MC^2: Multi-perspective Convolutional Cube for Conversational Machine Reading Comprehension

Xuanyu Zhang
College of Information Science and Technology
Beijing Normal University, Beijing, 100875, China
xyz@mail.bnu.edu.cn

Abstract

Conversational machine reading comprehension (CMRC) extends traditional single-turn machine reading comprehension (MRC) by multi-turn interactions, which requires machines to consider the history of conversation. Most of models simply combine previous questions for conversation understanding and only employ recurrent neural networks (RNN) for reasoning. To comprehend context profoundly and efficiently from different perspectives, we propose a novel neural network model, Multi-perspective Convolutional Cube (MC^2). We regard each conversation as a cube. 1D and 2D convolutions are integrated with RNN in our model. To avoid models previewing the next turn of conversation, we also extend causal convolution partially to 2D. Experiments on the Conversational Question Answering (CoQA) dataset show that our model achieves state-of-the-art results.

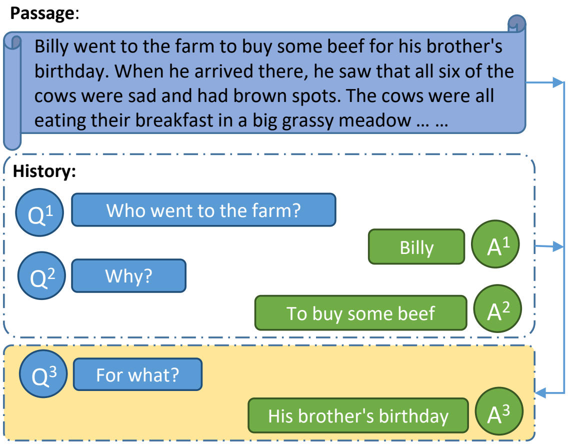
对话式MRC，需要考虑对话历史.

深度地

1 Introduction

Conversation is one of the most important approaches for humans to acquire information. Different from traditional machine reading comprehension (MRC), conversational machine reading comprehension (CMRC) requires machines to answer multiple follow-up questions according to a passage and dialogue history. However, these questions usually have complicated linguistic phenomena, such as co-reference, ellipsis and so on. Only considering conversation context profoundly can we answer the current question correctly.

Recently, many CMRC datasets, such as CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018), are proposed to enable models to understand passages and answer questions in dialogue. Here is an example from the CoQA dataset in Figure 1. We can observe that the second and third questions omit key information. It is impossible for both humans and machines to understand such questions without dialogue history.



Passage:

Billy went to the farm to buy some beef for his brother's birthday. When he arrived there, he saw that all six of the cows were sad and had brown spots. The cows were all eating their breakfast in a big grassy meadow ... ...

Figure 1: An example in the CoQA dataset.

Most of existing methods consider conversation history by prepending previous questions and answers to the current question, such as BiDAF++ (Yatskar, 2019), DrQA+PGNet (Reddy et al., 2019), SDNet (Zhu et al., 2018) and so on. However, the latent semantic information of dialogue history is neglected. And the model may confuse some unrelated questions and answers in a sentence. Although FlowQA (Huang et al., 2019) utilizes intermediate representations of previous conversation, the flow mechanism can not synthesize the information of different words in different turns of conversation simultaneously. Moreover, previous models only use recurrent neural network (RNN) as their main skeleton, which is not parallel due to recurrent nature. And RNN can only grasp information from two directions, either forward or backward. But for conversation, humans usually consider history from different perspectives and answer questions comprehensively.

