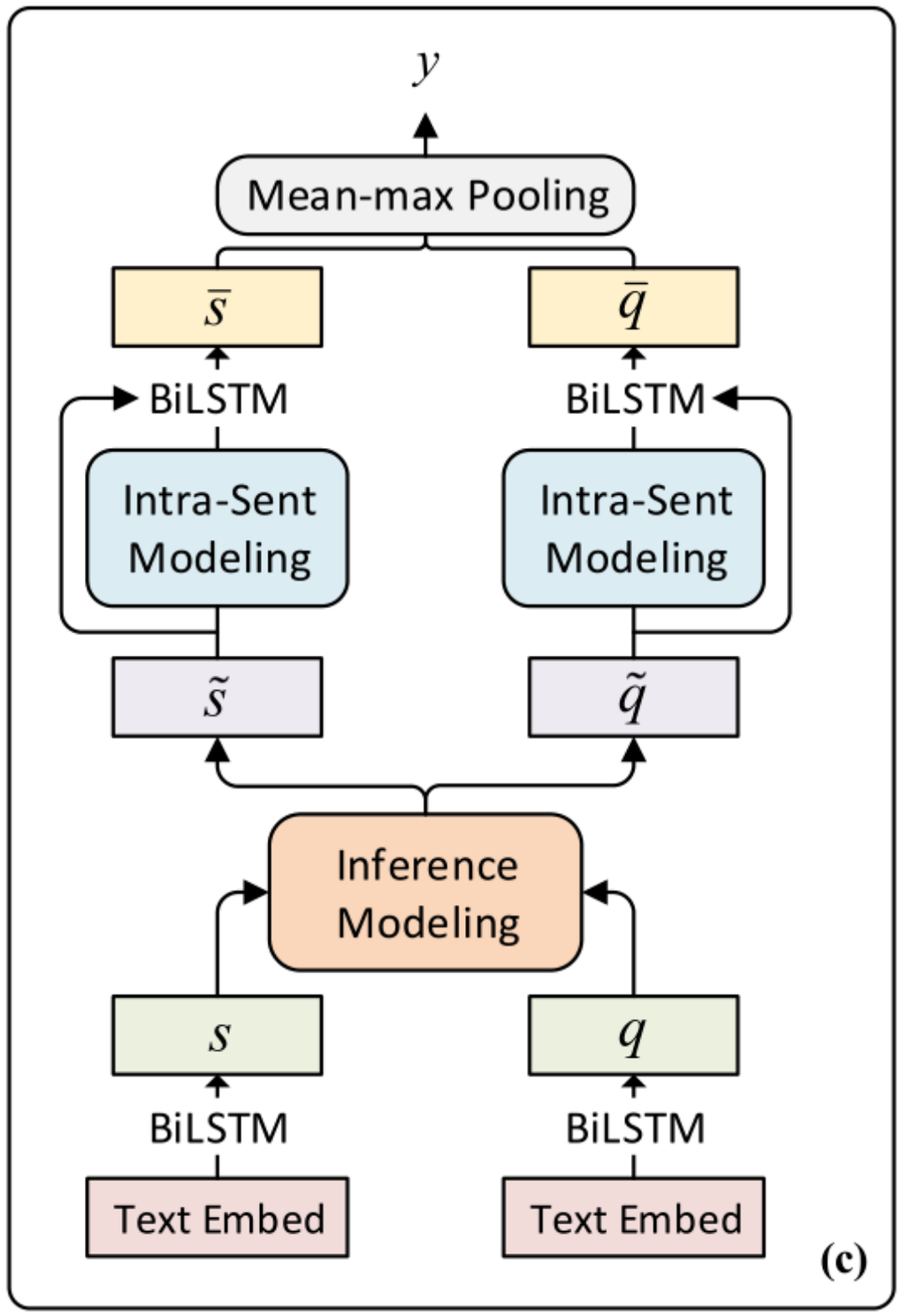
Modeling inferences with the following layers:

- Encoding

$$s_i = \text{BiLSTM}([\text{word}_i^s; \text{char}_i^s; \text{fea}_i^s]), \forall i \in [1, l_s]$$
$$q_j = \text{BiLSTM}([\text{word}_j^q; \text{char}_j^q; \text{fea}_j^q]), \forall j \in [1, l_q]$$

