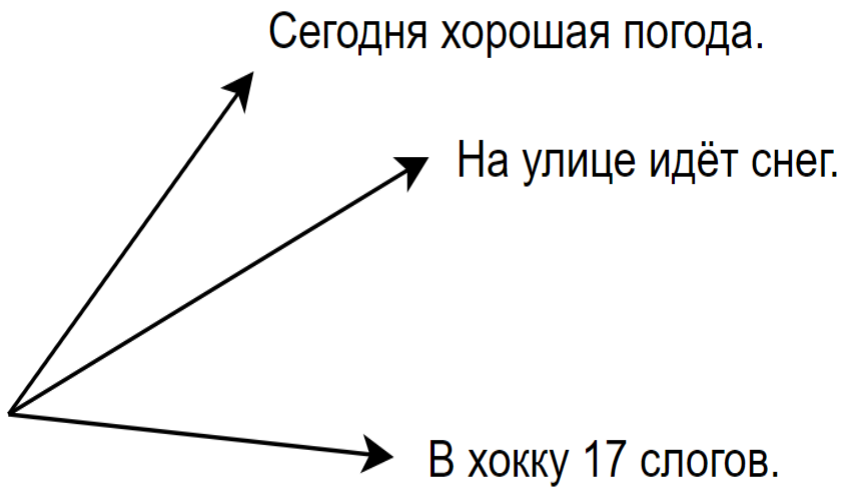
Векторные представления предложения и контекстно-зависимые модели слова

Белянова Марина Александровна, ИУ5-41А

Что такое векторное представление предложения?



- Вектор небольшой размерности (~300 компонент)
- Размерность вектора не зависит от количества слов в предложении
- Каждому предложению соответствует один вектор
- Чем ближе предложение по смыслу, тем ближе вектора предложений

Зачем нужны векторные представления предложения?

- Измерять близость между предложениями
  - Похожие заголовки новостей
  - Похожие реплики в диалоговой системе
- Хорошее признаковое пространство для классификации предложений

Усреднение векторных представлений слов

- Усреднение $\left\{\begin{matrix}[w_{11}, w_{12}, w_{13}, \ldots, w_{1n}], \\ [w_{21}, w_{22}, w_{23}, \ldots, w_{2n}], \\ [w_{31}, w_{32}, w_{33}, \ldots, w_{3n}]\end{matrix}\right\} = \left\{\begin{matrix}\overrightarrow{\text{мама}}, \\ \overrightarrow{\text{мыла}} \\ \overrightarrow{\text{раму}}\end{matrix}\right\}$

$$[\frac{w_{11}, w_{21}, w_{31}}{3}, \frac{w_{12}, w_{22}, w_{32}}{3}, \ldots, \frac{w_{1n}, w_{2n}, \ldots, w_{3n}}{3}]$$

Вектор предложения

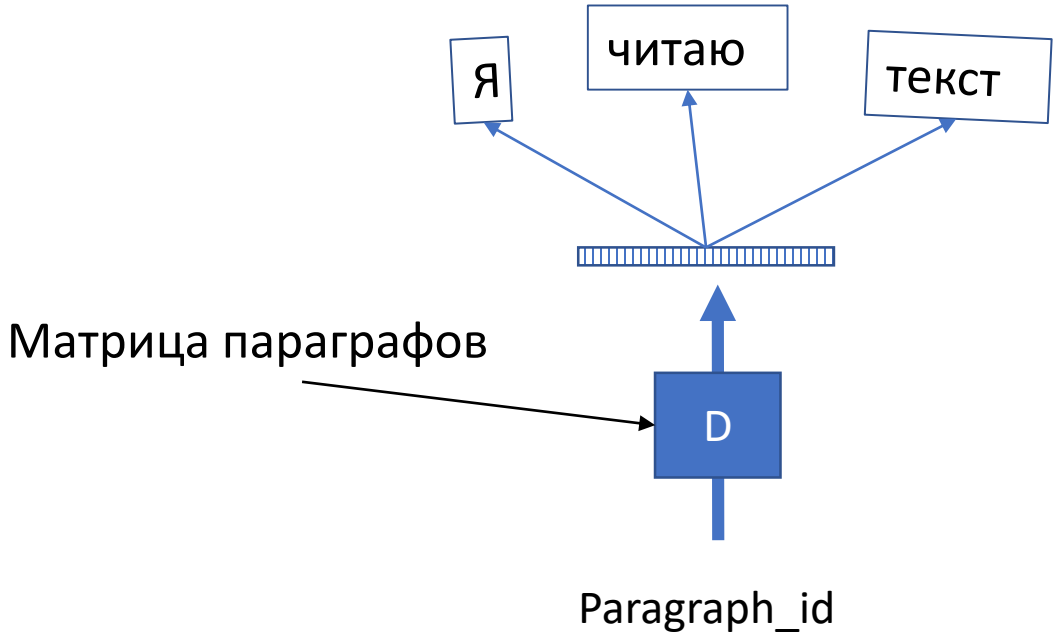
$\overrightarrow{\text{Мама мыла раму}}$

