Multi-Hop QA with Bi-Attention Processing and CNNs

Laura Cruz-Albrecht, Krishna Patel // {lcruzalb, kpatel7}@stanford.edu
CS224N: Natural Language Processing with Deep Learning, Winter 2019

CS224N

Problem Statement

Question answering systems are an exciting but challenging application of Natural Language Processing.

While much work has been done on general QA, there is a lack of work in the realm of QA requiring **multi-hop reasoning**, where the QA system has to reason over information from multiple documents to generate an answer.

We aimed to create a **multi-hop QA model** that utilized novel architecture building blocks to improve upon the publicly available HotpotQA baseline.

Goal:	**Train a model that takes in a question requiring multi-hop reasoning, and context paragraphs, and outputs an answer + the supporting facts.**

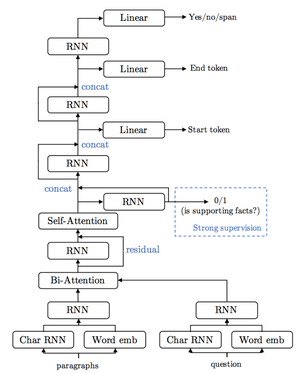
Data & Evaluation

Dataset: HotpotQA Dataset
Statistics: 89.8K train, 7.4K dev
Evaluation: F1, EM

Q: Mark Levenson toured the country with Stephen Colbert, Paul Dinello, and an actress that played what role in the series "Strangers with Candy"?
Context:
Paragraph 1: Distractor
[1] Wigfield: The Can Do Town That Just May Not is a satirical novel by comedians Amy Sedaris, Paul Dinello, and Stephen Colbert, three of the four creators of the Comedy Central show "Strangers with Candy." [2] It was first published on May 7, 2003 by Hyperion Books.
Paragraph 2: Gold
[3] Amy Louise Sedaris (born March 29, 1961) is an American actress, voice actress, singer, author, screenwriter and comedian. [4] **She is known for playing Jerri Blank in the Comedy Central television series "Strangers with Candy."** [5] She regularly collaborates with her older brother David, a humorist and author. [6] Since 2014, Sedaris has voiced the character Princess Carolyn in the Netflix animated series "BoJack Horseman."
Paragraph 3: Gold
…[6] Levenson composed music for David Sedaris's two Off Broadway shows and numerous recording projects. **[7] He recently toured the country with Stephen Colbert, Amy Sedaris and Paul Dinello in their production of Wigfield, which concluded its run at the U.S. Comedy Arts Festival in Aspen, Colorado.**
Supporting Facts: 4, 7
A: Jerri Blank

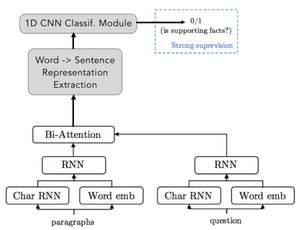
Approach

Existing HotpotQA Baseline



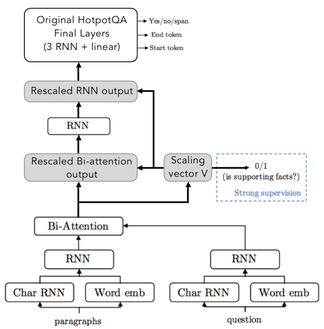
CNN Classification module

Independent module for supporting fact classification with 1D CNN



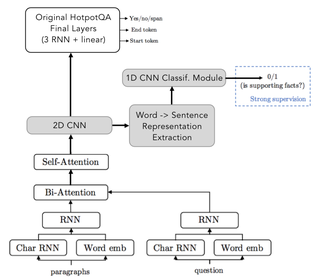
Processed Bi-Attention Model

Bi-attention post-processing + sigmoid/softmax for SP prediction and Q answering



2D CNN Model

Architecture leveraging word-level 2D CNN on self-attention output.



Experiments and Results

Table 2: Score comparison

Model	Split	Answer EM	Answer F1	Sup Fact EM	Sup Fact F1	Joint EM	Joint F1	Loss Overall	Loss Sup Fact
HotpotQA baseline	dev	44.44	58.28	21.95	66.66	11.56	40.86	-	-
	test	45.46	58.99	22.24	66.62	12.04	41.37	-	-
CNN classif. module	train	-	-	18.60	60.24	-	-	-	0.092
	dev	-	-	15.85	56.36	-	-	-	0.102
Bi-atten. + sigmoid	train	67.71	74.27	0	9.31	0	6.99	35.38	30.93
	dev	40.19	53.43	0	9.42	0	5.15	39.62	34.54
Integrated SP CNN	train	43.34	50.64	0	0	0	0	6.61	0.19
	dev	32.06	43.20	0	0	0	0	6.38	0.19
2D CNN, v1	train	79.96	85.54	3.49	15.69	2.91	13.87	3.62	0.17
	dev	40.78	53.77	2.59	14.17	1.24	8.30	6.03	0.17
2D CNN, v2	train	18.36	41.37	6.94	28.05	1.61	13.23	7.92	0.15
	dev	8.84	30.44	5.86	26.22	0.90	9.38	9.32	0.15

