Robust Question Answering: A study on Self-Attention Mechanism and Data Augmentation

Shicheng Xu, Yechao Zhang
{lukexu, yechaoz}@stanford.edu

CS224N 2022 Winter
Default Project (RobustQA)

Introduction

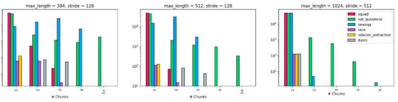
Build a robust QA system that can adapt to unseen domains with only a few training samples from the domain.
Our system focus on the **model limitation** of 512 tokens in full self-attention mechanism, and the **data limitation** of only 127 questions per OOD training set:
✓ Model limitation: Increasing attention sequence length beyond 512.
✓ Data limitation: Data augmentation at context, answer, and question level.
Our best single model achieves EM/F1 = **42.661/60.185** on the test set.

Immediate improvements over baseline

Combine ID and OOD.
- Baseline is only trained on ID.
- Further finetune on OOD does not always improve performance.

Increase max length to 512.
- Max length in baseline model is 384.
- Chunking does not allow model to learn long dependency across chunks.
- Increasing max length can reduce number of chunks per question.

Number of chunks at different max length and stride length.



DistilBertLongForQuestionAnswering

Longformer introduces sparse attention mechanisms to process long sequences and could replace self-attention of any Transformer-based model.
- Sliding window: each token attends to w tokens within sliding window.
- Global attention: questions tokens can attend to all context tokens.

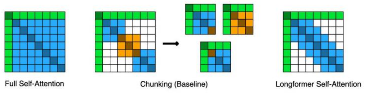
We implemented:
- `DistilBertLongSelfAttention`: reuse pretrained DistilBERT weights; extend embeddings size to 2048 by repeating.
- `DistilBertLongForQuestionAnswering`: Set global attention mask for question tokens.

Advantage over chunking:
- Attention weights are jointly learned from true label across sliding windows.
- Question token can attend to all tokens

Computation Complexity

	Full Self-Attention	Chunking	Longformer Self-Attention
Computation Complexity	$O(n^2)$	$O(m \times w^2)$	$O(n \times (w+q))$

Full self-attention vs. Chunking vs. Longformer self-attention



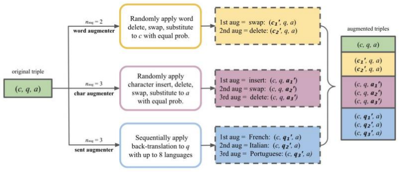
Training speed and memory comparison at different sequence length

Sequence length (Model)	Batch size	Train speed (it/s)	Train time	Memory (MB)
len=384 (DistilBERT)	16	82	50min	6639
len=512 (DistilBERT)	16	60	1h	8997
len=1024 (DistilBERTLong)	8	13	3h10min	11781
len=2048 (DistilBERTLong)	4	6.4	6h30min	12035

Data Augmentation

- We implement 3 types of augmenters:
  - **Word augmenter** for context
    - delete, swap, substitute
  - **Character augmenter** for answer
    - insert, delete, swap, substitute
  - **Sentence augmenter** for question
    - backtranslation
- **Hyperparameter** search for all techniques for each augmenter
  - Number of augmentations *n_aug*
  - Strength: (*p_word*, *p_char*)

DA Pipeline


original triple
(c, q, a)
n_{aug} = 2
word augmenter
Randomly apply word delete, swap, substitute to c with equal prob.
1st aug = swap: (c_1', q, a)
2nd aug = delete: (c_2', q, a)
n_{aug} = 3
char augmenter
Randomly apply character insert, delete, swap, substitute to a with equal prob.
1st aug = insert: (c, q, a_1')
2nd aug = swap: (c, q, a_2')
3rd aug = delete: (c, q, a_3')
n_{aug} = 3
sent augmenter
Sequentially apply back-translation to q with up to 8 languages
1st aug = French: (c, q_1', a)
2nd aug = Italian: (c, q_2', a)
3rd aug = Portuguese: (c, q_3', a)
augmented triples
(c, q, a)
(c_1', q, a)
(c_2', q, a)
(c, q, a_1')
(c, q, a_2')
(c, q, a_3')
(c, q_1', a)
(c, q_2', a)
(c, q_3', a)

DA Optimal Hyperparameters

Augmenter	n_{aug}	$(p_{\text{word}},\ p_{\text{char}})$			
		insert	delete	swap	substitute
word	1	(×, ×)	(0.40, ×)	(0.20, ×)	(0.05, ×)
character	1	(0.05, 0.10)	(0.20, 0.20)	(0.20, 0.40)	(0.05, 0.10)
sentence	6	(×, ×)	(×, ×)	(×, ×)	(×, ×)

QA Example

Question: Who is operated on and dies from his wounds?

