Методы предобработки текстовых данных для ускорения обучения языковых моделей

Сурков Максим Константинович

Научный руководитель: Ямщиков Иван Павлович

Санкт-Петербургская школа физико-математических и компьютерных наук
НИУ ВШЭ СПБ

17 марта 2021 г.

Обработка естественного языка в реальной жизни

- социальные сети
- электронная почта
- службы доставки
- голосовые помощники
- переводчики
- чат боты




Задачи обработки естественного языка

1. классификация последовательностей
   - спам
   - грубая речь[1]
2. генерация выходной последовательности из исходной
   - машинный перевод
   - ответы на вопросы
3. выделение информации из последовательностей
   - выделение именованных сущностей[2]

[1] G. H. Paetzold et al., SemEval'19 Task 5: Hate Speech Identification with RNN.

[2] Vikas Yadav et al., SemEval'19 Task 12: Deep-Affix Named Entity Recognition of Geolocation Entities. ACL'19

Современные методы решения задач обработки естественного языка

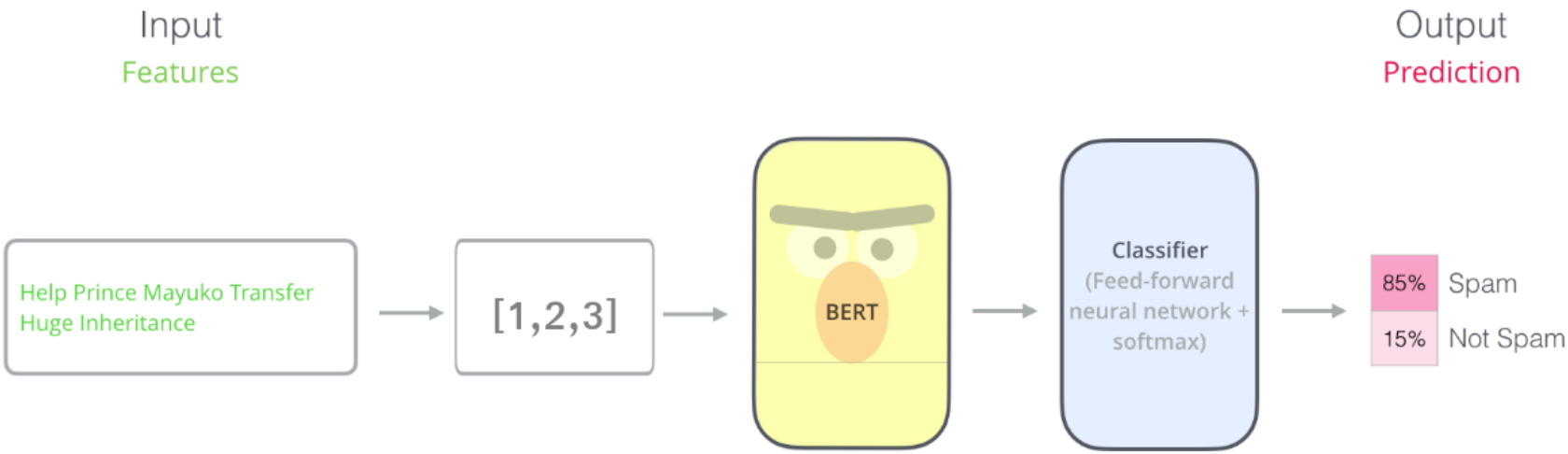
1. Механизм внимания[1]
2. **BERT** (Google)[2]
3. GPT-3 (OpenAI)[3]