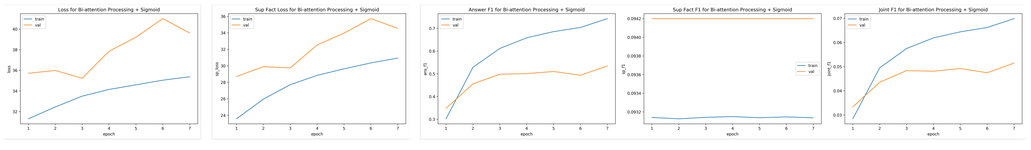
Qualitative analysis (2D CNN, v1)

- Rarely, the model correctly identified both the correct answer and all supporting facts
- Often, the model found the correct answer without identifying any suporting facts

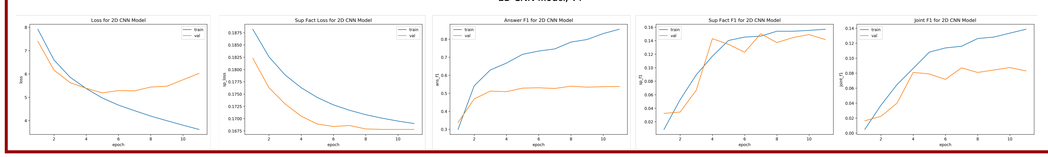
Q: The football manager who recruited David Beckham managed Manchester United during what timeframe?
A: from 1986 to 2013
Supporting Facts: [1] Their triumph was made all the more remarkable by the fact that Alex Ferguson had sold experienced players Paul Ince, Mark Hughes and Andrei Kanchelskis before the start of the season, and not made any major signings. [2] Instead, he had drafted in young players like Nicky Butt, David Beckham, Paul Scholes and the Neville brothers, Gary and Phil. [3] Sir Alexander Chapman Ferguson, CBE (born 31 December 1941) is a Scottish former football manager and player who managed Manchester United from 1986 to 2013.
Model Answer: 1986 to 2013
Model Supporting Facts: none identified

Figure 4.2: Often, the model found the correct answer without identifying any supporting facts.

Bi-attention Processing + Sigmoid Model



2D CNN Model, V1



Results

- CNN classification module trained in isolation performs comparable to baseline (though lower), lower performance when trained as part of model
- Bi-attention processing approach with sigmoid performs reasonably on QA task, but poorly on sup fact classification. Though F1 improves slightly over time, loss also diverges.
- 2D CNN, v1 best model (applies 2D CNN prior to SP classification rather than after)

Conclusions

CNNs seem to be a reasonable architectural building block for this task
Though we were not able to beat the HotpotQA baseline, our best model (2D CNN, v1) used a 2D CNN rather than an RNN, and attained lower but comparable overall Answer F1 / EM scores.

Explicit SP classification is not critical for the ultimate QA task
Despite having a low SP score, some models still had a high Answer F1 score, suggesting they were still able to identify supporting facts implicitly despite falling short in explicit identification.

Difficult to optimize SP classification with standard loss calculation
By utilizing a loss function such as CE loss, we end up minimizing loss by assigning no value to all of the sentences; this lowers the loss, but at the cost of rarely producing a true positive.

Future Work

- **Explore alternative techniques for combining question with context**. We currently use bi-attention to accomplish this; future work remains to try alternate methods.
- **Experiment with deeper 2D CNN layers**. We currently only use 1 layer; however, deep layers have proven effective in visual recognition and may also help with this task.
- **Hyperparameter tuning**. Experimenting with different hyperparameters for the CNN layers (among others) could improve performance.

References

[1] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016.

[2] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[3] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR*, abs/1705.03551, 2017.

[4] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

[5] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. *CoRR*, abs/1710.10723, 2017.

[6] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of CNN and RNN for natural language processing. *CoRR*, abs/1702.01923, 2017.

[7] Yoon Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.

[8] Yang et. al. Hotpotqa github. https://github.com/hotpotqa/hotpot.