Question Answering through Transfer Learning from Large Fine-grained Supervision Data

Sewon Min[1*], Minjoon Seo[2], Hannaneh Hajishirzi[2]
Seoul National University[1], University of Washington[2]
[*] All work was done while the author was an exchange student at University of Washington

ACL 2017 @Vancouver, Canada
Aug 1 (Tue) Poster Session 2

Abstract

- We show that coarser, sentence-level QA can significantly benefit from the transfer learning of BiDAF models trained on a large, fine-grained QA dataset.
- Span-level supervision leads to enriched representation of sentences, which can benefit other NLP tasks via weight transfer.
- Particularly we achieve state of the art on two smaller scale, sentence-level QA datasets by transferring BiDAF trained on span-level SQuAD.

Background and Data

- *Word2vec* and *GloVe* are some successful cases of transfer learning in NLP.
- Mou et al. [2] argue that transfer learning does not work when target task is different from source task. e.g. Sentiment Analysis → Question Classification, RTE → Paraphrase detection

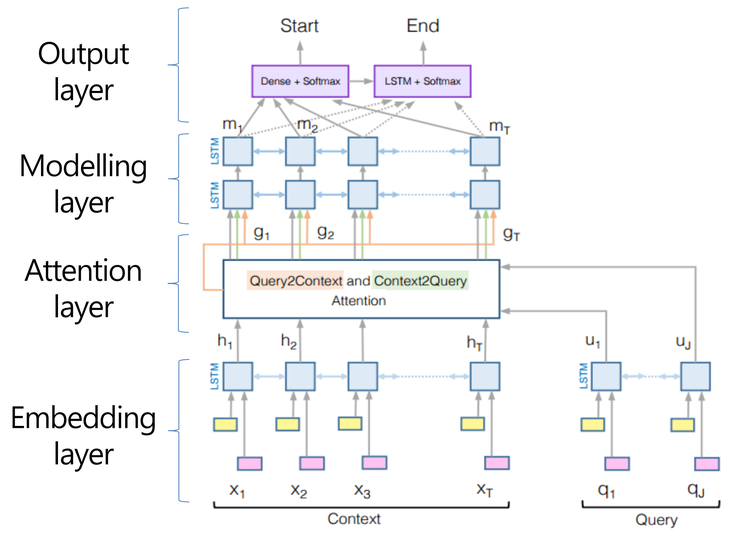
SQuAD	span-level	100k	context from Wikipedia, Q from human
SQuAD-T	sent-level	100k	our modification from SQuAD to sent-level

WikiQA	sent-level	2k	context from Wikipedia, Q from Bing
SemEval	sent-level	2k	conversation from community
SICK	RTE	10k	consists of a hypothesis and a premise

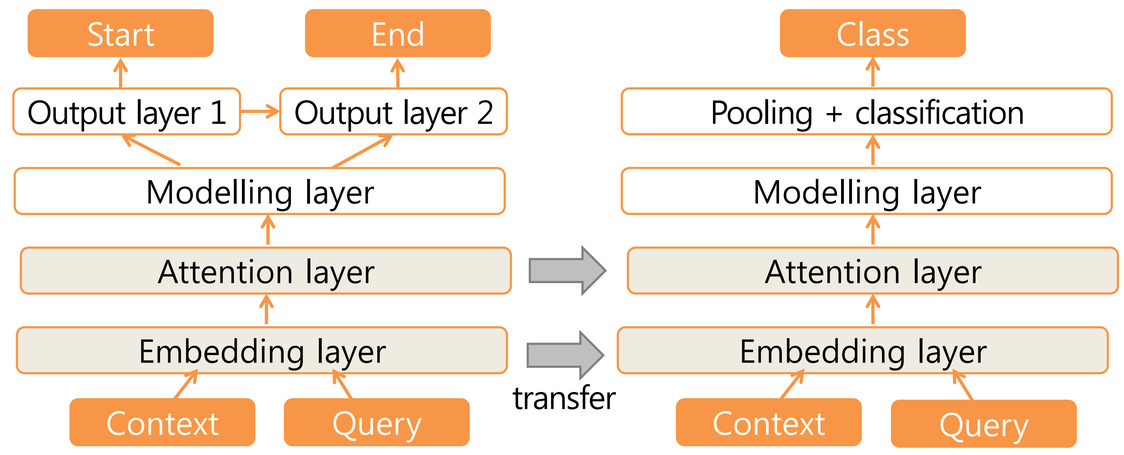
Dataset			Examples
Span-level	SQuAD	**Q**	Which company made Spectre
		C	Spectre (2015) is the 24th James Bond film produced by Eon Productions. It features …
		A	"Eon Productions"
Sentence-level	WikiQA	**Q**	Who made airbus
		C1	Airbus SAS is an aircraft manufacturing subsidiary of EADS, a European aerospace company.
		C2	Airbus began as an union of aircraft companies.
		C3	Aerospace companies allowed the establishment of a joint-stock company, owned by EADS.
		A	C1(Yes), C2(No), C3(No)
	SemEval	**Q**	I saw an ad, data entry jobs online. It required we give a fee and they promise fixed amount every month. Is this a scam?
		C1	well probably is so i be more careful if i were u. Why you looking for online jobs
		C2	SCAM!!!!!!!!!!!!!!!!!!!!
		C3	Bcoz i got a baby and iam nt intrested to sent him in a day care. thats y iam (...)
		A	C1(Good), C2(Good), C3(Bad)