To address these issues, we propose a novel model, i.e. **M**ulti-perspective **C**onvolutional **C**ube (MC^2). Every conversation is represented as a

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6185–6190
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

三个模块

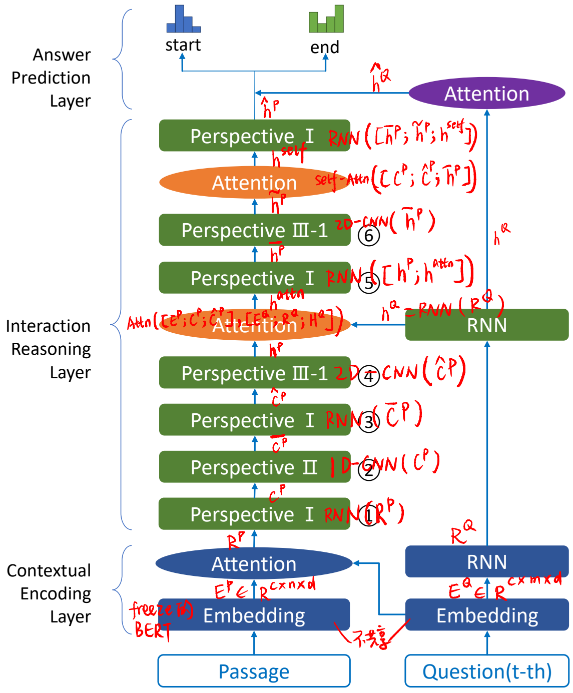


Figure 2: MC^2 structure overview.

cube, three dimensions of which are question answering (QA) turns, passage words and hidden states of words, separately. For one thing, convolutional neural networks (CNN) can extract local information effectively across dimensions in parallel. Introducing CNN to RNN allows the model to take into account local and global features efficiently. For another thing, machines can comprehend conversation history more deeply from different perspectives by fusing 1D and 2D convolutions in our model. In addition, to avoid information leakage of the next turn of dialogue, we extend causal convolution to 2D. Experiments on the Conversational Question Answering (CoQA) dataset show that our model improves the result of the published state-of-the-art model by 3.2%.

2 Approaches

In this section, we propose our novel model, MC^2, for the task of conversational machine reading comprehension, which can be formulated as follows. [For one conversation, given a passage with n tokens $P = \{p_i\}_{i=1}^n$ and multiple questions with c turns $Q = \{Q^t\}_{t=1}^c$, machines need to give the corresponding answers $A = \{A^t\}_{t=1}^c$. The t-th question with m tokens is $Q^t = \{q_j^t\}_{j=1}^m$. The neural network is required to model the probability distribution $p(A^t|Q^{\leq t}, P)$ for the t-th QA turn in the conversation.] As shown in Figure 2, there are three main layers in our model, i.e., contextual encoding layer, interaction reasoning layer and answer prediction layer. Our proposed cube is used in the middle layer. For convenience, we will illustrate our model from bottom to top.

问题描述

三个模块

2.1 Contextual Encoding Layer

The purpose of this layer is to extract useful information for upper layers. We embed questions and passages into a sequence of vectors with the latest contextualized model, BERT (Devlin et al., 2019), separately. Instead of fine-tuning BERT with extra scoring layers, we fix the weights of BERT like SDNet (Zhu et al., 2018) and aggregate L hidden layers generated by BERT as contextualized embedding for all BPE (Sennrich et al., 2016) tokens.

固定参数的 BERT.(不共享)

To introduce other linguistic features token by words and facilitate answer selection, we choose the first token of a word in BPE to represent the word. Generally, the first token is often the root of the word and can represent main meaning of the whole word. And it also contains information of rest tokens in the word with the bidirectional structure of BERT. Besides, we split the long sentence by shorter windows and combine them again when the sentence exceeds the maximum length of pre-trained BERT.

第一个piece作为词的代表.

对长句子进行切分及合并

第i个词的第1个piece的第l层表达

In detail, suppose $h_i^l \in \mathbb{R}^d$ is the l-th hidden layer of the first BPE token in the i-th word. We collapse all hidden layers generated by BERT into a single vector for each word following ELMo (Peters et al., 2018). The contextualized embedding for the i-th word is $e_i = \gamma \sum_{l=0}^{L} \alpha_l h_i^l$, where γ is designed to scale the vector and α_l is softmax-normalized weight for the l-th layer. These weights are all trainable. To be consistent with the number of turns of question $E^Q \in \{e_{t,j}^Q\}_{j=1}^m{}_{t=1}^c \in \mathbb{R}^{c\times m\times d}$, the passage e_i^P is expanded c times to $E^P = \{e_{t,i}^P\}_{i=1}^n{}_{t=1}^c \in \mathbb{R}^{c\times n\times d}$.