[1] Ashish Vaswani et al., Attention Is All You Need, 2017

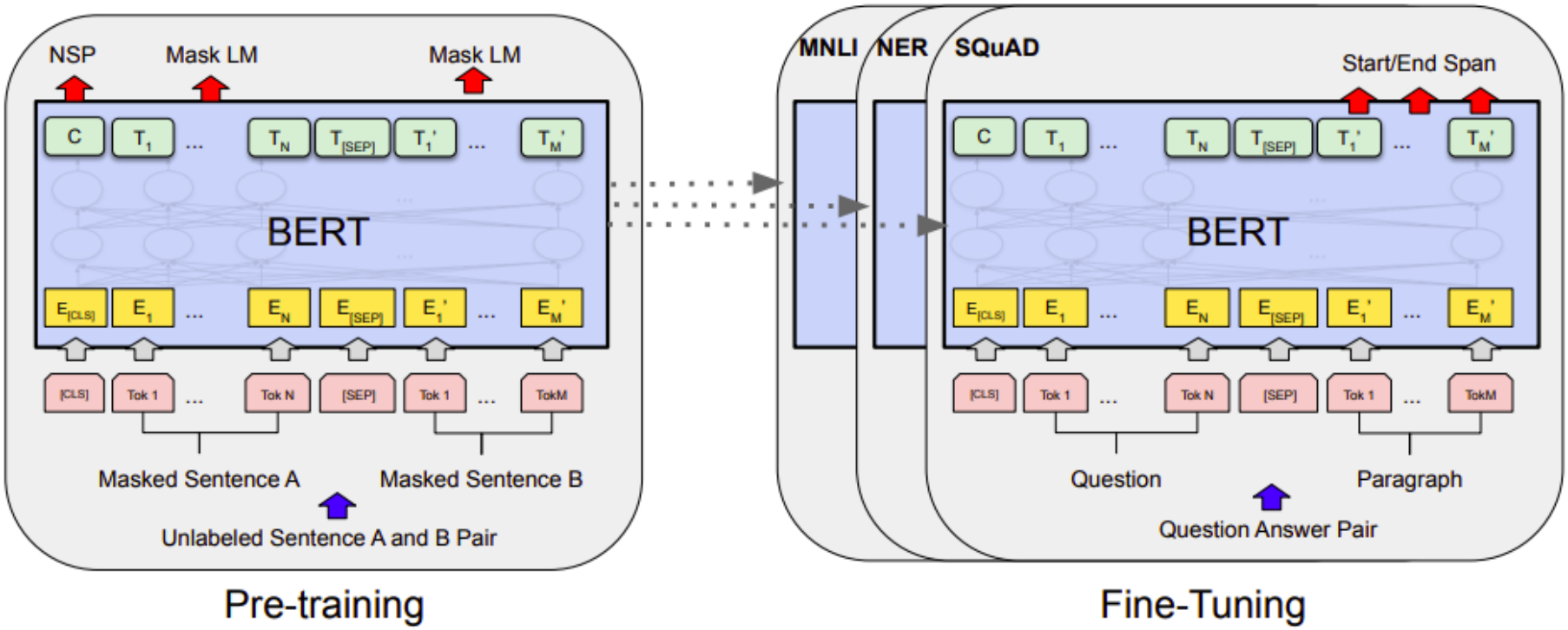
[2] Jacob Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019

[3] Tom B. Brown et al., Language Models are Few-Shot Learners, 2020

BERT. Использование



BERT. Обучение



BERT. Требуемые ресурсы

- количество параметров: $110M - 340M$
- время на предобучение: от 2-4 дней до 1-2 недель[1]
  - мировой рекорд: 47 минут на **1472** V100 GPU[2]
- время на дообучение: 1-2 дня
- размеры данных:

Датасет	Размер
Wikipedia	3-600M
HND	600k-2M
s140	1.6M
IWSLT	200-230k
QQP	364k
MNLI	393k