Усреднение векторных представлений слов

- Преимущества:
  - Быстрый метод
  - Можно использовать предобученные вектора слов
- Проблемы:
  - Равный вклад каждого слова в вес предложения (решается tf-idf весами)
  - Сохраняются все проблемы векторов слов
    - Многозначность
    - Близость синонимов и антонимов

Модель doc2vec

- Модель на основе word2vec:
  - Предсказать по id предложения (параграфа) и контексту следующее слово
- Проблема
  - Как получить вектор предложения, которого не было в обучающем множестве?

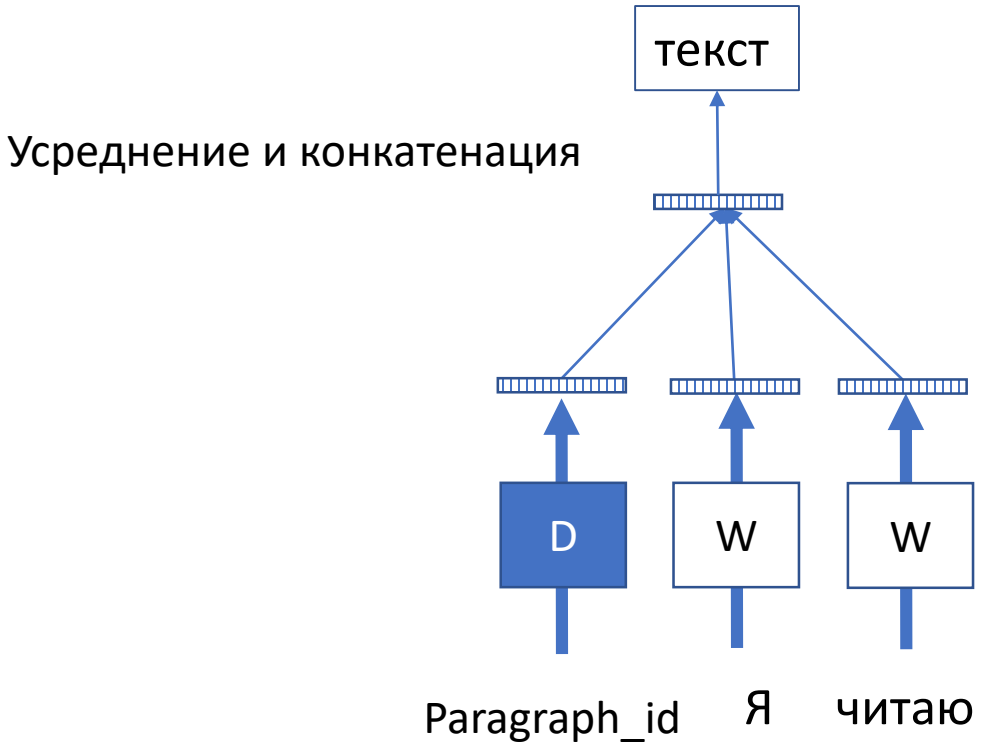


Распределённый мешок слов
(Distributed bag of words – DBoW)

https://arxiv.org/pdf/1405.4053.pdf

Модель doc2vec

- Модель на основе word2vec:
  - Предсказать по id предложения (параграфа) и контексту следующее слово
- Проблема
  - Как получить вектор предложения, которого не было в обучающем множестве?