可学习的参数

c×m×d

每次对话序列长度

历史对话轮数

把passage的表达重复c次.

To incorporate other linguistic information, three additional features are utilized for each word p_i in the passage following Chen et al. (2017), i.e. part-of-speech (POS) tags, named entity recognition (NER) tags and aligned question embeddings. The embeddings of POS e_i^{pos} and NER e_i^{ner} are learned for different tags, separately. And aligned question embeddings can be obtained in Eq. 1. Following Huang et al. (2018), we use $f(x, y) = \mathrm{ReLU}(Ux)^{\mathrm{T}} D\mathrm{ReLU}(Uy)$ as the attention score function between x, y, where D is a diagonal matrix and D, U are trainable.

对passage中的每个词,加额外三种特征

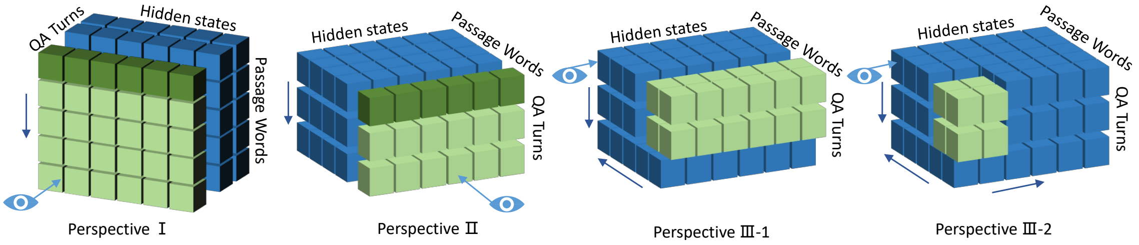


Figure 3: Different perspectives of the cube.

$f(x,y) = \mathrm{ReLU}(Ux) \cdot D \cdot \mathrm{ReLU}(Uy)$

第t轮第i个p词

第t轮第j个q词

$$s^i_j = f(e^P_{t,i}, e^Q_{t,j})$$
$$a^i_j = exp(s^i_j) / \sum_{k=1}^{m} exp(s^i_k) \in R \quad (1)$$
$$e^{attn}_{t,i} = \sum_{j=1}^{m} a^i_j e^Q_{t,j} \in R^d$$

We then concatenate these features and embeddings to $r^P_{t,i}$ for passages and employ bidirectional RNN to refine the question to $r^Q_{t,j}$.

$$r^P_{t,i} = [e^P_{t,i}; e^{pos}_{t,i}; e^{ner}_{t,i}; e^{attn}_{t,i}]$$
$$r^Q_{t,j} = \mathrm{BiRNN}(r^Q_{t,j-1}, e^Q_{t,j}) \quad (2)$$

2.2 Interaction Reasoning Layer

This layer plays an important role in our model, which aims to incorporate question information into passage representation further and reason from different perspectives by our proposed convolutional cube. The cube represents the hidden states of passages in a conversation. We will describe these perspectives in Figure 3 in the order of ① to ⑥ in Figure 2. To consider global context of each turn besides local information across different dimensions, *Perspective I* equipped with RNN is inserted before other CNN perspectives.

We first observe the cube from *Perspective I* and feed the hidden states of the cube $r^P_{t,i}$ to bidirectional RNN for each turn of conversation $c^P_{t,i} = \mathrm{BiRNN}(c^P_{t,i-1}, r^P_{t,i})$. Then the cube is viewed from *Perspective II* along QA turns for different words, separately. Since the $(t+1)$-th turn of information can not be used when processing the t-th turn, we employ 1D causal convolution (Oord et al., 2016) to the cube by moving the padding at the end to the beginning. And the representation of the cube can be updated from $c^P_{t,i}$ into $\bar{c}^P_{t,i}$ After viewed from these two perspectives (① ② in Figure 2), the hidden states of every word in passages grasp information from two dimensions of the cube.

LSTM

对每一轮，进行 BiRNN.