[1] При использовании 1x-4x GPU Nvidia Tesla V100 32Gb

[2] https://developer.nvidia.com/blog/training-bert-with-gpus

BERT. Существующие методы оптимизации

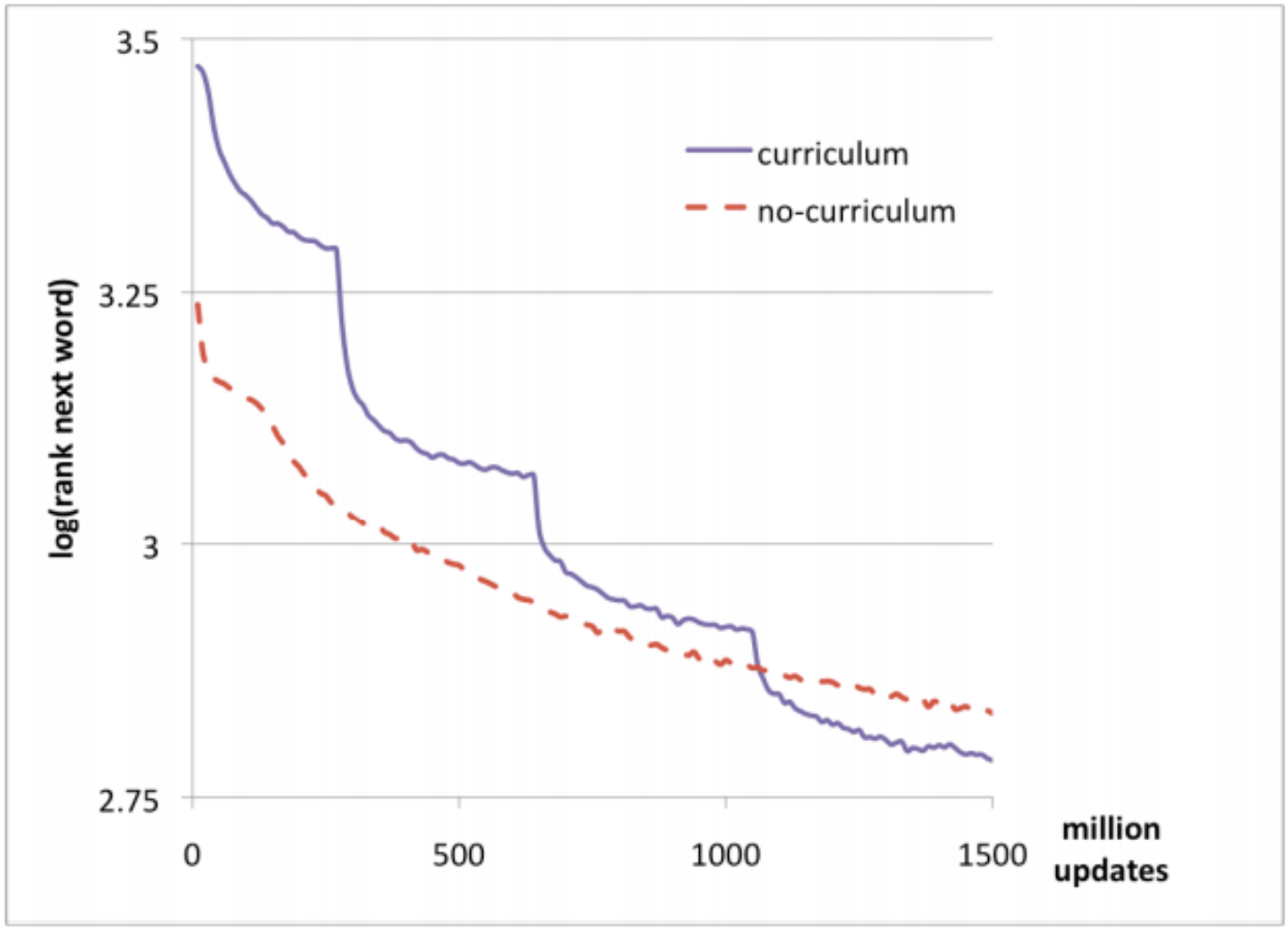
- квантизация[1]
- дистилляция[2]
- прунинг[3]

[1] Sheng Shen et al., Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT, 2019

[2] Victor Sanh et al., DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020

[3] Hassan Sajjad et al., Poor Man's BERT: Smaller and Faster Transformer Models, 2020

