Question Answering on SQuAD Dataset

Ожерельева София, Поконечный Эдуард

sofya.ozherelieva@yandex.ru, pokonechnyy.ep@phystech.edu

Moscow Institute of Physics and Technology

Поставленная задача

Задача Question answering заключается в поиске релевантной вопросу информации в документе (ответ на каждый вопрос представляет собой фрагмент текста или промежуток из соответствующего отрывка «контекста» или может не существовать в принципе).

Бейзлайновая модель

Начальная архитектура состояла из следующих слоёв:Word Embedding, Contextual Embedding, BiDAF, Self-Attention, Modeling Layer и Output Layer.

Улучшения

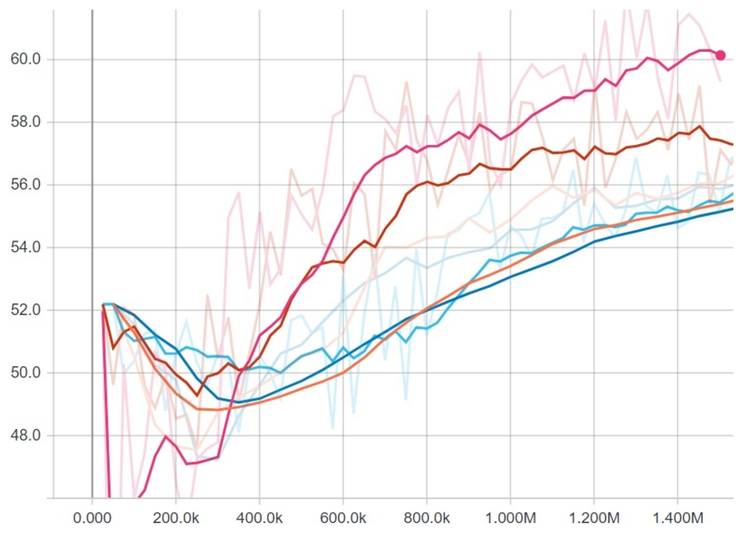
Рассматривалось влияние на качество следующих компонент:

- Char based embedding
- Свертки в качестве Contextual Embedding Layer (вместо LSTM). А также их комбинация
- Double Cross Attention вместо BiDirectional Attention Flow
- Self-attention после BiDirectional Attention Flow
- Размеры слоев, тюнинг гиперпараметров
- Подбор lr_scheduler

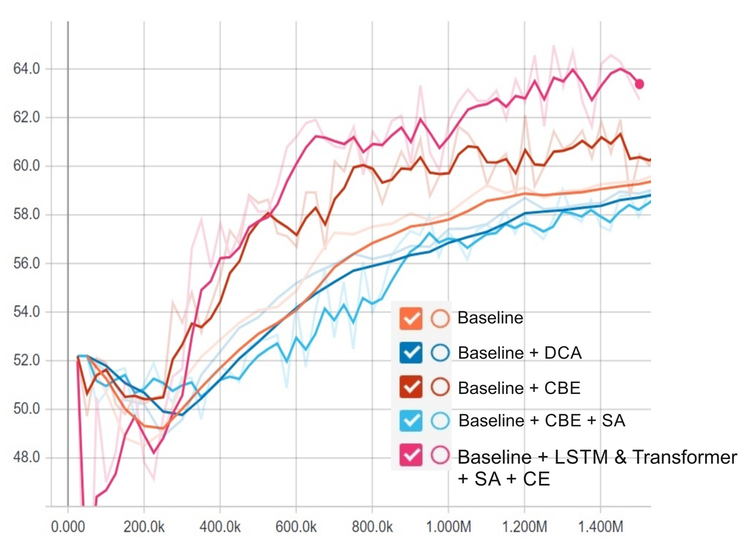
Метрики

Для оценки качества использовались метрики: Exact Match (строгая бинарнаяметрика true/false), F1 (более мягкая метрика, комбинация precision и recall).

EM score



F1 score



Набор данных

Данные разделены на три группы: train (129,941 примеров), dev (6078 примеров), test (5915 примеров). Данные состоят из контекста, вопроса и ответа (таргет), представляющего из себя диапазон контекста, в котором лежит ответ на вопрос, или NA, если ответа в контексте не содержится.