???

Next, we observe the cube from *Perspective I* again to fuse previous hidden states and generate global context $\hat{c}^P_{t,i}$ for each turn of conversation. To reason from more dimensions simultaneously, 2D CNN is utilized to generate hidden states of the cube $h^P_{t,i}$ along the dimension of both QA turns and passage words from *Perspective III-1*. Different from other models, three kinds of information can be considered comprehensively by this process: the same word in different QA turns, different words in the same QA turn and different words in different QA turns. Similar to 1D CNN above, the 2D CNN also requires to be unidirectional on the dimension of QA turns to avoid information leakage. But it is more reasonable to capture bidirectional information on the dimension of passage words. We thus extend traditional causal convolution partially to 2D CNN by moving padding only on one dimension. These two perspectives (③ ④ in Figure 2) strengthen the representation of our cube further.

2D causal Convolution

For questions in this layer, we pass them as the input to another RNN for reasoning $h^Q_{t,j} = \mathrm{BiRNN}(h^Q_{t,j-1}, r^Q_{t,j})$. Then we employ the attention score function mentioned above to integrate new information of questions to passages.

对question

attention

$$s^i_j = f([e^P_{t,i}; c^P_{t,i}; \hat{c}^P_{t,i}], [e^Q_{t,j}; r^Q_{t,j}; h^Q_{t,j}])$$
$$a^i_j = exp(s^i_j) / \sum_{k=1}^{m} exp(s^i_k) \quad (3)$$
$$h^{attn}_{t,i} = \sum_{j=1}^{m} a^i_j h^Q_{t,j}$$

As shown in Figure 2, we repeat the process of ③ ④ in ⑤ ⑥ for deeper understanding and reasoning. RNN takes $[h^P_{t,i}; h^{attn}_{t,i}]$ and generates $\bar{h}^P_{t,i}$ from *Perspective I*. Then 2D CNN generates $\tilde{h}^P_{t,i}$ from *Perspective III-1*. We use self-attention to enhance the current passage representation as follows:

$$s^i_j = f([c^P_{t,i}; \hat{c}^P_{t,i}; \bar{h}^P_{t,i}], [c^P_{t,j}; \hat{c}^P_{t,j}; \bar{h}^P_{t,j}])$$
$$a^i_j = exp(s^i_j) / \sum_{k=1}^{n} exp(s^i_k) \quad (4)$$
$$h^{self}_{t,i} = \sum_{j=1}^{n} a^i_j \bar{h}^P_{t,j}$$

Model	In-domain					Out-of-domain		Overall
	Child.	Liter	Mid-High.	News	Wiki	Reddit	Science	
PGNet	49.0	43.3	47.5	47.5	45.1	38.6	38.1	44.1
DrQA	46.7	53.9	54.1	57.8	59.4	45.0	51.0	52.6
DrQA+PGNet	64.2	63.7	67.1	68.3	71.4	57.8	63.1	65.1
Augmt. DrQA	66.0	63.3	66.2	71.0	71.3	57.7	63.0	65.4
BiDAF++	66.5	65.7	70.2	71.6	72.6	60.8	67.1	67.8
FlowQA	73.7	71.6	76.8	79.0	80.2	67.8	76.1	75.0
SDNet	75.4	73.9	77.1	80.3	83.1	69.8	76.8	76.6
MC^2	**78.4**	**76.7**	**81.1**	**83.0**	**84.8**	**73.8**	**80.6**	**79.8**
Human	90.2	88.4	89.8	88.6	89.9	86.7	88.1	88.8

Table 1: Model and human performance (% in F1 score) on the CoQA test set.