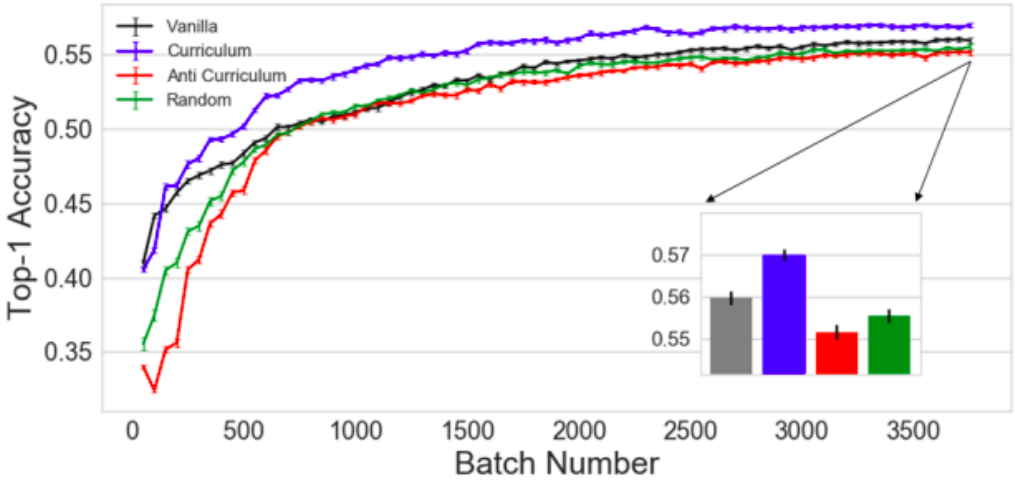
Обучение с расписанием. Начало



Y. Bengio et al., Curriculum learning, 2009

Обучение с расписанием. Применение

- компьютерное зрение[1]



- обучение с подкреплением[2]
- глубокое обучение[3]

---

[1] Guy Hacohen, Daphna Weinshall, On The Power of Curriculum Learning in Training Deep Networks, 2019

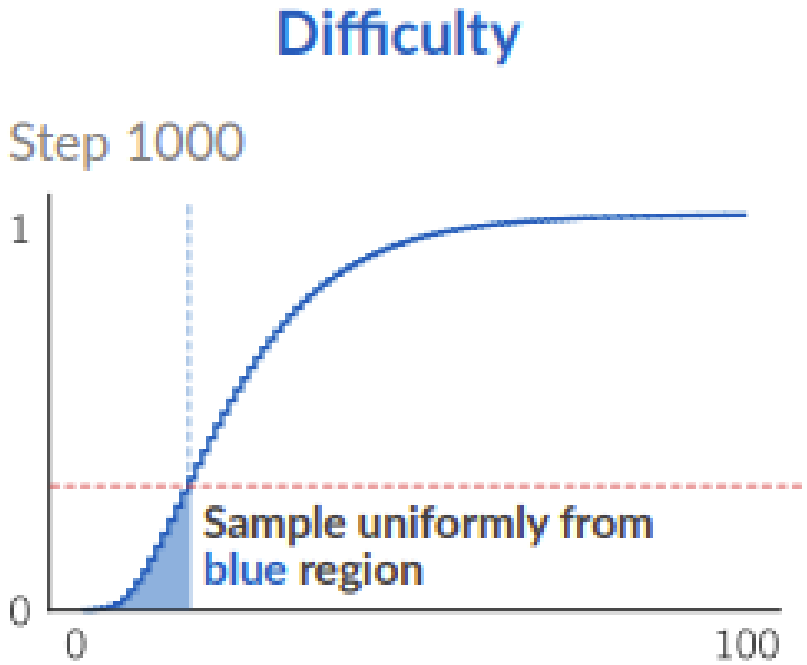
[2] Sanmit Narvekar et al., Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey, 2020

[3] Mermer et al., Scalable Curriculum Learning for Artificial Neural Networks, 2017