- Inference Modeling

$$a_{ij} = s_i^{\mathsf{T}} q_j, \forall i \in [1, l_s], \forall j \in [1, l_q]$$
$$b_i = \sum_{j=1}^{l_q} \frac{e^{a_{ij}}}{\sum_{k=1}^{l_q} e^{a_{ik}}} q_j \ , \ c_j = \sum_{i=1}^{l_s} \frac{e^{a_{ij}}}{\sum_{k=1}^{l_s} e^{a_{kj}}} s_i$$
$$\tilde{s}_i = F(s_i, b_i) \ , \ \tilde{q}_j = F(q_j, c_j)$$



Interactive Model

Modeling inferences with the following layers:

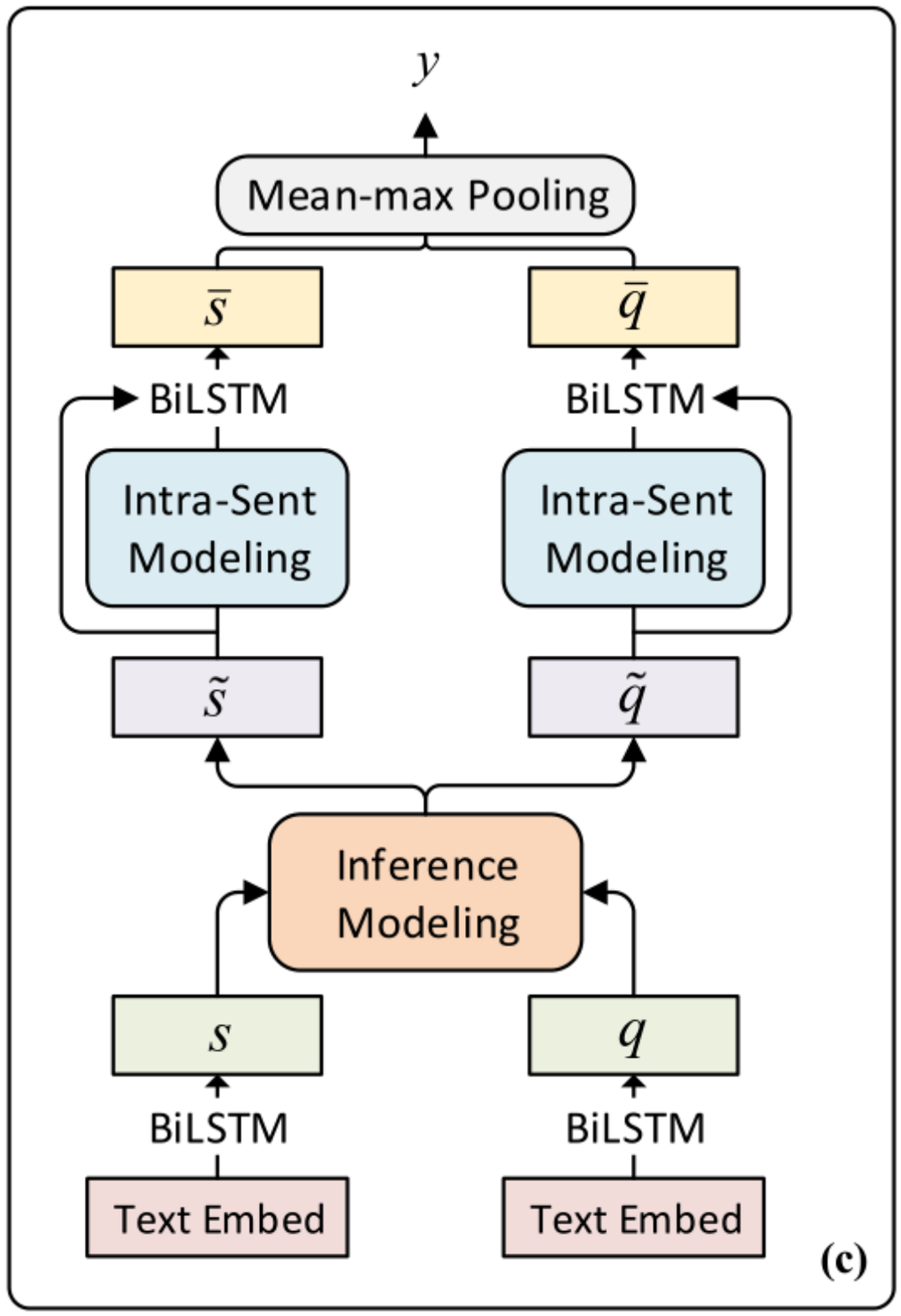
- Encoding

$$s_i = \text{BiLSTM}([\text{word}_i^s; \text{char}_i^s; \text{fea}_i^s]), \forall i \in [1, l_s]$$
$$q_j = \text{BiLSTM}([\text{word}_j^q; \text{char}_j^q; \text{fea}_j^q]), \forall j \in [1, l_q]$$