Model

BiDAF [1] is a span-level QA model trained on SQuAD, achieved state of the art on SQuAD
- code: https://github.com/allenai/bi-att-flow
- demo: https://allenai.github.io/bi-att-flow/demo/



BiDAF-T (sentence-level QA model, by modifying BiDAF) **with Weight Transfer**

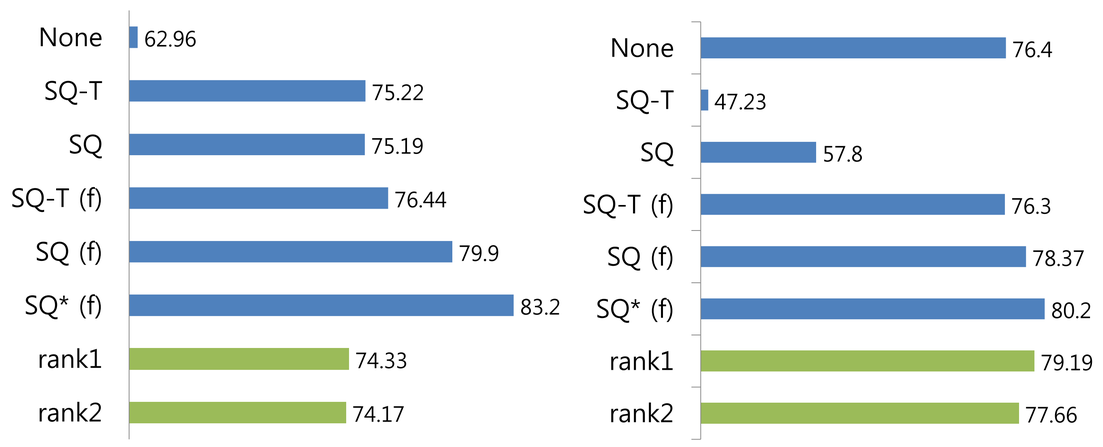


BiDAF outputs start and end position of span.

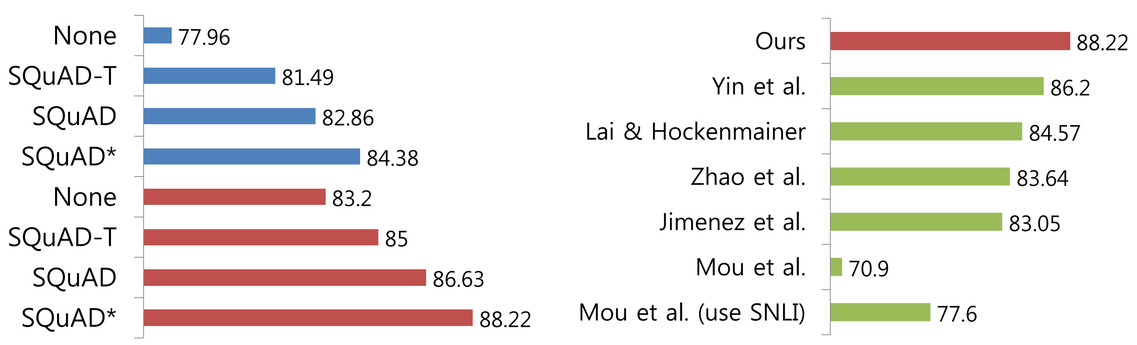
BiDAF-T outputs classification result.