Обучение с расписанием в обработке языка

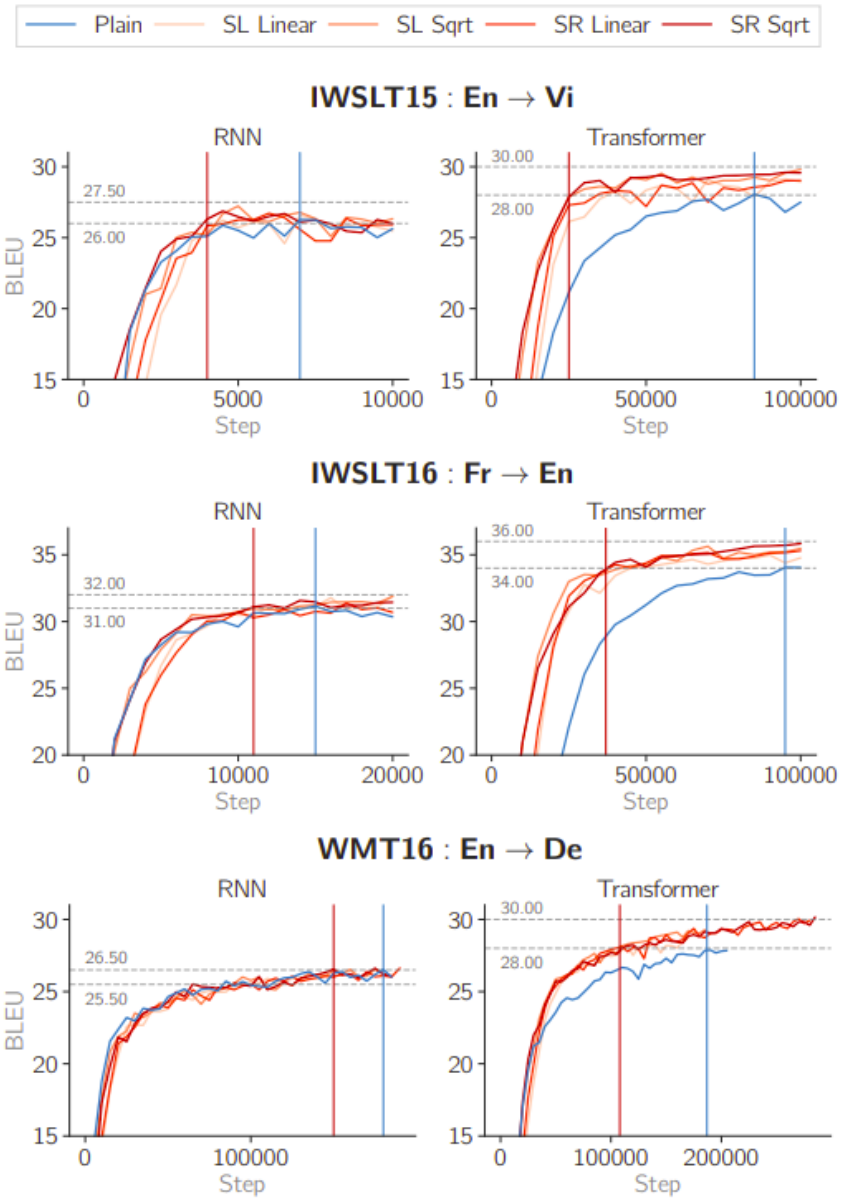
- Задача: машинный первод
- Модель: BERT, LSTM
- Датасеты: IWSLT'15, IWSLT'16, WMT'16
- Алгоритм:

1. сортируем тексты по сложности (длина, логарифм веротности правдоподобия)
2. в течение T шагов (рассмотрим шаг t)
   - считаем $c(t) \in [0, 1]$
   - строим батч из $c(t)$ **первых** текстов корпуса
   - шаг обучения



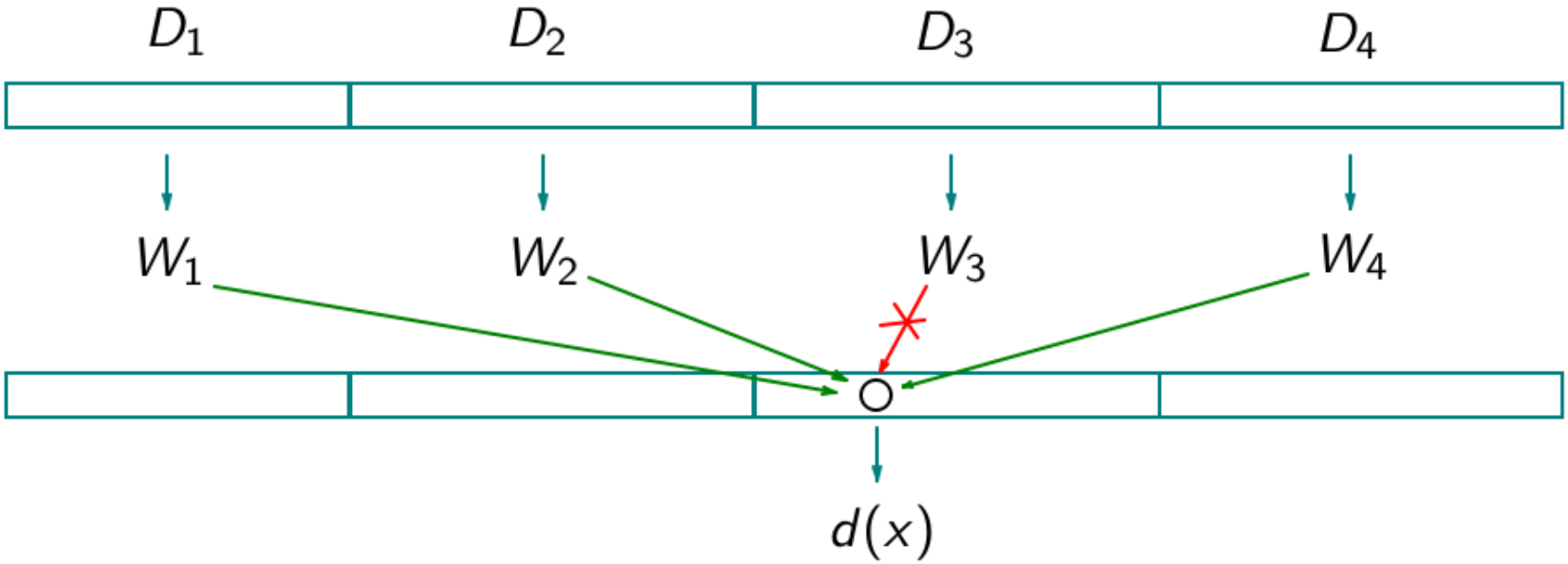
E. A. Platanios et al., Competence-based Curriculum Learning for Neural Machine Translation, ACL'19