Таблица 1: Experiments results on the SQuAD 2.0 dataset

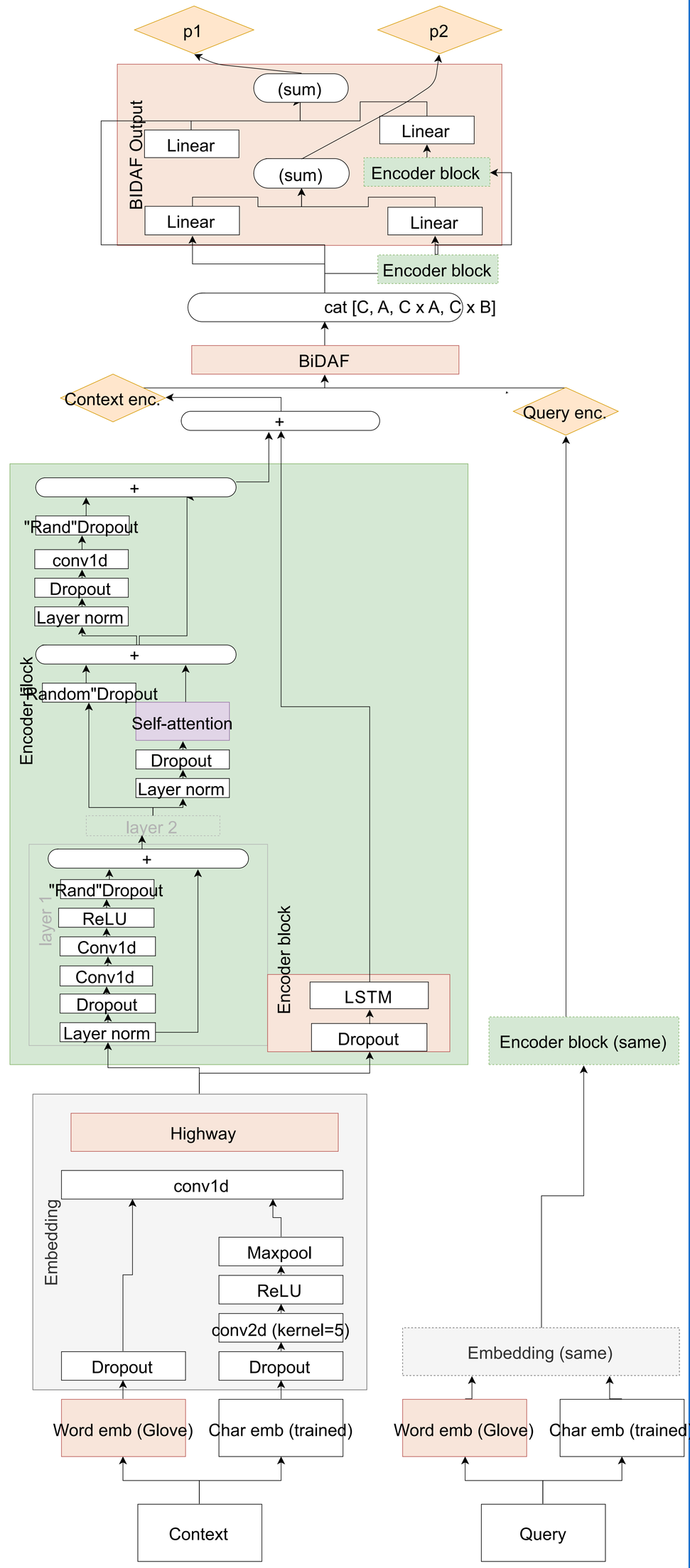
	Dev set		Test set	
Single model	F1	EM	F1	EM
Baseline	60.93	57.39		
Baseline + Char Emb	60.72	57.33		
Baseline + Char Emb v.2 + fine tuning	62.10	59.30		
Baseline + DCA	60.19	57.20		
Baseline + SA + Char Emb (monster)	59.20	57.02		
Baseline + LSTM & Transformer combo + SA + CE	> 64.70	> 61.82		

Таблица 2: Results on the SQuAD 2.0 dataset

	Dev Set	Test Set
Single model	F1/EM	F1/EM
Logistic Regression Baseline	51.0/40.0	51.0/40.4
BiDAF	?/?	59.33/62.3
QANet(https://github.com/BangLiu/QANet-PyTorch)	80.49/71.24	?/?
BiDAF++	?/?	65.65/68.87
Our Best Model	64.70/61.82	?/?
Ensemble	F1/EM	F1/EM
SQuAD competition best (BERT + DAE + AoA)	?	87.15/89.47

Список литературы

1. QANet: Combining Local Convolution with Global Self-Attention for ReadingComprehension. 2018.
2.Convolutional Neural Networks for Sentence Classification. 2014.



Полученные результаты

В рамках данного проекта нам удалось:

- Разобраться в имеющихся на данный момент базовых подходах к решению question answering
- Имплементировать базовые подходы, понять основные особенности их работы, а также степень их влияния на качество решения задачи

Дальнейшие исследования

Появилась почва для создания конкурирующих с SotA решений

- Попробовать новые архитектуры (комбинации слоев, ансамбли и прочее)
- Изучить взаимодействие с BERT
- Рассмотреть конкурирующие лучшие модели, изучить их внутреннее устройство, недостатки, подумать над способами их устранения
- Рассматривать влияние ответов без вопросов, научить модель работать с такими данными