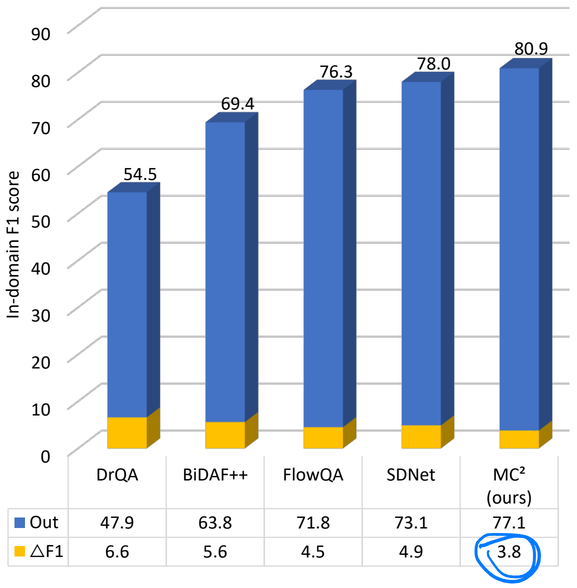


Figure 4: F1 score of models on in-domain and out-of-domain parts of the CoQA test set.

At last, we view the cube from *Perspective I* again to synthesize the global information $\hat{h}^P_{t,i} = \text{BiRNN}(\hat{h}^P_{t,i-1}, [\bar{h}^P_{t,i}; \tilde{h}^P_{t,i}; h^{self}_{t,i}])$.

2.3 Answer Prediction Layer

This layer is the top one of our model. We use similar methods (Chen et al., 2017; Huang et al., 2019; Zhu et al., 2018) to predict the position of the answer in the passage. We project the question representation into one vector for each turn of dialogue $\hat{h}^Q_t = \sum_{j=1}^{m} a_{t,j} h^Q_{t,j}$, where $a_{t,j} = exp(W h^Q_{t,j}) / \sum_{k=1}^{m} exp(W h^Q_{t,k})$ and W is trainable. Then two different bilinear attention functions are used to estimate the probability of the start and end according to $\hat{h}^P_{t,i}$ and $\hat{h}^Q_t$. We choose the position of the maximum product of these two probabilities as the best span. For other answer types, such as *yes*, *no* and *unknown*, we condense the passage representation $\hat{h}^P_{t,i}$ to $\hat{h}^P_t$ like questions and classify the answer according to $[\hat{h}^P_t; \hat{h}^Q_t]$.

每轮的Q转成一个定长向量

双线性attention计算起止位置概率

起止位置乘积最大的为预测span

???

对其他类型的答案

To train the cube, we minimize the sum of the negative log probabilities of the ground truth start position, end position and answer type by the predicted distributions.

loss

3 Experiments

3.1 Data and Metric

We conduct our experiments on the CoQA (Reddy et al., 2019), a large-scale CMRC dataset annotated by human. It consists of 127k questions with answers collected from 8k conversations over text passages. As shown in Table 1, it covers seven diverse domains (five of them are in-domain and two are out-of-domain). The out-of-domain passages only appear in the test set. Aligned with the official evaluation, F1 score is used as the metric, which measures the overlap between the prediction and the ground truth at word level.

3.2 Implementation Details

We use pre-trained BERT$_{\text{LARGE}}$ model for contextualized embeddings, the dimension of which is 1024. And spaCy is applied for tokenization, part-of-speech and named entity recognition. The last turn of the answer is added to the next turn as guidance in the dataset. Each batch contains one cube for one conversation. We employ LSTM as the structure of RNN, the hidden size of which is 250 throughout our model. The kernel size is set to 5 and 3 for 1D and 2D CNN, respectively. And the dropout rate is set to 0.4. The Adamax (Kingma and Ba, 2015) is used as our optimizer with 0.1 learning rate.

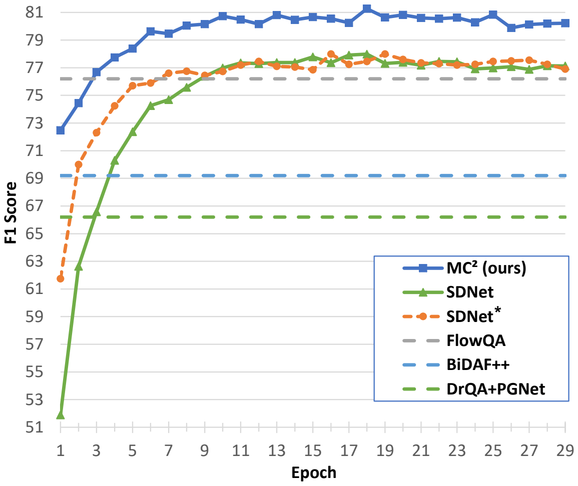


Figure 5: F1 score on the CoQA dev set under different training epochs. [1]