Context: … Stubby's cell-mates are Emmanuelle 'Bunny' O'Neill (Lynne Frederick) a young prostitute; Bud (Harry Baird), a mentally disturbed black man who is obsessed with the dead; and Clem (Michael J. Pollard), the town drunkard. During the night, a masked posse of men wreck havoc throughout the settlement. Fearing for their safety as the massacre progresses, the prisoners demand to be realeased, but the sheriff refuses. The following morning, the sheriff, amused by Stubby's ways, allows them to leave town, pointing them to an abandoned wagon as their means of transportation. … That evening, Chaco persuades the group to take peyote with him. Stubby, secretly spits out the hallucinogenic cactus, suspicious alerted, but the rest of the party accepts. Chaco then amuses himself with Clem, humiliating the drug-addled alcoholic by ordering that he crawl and bark like a dog for some liquor. When Stubby attempts to make a run for it, Chaco subdues him and ties up the four travelers at gunpoint. As day breaks, Chaco rapes Bunny despite seeing her condition. Stubby swears to kill Chaco one day if he survives. The evil bandit then rides off on the wagon with all their gear, leaving them tied up in the desert, except for Clem who he shoots in the leg. Clem frees the others and they continue their trek on foot. As Clems wound worsens, Stubby is forced to remove the bullet and he and Bud carry him on a makeshift stretcher. Later, the group comes upon the aftermath of another attack by Chaco and two henchmen. The Christian wagoners, children and all, have all been slaughtered.The group's journey continues. Eventually the four friends see buildings ahead. … That evening, Clem gets more sicker from the infection of his wound. But he makes dying request that Stubby and Bunny to get married. The couple profess their love for one another and they kiss for the first time. The next morning, the couple decide to leave the ghost town. Bud elects to stay behind having learned to hunt animals for food. But Stubby and Bunny discover Clems mutilated dead body and learn that Bud has taken a liking to cannibalism as well. …

Answer: Clem
Prediction (word augmenter): Stubby Preston
Prediction (character augmenter): Clem
Prediction (sentence augmente): Stubby
Prediction (len=512): The Christian wagoners
Prediction (len=1024): Clem

Conclusion

- **Cotrain ID+OOD** is a simple but effective way to train a robust model.
- **Increasing the sequence length** in self-attention mechanism should be **prioritized** given available accelerator resource.
  - Increase stride from 128 to 512 significantly improves our final model performance (len=1024).
- **Data augmentation** is effective if you have a **tight budget** on accelerator resource.
  - Character augmenter for answer spans, forcing the model to learn surrounding context, tends to work better for long-sequence context

Experiments and Results

Validation F1/EM

ID	Methods	F1	EM
0	Baseline (ID)	47.72	30.63
0	Baseline (ID)+Finetune(OOD)	**48.49**	**32.46**
1	Cotrain (ID+OOD)	**51.53**	**35.86**
1	Cotrain (ID+OOD)+Finetune(OOD)	50.81	34.82
2	len=512 (ID)	50.67	34.29
2	len=512 (ID)+len=512 (OOD)	50.17	34.29
2	len=512 (ID+OOD)	**51.17**	**36.91**
2	len=512 (ID+OOD)+len=384 (OOD)	50.27	35.86
2	len=512 (ID+OOD)+len=512 (OOD)	50.96	36.65
3	len=1024 (OOD)	**47.56**	**34.03**
3	len=1536 (OOD)	47.41	32.72
3	len=2048 (OOD)	47.41	32.98
4	len=1024, stride=128 (ID+OOD)	50.07	35.08
4	len=1024, stride=512 (ID+OOD)	**53.70**	**38.48**
5	DA (OOD) - Word Augmenters	48.57	33.51
6	DA (OOD) - Character Augmenters	49.24	31.94
7	DA (OOD) - Sentence Augmenters	49.10	34.82
8	DA (OOD) - Combined Augmenters	**49.39**	**35.08**