Обучение с расписанием в обработке языка



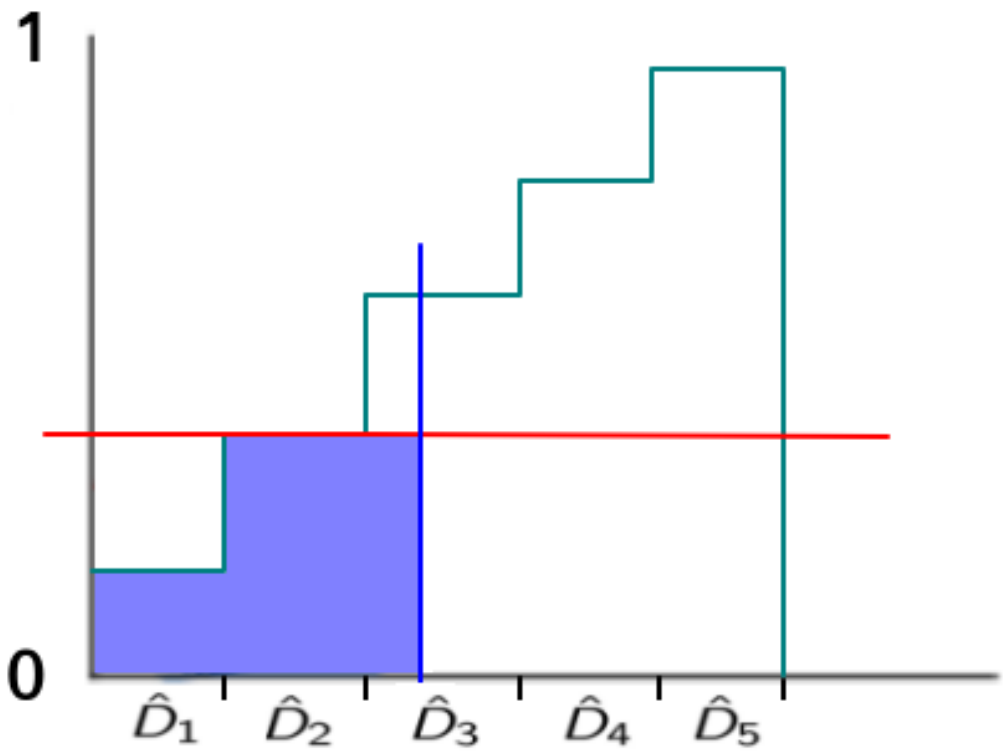
Обучение с расписанием в обработке языка

- Задача: классификация
- BERT
- Датасеты: SQuAD 2.0, NewsQA, GLUE
- Алгоритм: в течение T шагов