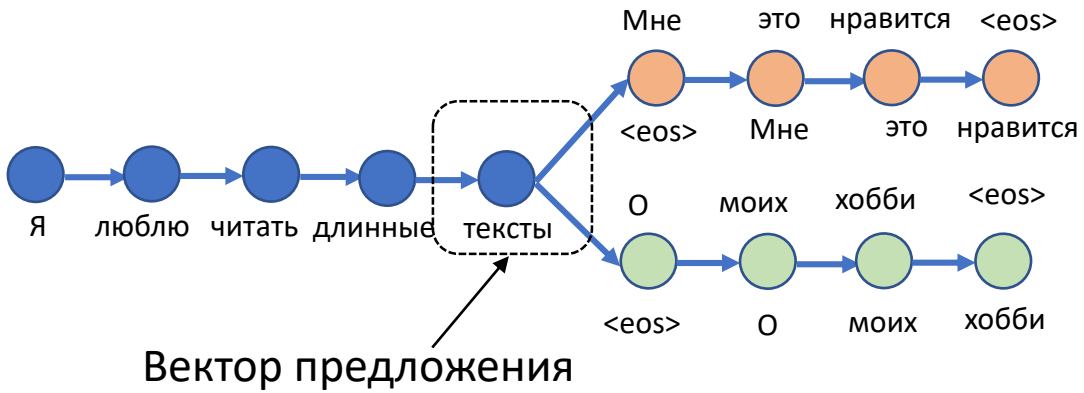


Распределённая память (Distributed memory – DM)

https://arxiv.org/pdf/1405.4053.pdf

Модель skip-thought

- Модель на основе архитектуры кодировщик-декодировщик

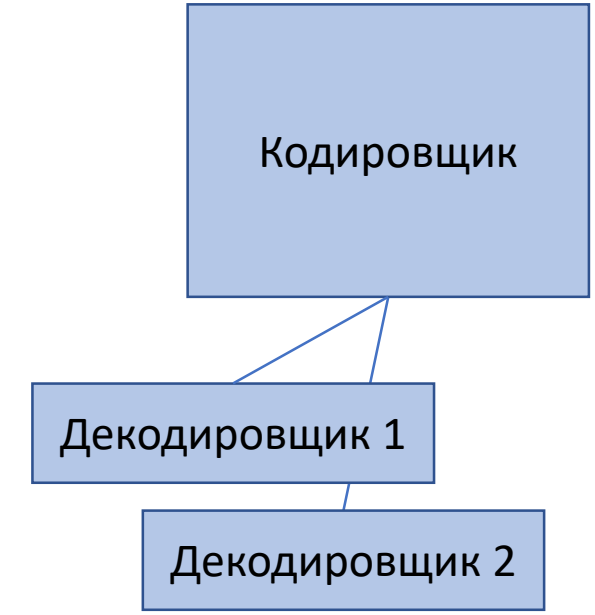


- По предложению s_i предсказываем предыдущее предложение s_{i-1} и следующее s_{i+1}

