Exploring the Impacts of Character Modeling and CoAttention on BiDAF Performance

Sophia Andrikopoulos [1]

[1]Stanford Department of Computer Science

Introduction

Question-answering (QA) is a highly relevant area of focus in the Natural Language Processing (NLP) community. QA systems are both useful tools in themselves, and provide us with a better understanding of how well machines encode human language. Massive improvements have been made to QA models in recent years with the introduction of neural attention mechanisms. I focus on one such model, BiDirectional Attention Flow (BiDAF). While I do not propose any significant novel components, I explore the addition of character-level embeddings to BiDAF. Further, I examine the impacts of CoAttention, a secondary attention model, on BiDAF performance in order to better understand the functions of both BiDirectional and CoAttention and the interaction between the two. The models are evaluated on the Stanford Question Answering 2.0 (SQuAD 2.0) dataset.[3]

Related Work

This exploration interprets the following previous work in the QA domain:

- **Character embeddings** use a single convolutional layer and max-pooling to build character representations [1]
- Previously, QA models typically attended to small portions of the context with **uni-directional attention**
- **BiDirectional attention** represents direct query-to-context and context-to-query attention, but only performs a single pass [4]
- **CoAttention** attends over first-level bidirectional attention, outputting a second-level attention representation [5]

BiDAF Baseline

The baseline model is based on the BiDirectional Attention Flow model (BiDAF). BiDAF consists of a word-Embedding layer, an Encoder layer, a BiDirectional Attention layer, a Modeling layer, and an Output layer. The BiDirectional attention layer represents query-to-context and context-to-query attention. The original BiDAF model also includes character embeddings that are not included in the baseline, highlighted in green in Figure 1.

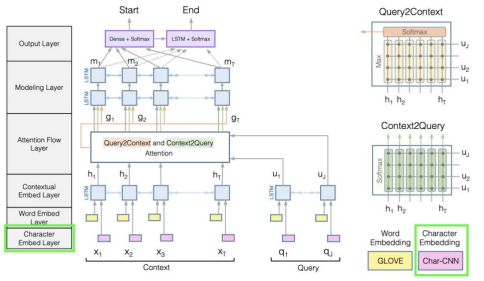

Figure 1. BiDirectional Attention Flow model. This paper's baseline model does not include the character-level embeddings highlighted in green. [4]

Approaches

1. Re-integrating Character embeddings to emulate original BiDAF model.[1],[4] **Character embeddings are used in further explorations.**
2. Replacing Bidirectional attention layer with a CoAttention layer described below.[5]
3. Creating the Dual Attention layer by concatenating Bidirectional and CoAttenion outputs.

CoAttention Layer

CoAttention includes a secondary attention computation that attends over the first-level attention output. For question hidden states $q_1, \ldots, q_M$ and context hidden states $c_1, \ldots, c_N$:

1. $q^{'=\tanh(Wq+b)}$
2. Randomly-initialized sentinel vectors are concatenated to c and $q^{'}$
3. Affinity matrix $\mathbf{L} = c^T q$ is computed
4. Context-to-Question attention distributions $\alpha = \text{softmax}(\mathbf{L}_{i,:})$
5. Context-to-Question attention outputs $a = \sum_{j=1}^{M+1} \alpha_j q_j$
6. Question-to-Context attention distributions $\beta = \text{softmax}(\mathbf{L}_{:,j})$
7. Question-to-Context attention outputs $b = \sum_{i=1}^{N+1} {}_i c_i$
8. $s = \sum_{j=1}^{M+1} \alpha_j b_j$, $[s; a][:N]$ fed through 2-layer GRU to obtain CoAttention output u

Results

- Character embeddings and Dual Attention outperform BiDAF baseline
- CoAttention outperforms BiDAF baseline but reduces performance of character embedding model with BiDirectional attention
- Dual Attention outperforms all other models explored

Figure 2. Model performance leveled out around Epoch 30 for all models.

	Dev		Test	
	F1	EM	F1	EM
BiDAF Baseline	60.412	56.898		
BiDAF with Character-Level Embeddings	63.445	60.208	62.490	58.969
BiDAF with CoAttention	62.012	58.578		
BiDAF with Dual Attention	**64.396**	**60.931**	**63.926**	**60.338**

Table 1. Model built off of the BiDAF baseline implemented in pytorch.[2] Performance on the SQuAD 2.0 dataset evaluated using both F1 and EM scores.[3]

Key Findings

1. Character embeddings improved baseline as expected [4]
   - Likely improve upon the model by capturing sub-word meanings, allowing the model to handle previously unseen words and therefore to generalize outside of the training set
2. CoAttention layer worsened performance of character embedding model
   - could be attributed to a poor implementation of the CoAttention layer
   - second-level attention output likely passed less useful information to future layers compared to BiDirectional Attention, indicating that first-level attention is an important aspect of BiDAF
3. Dual Attention model outperforms all other models explored
   - supports theory that information passed to future layers using first-level attention improves model performance
   - suggests that CoAttention layer was likely implemented correctly
   - suggests that CoAttention contains useful secondary information that supports BiDirectional attention output
   - layer size is double that of BiDirectional or Coattention, passing on significantly more information to future layers and improving model performance
   - addition of sentinel vectors may improve AvNA performance

Next Steps

This exploration highlights important attributes of the BiDAF model. Primarily, character embeddings allow the model to generalize outside of the training set to unseen words by capturing sub-word meaning, greatly improving BiDAF performance. Secondly, BiDirectional attention is an integral component of BiDAF–as the name suggests–because it provides future layers with important first-level context-to-question and question-to-context information. On its own, CoAttention negatively impacts BiDAF performance, likely because second-level attention does not carry enough information on its own to support the model. This is supported by the high performance of Dual Attention, suggesting the second-level information of CoAttention positively complements the more powerful BiDirectional attention. While I have improved upon the BiDAF baseline with character embeddings and Dual Attention, there is still great room for exploration and improvement within BiDAF and the SQuAD 2.0 question-answering task:

- exploring the effects of larger embeddings, or n-gram embeddings
- explore other attention models such as Self Attention by substituting or concattenating similar to Dual Attention
- with greater resources, hyperparameter tuning in order to maximize model performance

References

[1] Yoon Kim.
Convolutional neural networks for sentence classification.
volume abs/1408.5882, 2014.

[2] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala.
Pytorch: An imperative style, high-performance deep learning library.
In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[3] Pranav Rajpurkar, Robin Jia, and Percy Liang.
Know what you don't know: Unanswerable questions for SQuAD.
In *Association for Computational Linguistics (ACL)*, 2018.

[4] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi.
Bidirectional attention flow for machine comprehension.
volume abs/1611.01603, 2016.

[5] Caiming Xiong, Victor Zhong, and Richard Socher.
Dynamic coattention networks for question answering.
volume abs/1611.01604, 2016.