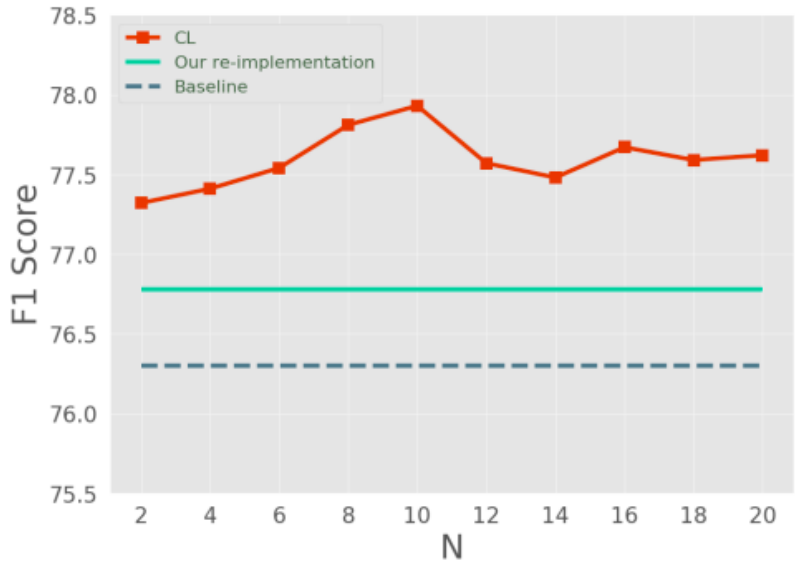
Benfeng Xu et al., Curriculum Learning for Natural Language Understanding, ACL'20

Обучение с расписанием в обработке языка



Обучение с расписанием в обработке языка

	MNLI-m	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B	Avg
results on dev									
BERT Large	**86.6**	92.3	91.3	70.4	93.2	88.0	60.6	90.0	84.1
BERT Large*	**86.6**	92.5	91.5	74.4	93.8	91.7	63.5	90.2	85.5
BERT Large+CL	**86.6**	**92.8**	**91.8**	**76.2**	**94.2**	**91.9**	**66.8**	**90.6**	**86.4**
results on test									
BERT Large	**86.7**	91.1	89.3	70.1	**94.9**	89.3	60.5	87.6	83.7
BERT Large*	86.3	92.2	**89.5**	70.2	94.4	89.3	60.5	87.3	83.7
BERT Large+CL	**86.7**	**92.5**	**89.5**	**70.7**	94.6	**89.6**	**61.5**	**87.8**	**84.1**



Обучение с расписанием в обработке языка. Направления для исследований

- Много важных задач обработки естественного языка с **большими** корпусами тренировочных данных
- Решаются с помощью **тяжелых** моделей, которые **долго** учатся
- Не исследованы метрики оценки сложности текста (длина - текущий предел)
- Эксперименты проведены только на определенных задачах
  - ACL'19 - только задача машинного перевода
  - ACL'20 - только задача классификации[1]
- Не исследовано влияние обучения с расписанием на этапе предобучения

[1] Не совсем честное обучение с расписанием; Не ускоряет; Требует еще больших ресурсов

Цели и задачи

Цель: ускорить обучение языковой модели BERT с помощью обучения с расписанием за счет метрики оценки сложности текстовых данных на задачах предобучения, классификации и машинного перевода

Задачи:

1. Найти эффективные[1] метрики оценки сложности текста
2. Реализовать механизм подсчета найденных метрик на больших датасетах
3. Сравнить найденные метрики с существующими метриками оценки сложности текста
4. Исследовать влияние найденных метрик на скорость обучения языковой модели BERT

[1] с точки зрения сокрости обучения модели

Поиск метрик

1. **длина**, вероятность правдоподобия[1]
2. информационный поиск
   - **tf-idf**
   - энтропия, семантическая сложность[2]
3. средняя частота слова, самое редкое слово в предложении[3]
4. число определенных частей речи[4]
5. **теория информации**

---

[1] E. A. Platanios et al., Competence-based Curriculum Learning for Neural Machine Translation, ACL'19

[2] Frans van der Sluis et al., Using Complexity Measures in Information Retrieval, 2010

[3] Xuan Zhang et al., An Empirical Exploration of Curriculum Learning for Neural Machine Translation, 2018

[4] Tom Kocmi, Ondrej Bojar, Curriculum Learning and Minibatch Bucketing in Neural Machine Translation, 2017

Поиск метрик

метрика	формула
Multi-information	$\sum_{v \in V} H_p(X_v) - H_p(X_V)$
Excess Entropy (EE)	$\left[\sum_{v \in V} H(X_{V \setminus \{v\}})\right] - (N-1)H(X_V)$
TSE	$\sum_{k=1}^{N-1} \frac{k}{N} C^{(k)}(X_V)$, где $C^{(k)}(X_V) = \frac{N}{k\binom{N}{k}} \sum_{A \subseteq V, \lvert A \rvert = k} H(X_A) - H(X_V)$
Transient information	:(

$$V = \{1, \ldots, N\}$$

$$X_V = (X_1, \ldots, X_N)$$

Nihat Ay et al., A **Unifying** Framework for Complexity Measures of Finite Systems, 2006