https://arxiv.org/pdf/1506.06726.pdf

Модель универсального кодировщика

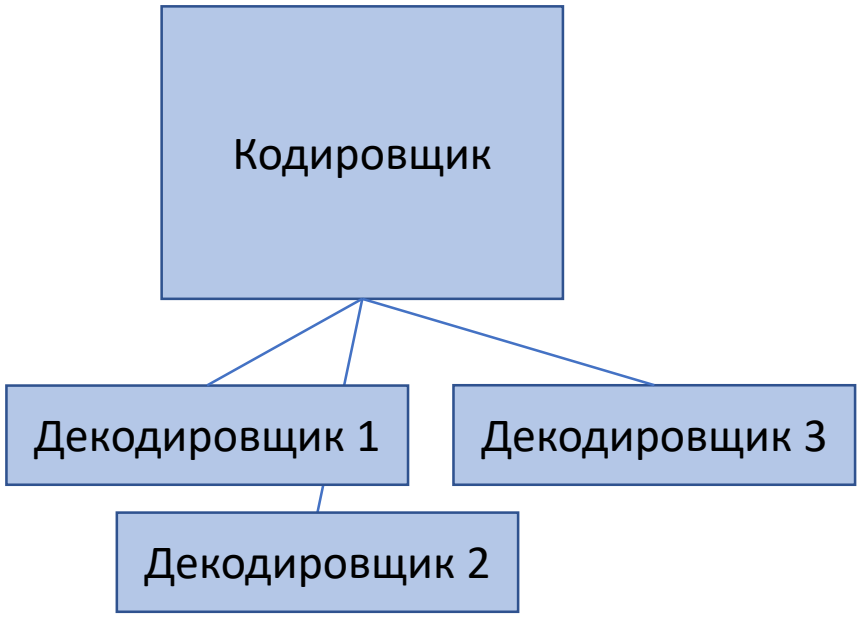
- Модель на основе архитектуры кодировщик-декодировщик
- Кодировщик состоит из блоков-трансформеров
- Несколько декодировщиков для
  - Частичного обучения: предсказания следующего предложения (частичное обучение)
  - Обучения с учителем: предсказание логической связи между предложениями (natural language inference, NLI)



https://arxiv.org/pdf/1803.11175.pdf

Модель универсального кодировщика

- Перенос обучения
  - Допустим, что кодировщик уже обучен
  - Добавим новый декодировщик 3 для определения тональности предложения
  - Декодировщики 1 и 2 играли вспомогательную роль в обучении кодировщика
- Декодировщик 3 используется для решения практической задачи
- Как правило, даёт прирост качества по сравнению с обучением с нуля

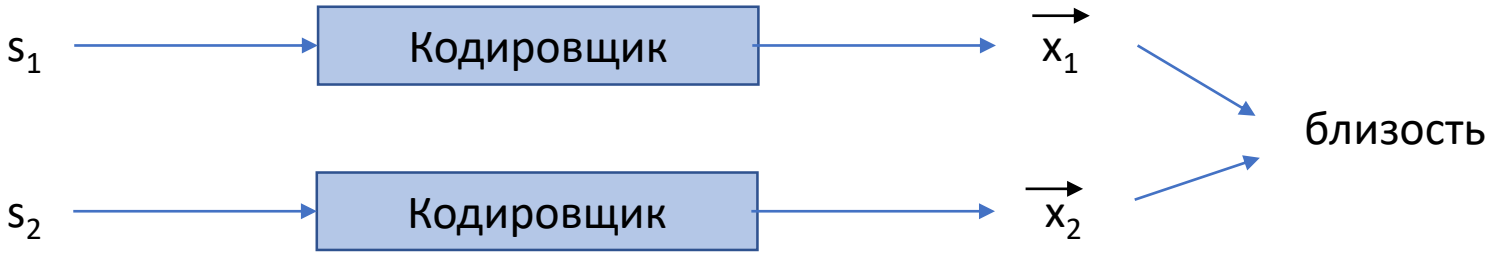


https://arxiv.org/pdf/1803.11175.pdf

Модель универсального кодировщика

- Кодировщик независимо выдаёт вектора предложений
- Близость между предложениями определяется как