去掉一些 CNN层

1D 2D

1D→2D

2D→1D

2D-CNN换成3D-CNN

Configuration	F1	Δ F1
MC^2	81.266	-
w/o ② ④ ⑥	77.363	-3.903
w/o ②	80.718	-0.548
w/o ④	80.867	-0.399
w/o ⑥	80.849	-0.417
replace ② with ④	80.932	-0.334
replace ④ with ②	80.473	-0.793
replace ⑥ with III-2	81.087	-0.179
exchange ② with ④	81.102	-0.164

Table 2: Ablation study on the CoQA dev set. (② ④ ⑥ come from Fig. 2. III-2 comes from Fig. 3.)

3.3 Result

We compare our MC^2 with other baseline models [2] in Table 1: PGNet (See et al., 2017), DrQA (Chen et al., 2017), DrQA+PGNet (Reddy et al., 2019), Augmented DrQA (Reddy et al., 2019), BiDAF++ (Yatskar, 2019), FlowQA (Huang et al., 2019) and SDNet (Zhu et al., 2018). Our model achieves significant improvement over these published models. Comparing with the previous state-of-the-art model, SDNet, our model outperforms it by 3.2% on F1 score. And SDNet also takes pre-trained BERT as embedding without fine-tuning. Especially, our single model surpasses the ensemble model of both FlowQA and SDNet.

Figure 4 shows the gap between in-domain and out-of-domain on the test set. Although all models perform worse on out-of-domain datasets compared to in-domain datasets, our model only drops 3.8% on F1 score. It is the smallest drop between in-domain and out-of-domain among all models, which proves that our model has very good generalization ability. Besides, our model achieves the best performance on both in-domain and out-of-domain datasets.

Figure4

泛化性好

The learning curve is shown in Figure 5. It reflects the performance of models under different training epochs on the development set. We can observe that our model completely surpasses SDNet at every epoch. And it outperforms all baseline models only after 5 epochs and achieves the best performance after 18 epochs. Especially, our model achieves 72.472% on F1 score only after the first epoch, which is about 10% to 20% higher than SDNet. Thus with fewer training epochs, our model still can perform well.

Figures

3.4 Ablation Studies

To study how each perspective of our proposed cube contributes to the performance, we conduct an ablation analysis on the development set in Table 2. The results show that removing all CNN perspectives of the cube, i.e. ② ④ ⑥ in Figure 2, will cause a substantial performance drop (3.90% on F1 score). And removing any of them also results in marginal decrease in performance. It is clear that the improvement of reading from different perspectives simultaneously is larger than that of the sum of reading from single perspective separately. Besides, replacing 2D CNN (*Perspective III-1*) with 1D CNN (*Perspective II*) also causes a significant decline of performance (0.79% on F1 score). We also explore 3D CNN (*Perspective III-2*), but it brings no improvement as expected.

4 Conclusion

In this paper, we introduce Multi-perspective Convolutional Cube (MC^2), a novel model for conversational machine reading comprehension. The cube is viewed from different perspectives to fully understand the history of conversation. By integrating CNN with RNN, fusing 1D and 2D convolutions, extending causal convolution to 2D, our model achieves the best results among published models on the CoQA dataset without fine-tuning BERT. We will study further the capability of our approaches on other datasets and tasks in the future work.

[1] SDNet comes from experiments of the original author. SDNet* refers to the proportion of Fig. 2 in the original paper.

[2] We only consider published models on the CoQA. Although some models perform better on the leaderboard recently, they usually focus on fine-tuning BERT model.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2019. FlowQA: Grasping flow in history for conversational machine comprehension. In *Proceedings of the 7th International Conference on Learning Representations*.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2018. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. In *International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *SSW*, 125.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Mark Yatskar. 2019. A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. SDNet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.