- Inference Modeling

$$a_{ij} = s_i^{\mathsf{T}} q_j, \forall i \in [1, l_s], \forall j \in [1, l_q]$$
$$b_i = \sum_{j=1}^{l_q} \frac{e^{a_{ij}}}{\sum_{k=1}^{l_q} e^{a_{ik}}} q_j \ , \ c_j = \sum_{i=1}^{l_s} \frac{e^{a_{ij}}}{\sum_{k=1}^{l_s} e^{a_{kj}}} s_i$$
$$\tilde{s}_i = F(s_i, b_i) \ , \ \tilde{q}_j = F(q_j, c_j)$$

- Intra-Sentence Modeling



Interactive Model

Modeling inferences with the following layers:

- Encoding

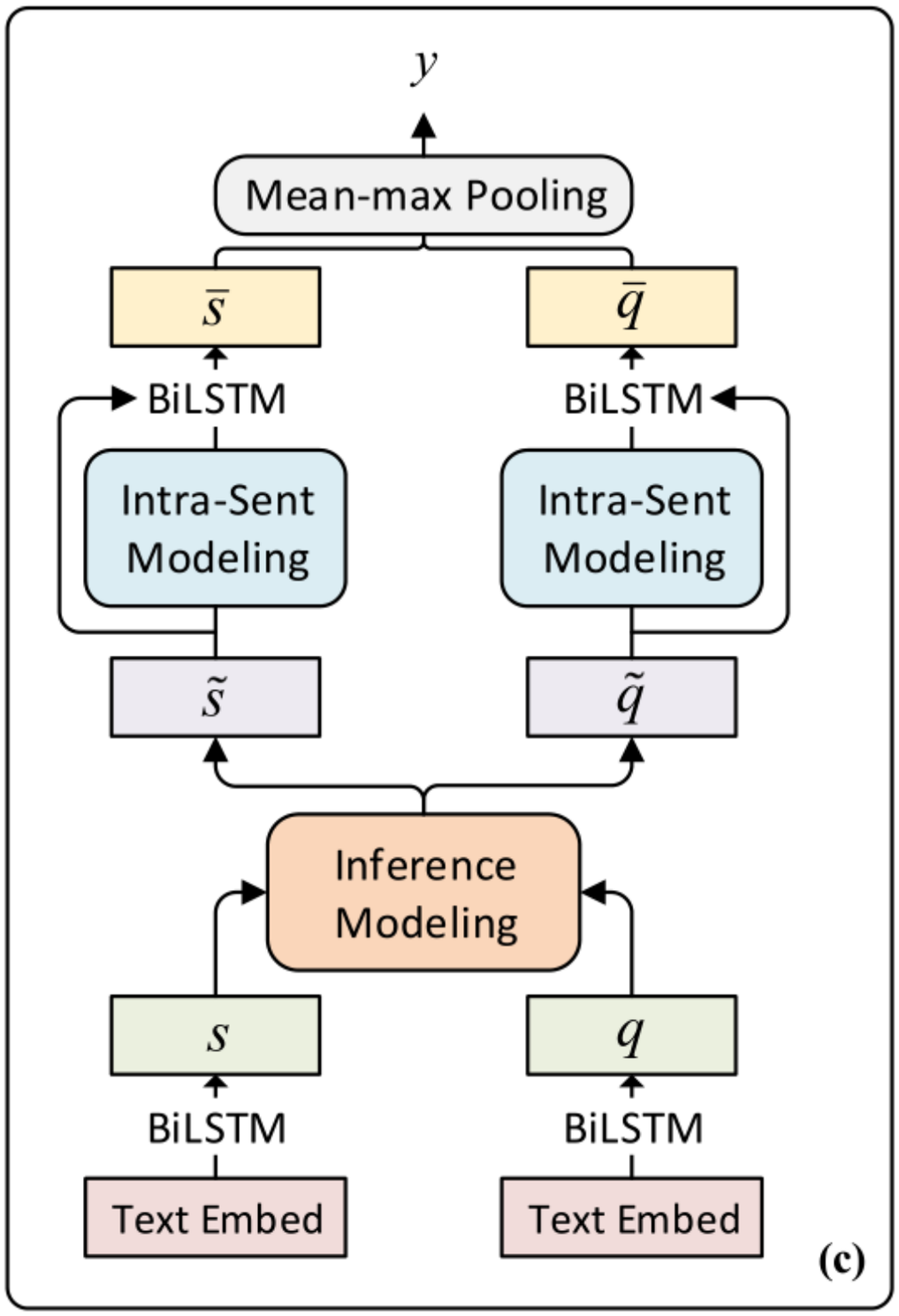
$$s_i = \text{BiLSTM}([\text{word}_i^s; \text{char}_i^s; \text{fea}_i^s]), \forall i \in [1, l_s]$$
$$q_j = \text{BiLSTM}([\text{word}_j^q; \text{char}_j^q; \text{fea}_j^q]), \forall j \in [1, l_q]$$