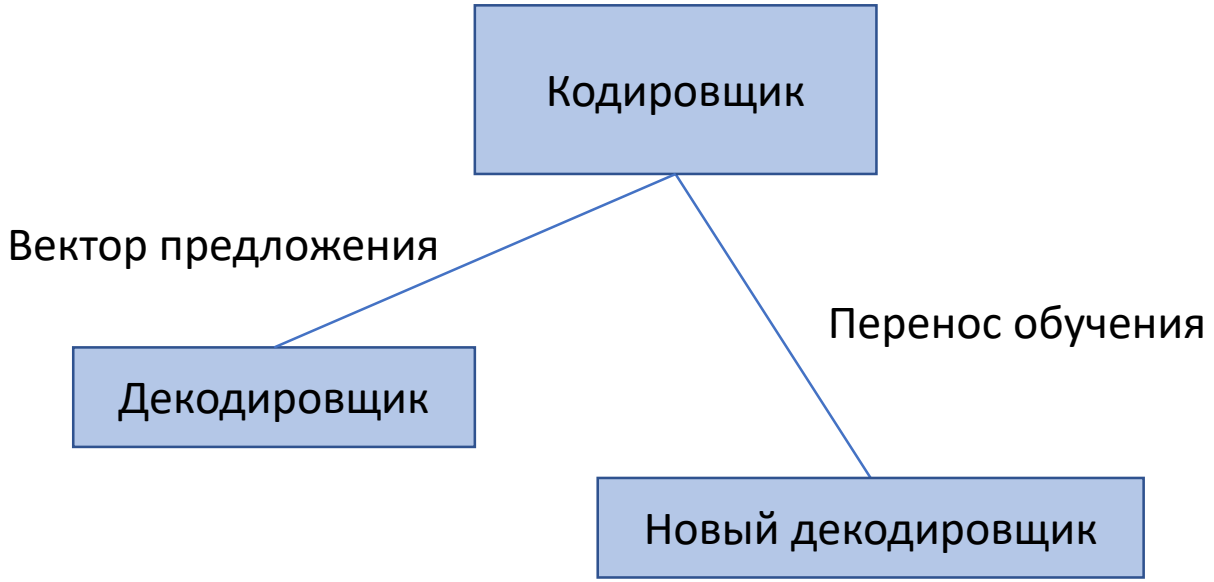
$$близость(s_1, s_2) = 1 - \arccos\left(\frac{\overrightarrow{x_1}, \overrightarrow{x_2}}{\left\|\overrightarrow{x_1}\right\|, \left\|\overrightarrow{x_2}\right\|}\right) / \pi$$



https://arxiv.org/pdf/1803.11175.pdf

Векторное представление предложения

- Архитектура кодировщик-декодировщик
- Используются как целевые функции с учителем, так и без учителя
- Показывают хорошие результаты при переносе обучения



Вопрос

Применение модели doc2vec на практике затруднено...

а) медленным обучением модели

б) необходимостью использовать слишком большие данные для обучения

в) недостатком обучающих данных

г) тем, что модель не умеет работать с предложениями, которых не видела в процессе обучения

Перенос обучения

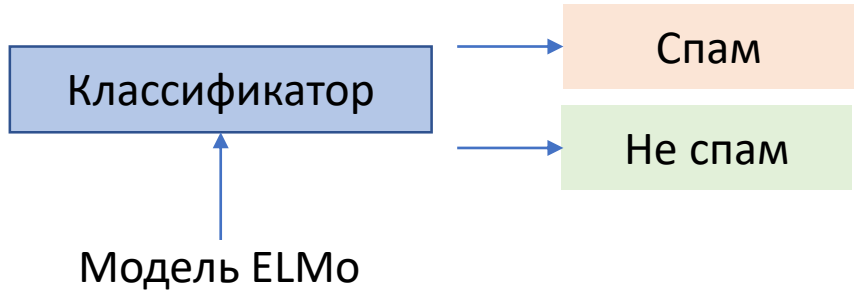
1) Частичное обучение

Модель ELMo

Данные: неразмеченные тексты

Задача: языковое моделирование

2) Обучение с учителем



Данные: размеченные тексты

Модель ELMo

- Вход модели – слова
- Выход модели – контекстно-зависимые векторные представления слов
- Векторное представление предложения: усреднение векторных представлений слов

https://arxiv.org/pdf/1802.05365.pdf

Модель ELMo. Обучение модели

- Два LSTM-слоя
- Языковая модель $p(word_i|word_{1:i-1})$

https://arxiv.org/pdf/1802.05365.pdf

Модель ELMo. Как получить контекстно-зависимое представление слова?