Адаптация EE и TSE под задачи обработки языка

1. Образование совместной случайной величины

$$T = (t_1, t_2, \ldots, t_{i-1}, t_i, \ldots, t_n)$$

$$t_i \to \xi^i_{t_i} =: \mu_i - \text{бинарная случайная величина}$$

$$\downarrow$$

$$\xi = (\xi^1_{t_1}, \xi^2_{t_2}, \ldots, \xi^{i-1}_{t_{i-1}}, \xi^i_{t_i}, \ldots, \xi^n_{t_n})$$

2. Вычисление энтропии

$$H(\mu) = \sum_{i=1}^{n} H(\mu_i|\mu_1, \mu_2, \ldots, \mu_{i-1}) = \sum_{i=1}^{n} H(\mu_i|\mu_{i-k}, \ldots, \mu_{i-1})$$

3. $k = 1$

$$H(\mu) = H(\mu_1) + H(\mu_2|\mu_1) + \ldots + H(\mu_i|\mu_{i-1}) + \ldots + H(\mu_n|\mu_{n-1})$$

Вычисление метрик

1. длина
2. tf-idf

$$\sum_{i=1}^{n} f(X_i) \log \frac{|D|}{\{j : X_i \in X^{(j)}\}}$$

   - $x_i \rightarrow$ число текстов, в которых есть x_i
3. энтропия для вычисления EE, TSE
   - длина $\rightarrow$ число текстов с такой длиной
   - $(i, x_i) \rightarrow$ число текстов, где $t_i = x_i$
   - $(x_i) \rightarrow$ число текстов, где x_i является последним токеном
   - $(i, x_{i-1}, x_i) \rightarrow$ число текстов, где на $(i-1)$-й позиции стоит x_{i-1}, а на i-й позиции стоит x_i
4. EE,TSE - ?

Вычисление EE

$$EE(X) = \left[\sum_{v \in V} H(X_{V \setminus \{v\}})\right] - (N-1)H(X_V) =$$

$$\left[\sum_{i=1}^{n} H(\mu_1, \ldots, \mu_{i-1}, \mu_{i+1}, \ldots, \mu_n)\right] - (n-1)H(\mu)$$

- $\mathcal{O}(n^2)$
- $\mathcal{O}(n)$

$$\sum_{i=1}^{n} H(\mu_1, \ldots, \mu_{i-1}, \mu_{i+1}, \ldots, \mu_n) =$$

$$= \sum_{i=1}^{n} H(\mu) - H(\mu_i|\mu_{i-1}) - H(\mu_{i+1}|\mu_i) + H(\mu_{i+1})$$

$$EE(X) = \sum_{i=2}^{n} H(\mu_i) - H(\mu_i|\mu_{i-1}) = \sum_{i=2}^{n} I(\mu_{i-1} : \mu_i)$$

Вычисление TSE

$$\sum_{k=1}^{N-1} \frac{k}{N} C^{(k)}(X_V)$$

$$C^{(k)}(X_V) = \frac{N}{k\binom{N}{k}} \sum_{A \subseteq V, |A|=k} H(X_A) - H(X_V) =$$

$$= \frac{N}{k} \left[\frac{1}{\binom{N}{k}} \sum_{A \subseteq V, |A|=k} H(X_A) \right] - \frac{N}{k} H(X_V)$$

Вычисление TSE

$$C^{(k)}(X_V) = \frac{1}{\binom{n}{k}} \sum_{A \subseteq V, |A|=k} H(X_A) = \frac{1}{\binom{n}{k}} \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} H(\mu_{i_1}, \mu_{i_2}, \ldots, \mu_{i_k})$$

1. $\mathcal{O}^*(2^n)$
2. $\mathcal{O}(n^2)$ - динамическое программирование
3. $\mathcal{O}(n)$

$$C^{(k)}(X_V) = \sum_{i=1}^{n} A_i H(\mu_i) + \sum_{i=2}^{n} B_i H(\mu_i | \mu_{i-1})$$

$$A_i = \frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n}$$

$$B_i = \frac{\binom{n-2}{k-2}}{\binom{n}{k}} = \frac{k(k-1)}{n(n-1)}$$