Solving “Contradictory, My Dear Watson” competition on Kaggle

—

Кузьмин Сергей, Андреев Никита, Федоркина Мария

Natural Language Inference

Два предложения:

- посылка (premise)
- гипотеза (hypothesis)

Задача: понять, правда ли что гипотеза следует из (entailment), противоречит (contradiction) или не следует из (neutral) посылки.

Посылка	Гипотеза	Вывод
These are issues that we wrestle with in practice groups of law firms, she said.	Practice groups are not permitted to work on these issues.	contradiction
An older and younger man smiling.	Two men are smiling and laughing at the cats playing on the floor.	neutral
A black race car starts up in front of a crowd of people.	A man is driving down a lonely road.	contradiction
A soccer game with multiple males playing	Some men are playing a sport.	entailment
A smiling costumed woman is holding an umbrella.	A happy woman in a fairy costume holds an umbrella.	neutral

Datasets

Соревнование “Contradictory, My Dear Watson” на Kaggle

https://www.kaggle.com/c/contradictory-my-dear-watson/data

- Тестовый датасет (5195 значений)
- Тренировочный датасет (12120 значений)
- 15 языков

Baseline algorithms

- LSTM

LSTM

- Модель: учим 3 линейных слоя поверх GloVe эмбеддингов
- Оптимизатор: Adam
- 3 эпохи
- Счет на Kaggle: 0.40923
- За 3 эпохи точность на тренировочной выборке большая, на валидационной выборке низкая => переобучение
- Возьмем SGD, побольше эпох, будем следить за точностью на валидационной выборке => 0.44812

Выбор трансформера

- bert-base-multilingual-cased
- tf-xlm-roberta-large
- xlm-roberta-large-xnli
- distilbert-base-multilingual-cased

bert-base-multilingual-cased

- Adam optimizer, categorical crossentropy loss
- 10 эпох
- Счет на Kaggle: 0.64581
- Код на гитхабе

tf-xlm-roberta-large

- Adam optimizer, categorical crossentropy loss
- 4 эпохи
- Счет на Kaggle: 0.80500
- Код на гитхабе

xlm-roberta-large-xnli

- Adam optimizer, categorical crossentropy loss
- 3 эпохи
- Счет на Kaggle: 0.92050
- Код на гитхабе

xlm-roberta-large-xnli

Special thanks to Tensorflow Datasets (TFDS) for providing this and many other useful datasets! (c)

Код сравнения датасетов:

- train premises not in xnli: 6863 / 12120
- train гипотезы not in xnli: 6870 / 12120
- test premises not in xnli: 2944 / 5195
- test гипотезы not in xnli: 2943 / 5195

Эксперименты

- data augmentation
- optimizers
- diff pruning
- different approaches

Data augmentation

- Data augmentation в NLP -- сложно!
- translation
- contextual augmentation
- synonyms
- syntactic data augmentation (ссылка)
- nlpaug library

Translation

- не получилось =(
- проблемы с библиотекой googletrans -- где-то в ноябре все упало, и с тех пор не поднялось (по крайней мере, у нас)

Contextual augmentation

- заменим или добавим слова, основываясь на контексте (например, используя BERT)
- есть вероятность, что все сломается -- контекст при замене или добавлении может не измениться, но смысл будет противоположный
- только гипотеза, не проверяли (ресурсозатратно)

Synonyms

- будем заменять случайные слова на их синонимы вне зависимости от контекста
- иногда получается бессмысленный текст
- более безопасно, чем контекстуальные замены
- добавили около 6000 предложений, Kaggle score: $0.56034 => 0.57189$

Syntactic data augmentation

- На стандартных датасетах претренированные модели показывают хорошие результаты, но на специальных часто ошибаются
- Гипотеза: проблема в датасетах, а не в моделях
- Пример:

  The lawyer saw the actor. =/=> The actor saw the lawyer.

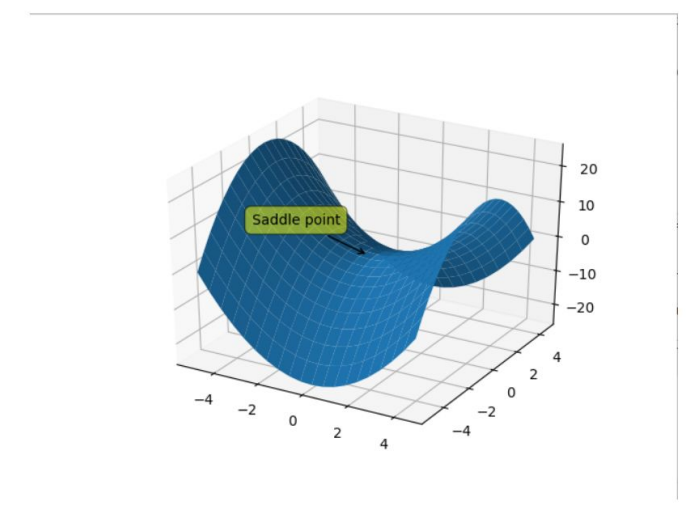
- Построим специальные примеры, в которых будем менять, например порядок слов в предложении, чтобы модель лучше выучила язык
- Было решено не делать, т.к. почти все данные взяты из стандартных датасетов

Optimizers

- цикличный learning rate
- stochastic weight averaging

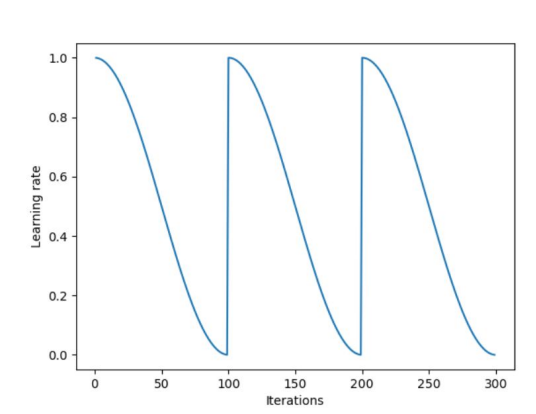
Cyclic learning rate

- Будем то увеличивать, то уменьшать learning rate
- Хотим избежать ситуации, в которой мы застрянем в седлах
- SGDR (ссылка)



Stochastic Weight Averaging

- ссылка простым языком, ссылка сложным языком
- ансамбли сетей -- агрегация нескольких сетей и усреднение результата
- будем делать по-другому -- тренируем одну сеть и используем циклический learning rate для получения разных моделей с помощью усреднения весов моделей, полученных в результате обучения



$$w_{\text{SWA}} \leftarrow \frac{w_{\text{SWA}} \cdot n_{\text{models}} + w}{n_{\text{models}} + 1},$$

Stochastic Weight Averaging

- эксперименты проводились на DistilBERT, чтобы не тратить много ресурсов
- 24 эпохи, оптимизатор -- SGDR, длина цикла -- 4
- Score: 0.57189 $=>$ 0.57901
- почти бесплатное улучшение результатов модели
- проведено мало экспериментов, возможно, можно добиться большего улучшения точности