Experiments

- Models with transfer learning outperform those without.
- Weight-transferred models achieve better result with pretraining on SQuAD than SQuAD-T.

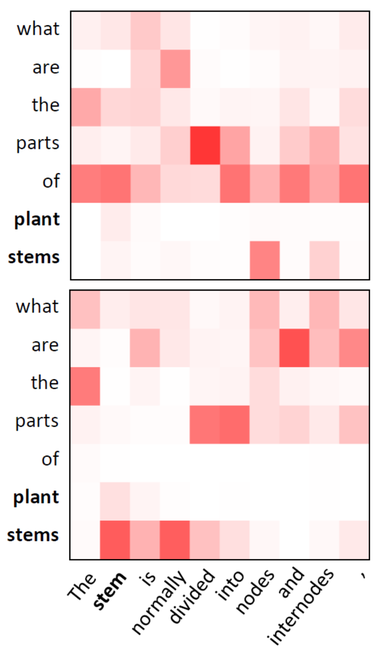


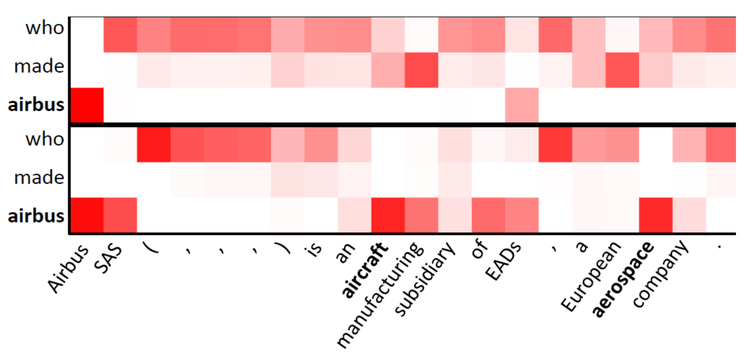
We achieve state of the art on WikiQA and SemEval-2016.
(f) indicates the model is fine-tuned and * indicates ensemble method. Metric is Mean Average Precision (MAP).



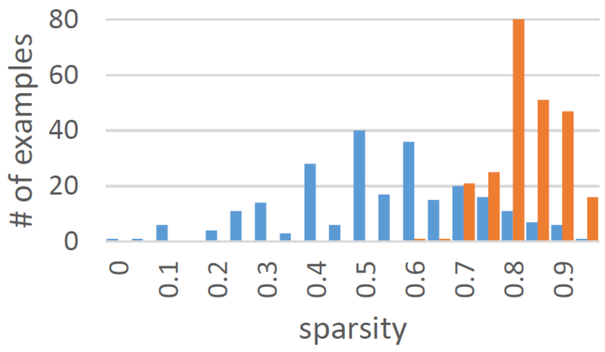
We achieve state of the art on SICK. Result with no use of SNLI, with use of SNLI, and from previous works, respectively.

Qualitative Results

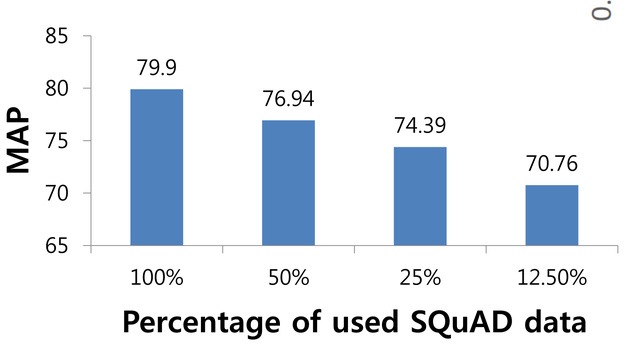




Attention maps. Correspondence between the words in WikiQA for SQuAD-T-pretrained (top) and SQuAD-pretrained (bottom).
- correspondence between words with similar meaning
- sparsity of the graph



Sparsity Graph. Pretrained on SQuAD-T and on SQuAD (Avg 0.56 vs 0.84)



Results with varying sizes of SQuAD dataset used during pretraining finetuned and tested on WikiQA

[1] Minjoon Seo et al. 2017. Bidirectional attention flow for machine comprehension. In ICLR.
[2] Lili Mou et al. 2016. How transferable are neural networks in nlp applications? In EMNLP.