- Inference Modeling

$$a_{ij} = s_i^{\mathsf{T}} q_j, \forall i \in [1, l_s], \forall j \in [1, l_q]$$
$$b_i = \sum_{j=1}^{l_q} \frac{e^{a_{ij}}}{\sum_{k=1}^{l_q} e^{a_{ik}}} q_j \ , \ c_j = \sum_{i=1}^{l_s} \frac{e^{a_{ij}}}{\sum_{k=1}^{l_s} e^{a_{kj}}} s_i$$
$$\tilde{s}_i = F(s_i, b_i) \ , \ \tilde{q}_j = F(q_j, c_j)$$

- Intra-Sentence Modeling

- Predicting



Hybrid Model

Combine sequential model with interactive model

- Merge the output vectors of two models with a simple concatenation
- Use a unified feed-forward classifier to output no-answer probability
- Initialize with pre-trained parameters from two models, and then finetune on the target task

Evaluation

SQuAD2.0
The Stanford Question Answering Dataset

- The Stanford Question Answering Dataset 2.0 (SQuAD 2.0)
  - MRC benchmark augmented with unanswerable questions
  - 53,775 negative examples with the same passages
  - the passage contains plausible answers for these questions
  - the unanswerable questions are relevant to the passages

- Use Reinforced Mnemonic Reader as the base model

Evaluation

Our approach outperforms previous approaches at the time of submission

SQuAD 2.0	Test EM	Test F1
BNA	59.2	62.1
DocQA	59.3	62.3
VS^3-Net	68.4	71.3
SAN	68.6	71.4
FusionNet++ (*ensemble*)	70.3	72.6
Reinforced Mnemonic Reader + Answer Verifier	**71.7**	**74.2**

(Results extracted form https://rajpurkar.github.io/SQuAD-explorer/ on Aug 23, 2018)

Evaluation

Ablation study on readers with different auxiliary losses

Configuration	HasAns EM	HasAns F1	All EM	All F1	NoAns ACC
Reinforced Mnemonic Reader	79.4	86.8	71.4	73.7	77.0
- Independent Span Loss	78.9	86.5	71.2	73.5	76.7
- Independent No-Answer Loss	79.5	86.6	69.4	71.4	75.1
- Both Losses	78.7	86.2	70.0	71.9	75.3

Evaluation

Ablation study on different answer verifiers

Configuration	NoAns ACC
Sequential Model	74.5
Interactive Model	74.6
Interactive Model + ELMo	75.3
Hybrid Model	**76.2**
Hybrid Model + ELMo	76.1

Configuration	All EM	All F1	NoAns ACC
Reinforced Mnemonic Reader	71.4	73.7	77.0
+ Sequential Model	71.8	74.4	77.3
+ Interactive Model	71.8	74.2	78.1
+ Interactive Model + ELMo	72.0	74.3	78.2
+ Hybrid Model	**72.3**	**74.8**	**78.6**
+ Hybrid Model + ELMo	71.8	74.3	78.3

Conclusion

- Limitations of current extractive MRC models
  1. *Efficiency*: although effective, the ensemble models are not efficient
  2. *Robustness*: current approaches are vulnerable to adversarial sentences
  3. *Robustness*: existing models are easily fooled by unanswerable questions

Conclusion

- Limitations of current extractive MRC models
  1. *Efficiency*: although effective, the ensemble models are not efficient
  2. *Robustness*: current approaches are vulnerable to adversarial sentences
  3. *Robustness*: existing models are easily fooled by unanswerable questions

- Our contributions
  1. *Attention-guided answer distillation* that compresses the ensemble model into a single model without losing performances
  2. *Read-then-verify system* that validates the question via answer verification

Thanks! Questions?