- Конкатенируем скрытые слои

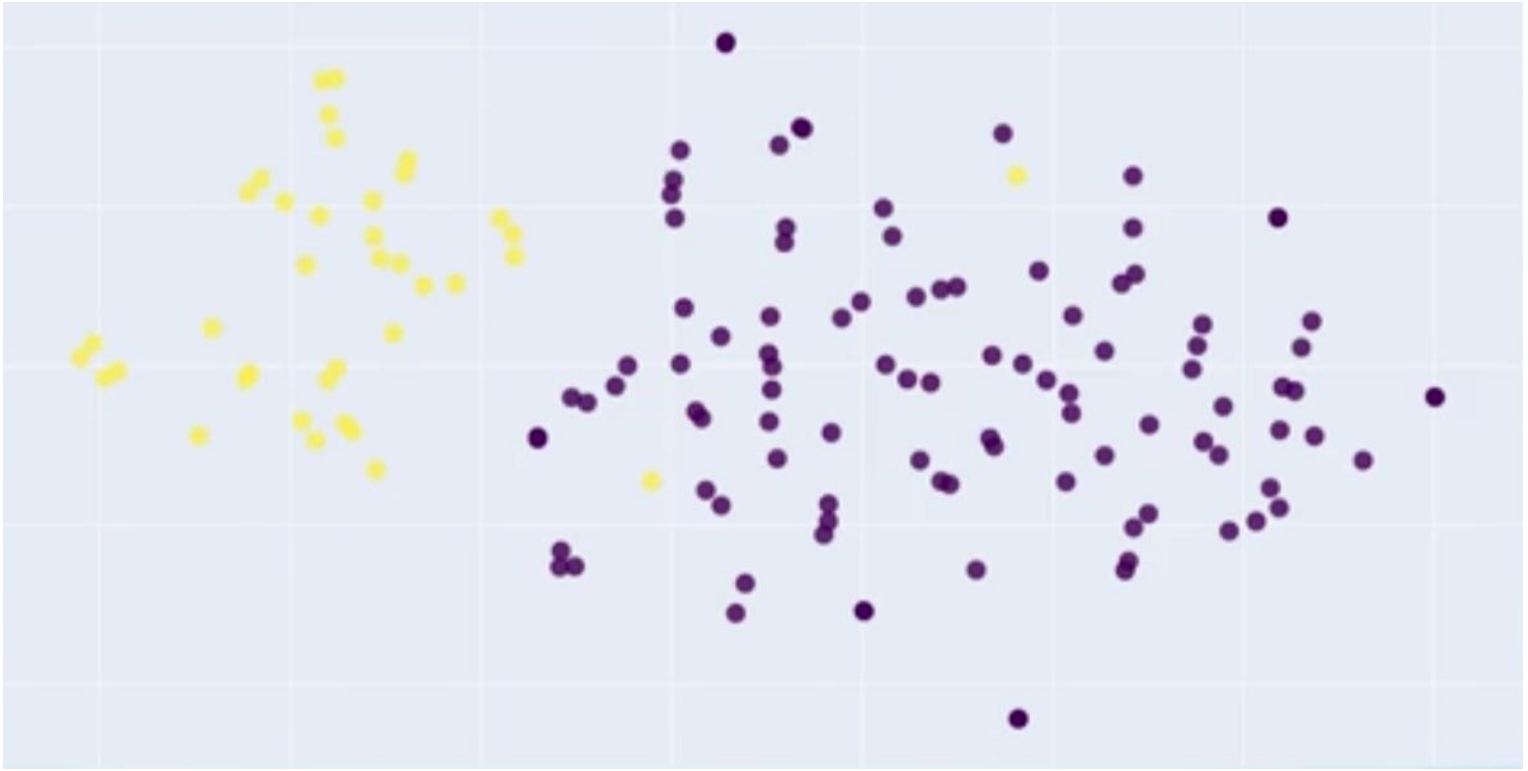
- Взвесим каждый слой

- Сложим с весами

Веса специфичны для задачи и обучаются вместе с моделью

Модель ELMo

- Проекция векторов ELMo слова «замок» на двумерную плоскость: мы видим, как разделяются смыслы



Модель ELMo

Модель ELMo

Модель ELMo


воен. Приспособление для производства выстрела в некоторых видах огнестельного оружия
Бойцы положили на траву пять пулеметов и замки с трех орудий

Модель ELMo

Демо модель ELMo

demo.allennlp.org

Вопрос

Основной тип слоев, использованных в модели ELMo — это:

а) сверточные

б) рекуррентные

3) полносвязные

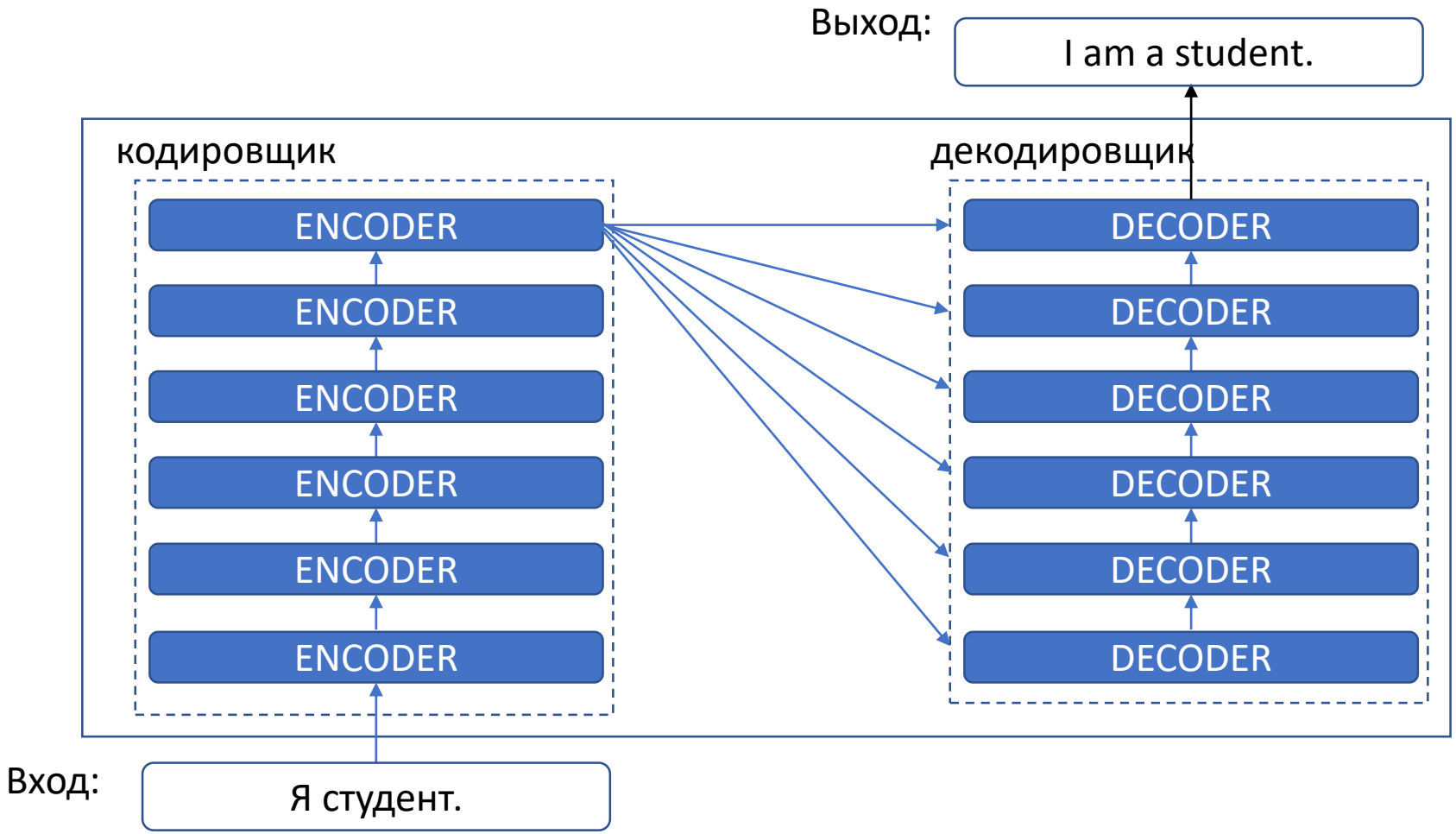
4) трансформеры

Модель BERT

- Появилась в конце 2018 года и произвела настоящий прорыв, установив новые рекорды в большей части задач
- Новая парадигма в обработке текстов: дообучаем большую предобученную модель для частных задач
- Предобученная модель уже много знает о языке
- Использование предобученных моделей открывает новые перспективы: меньше данных, показатели качества выше, время дообучения меньше
- Однако предобучение может длиться до нескольких недель

https://arxiv.org/pdf/1810.04805.pdf

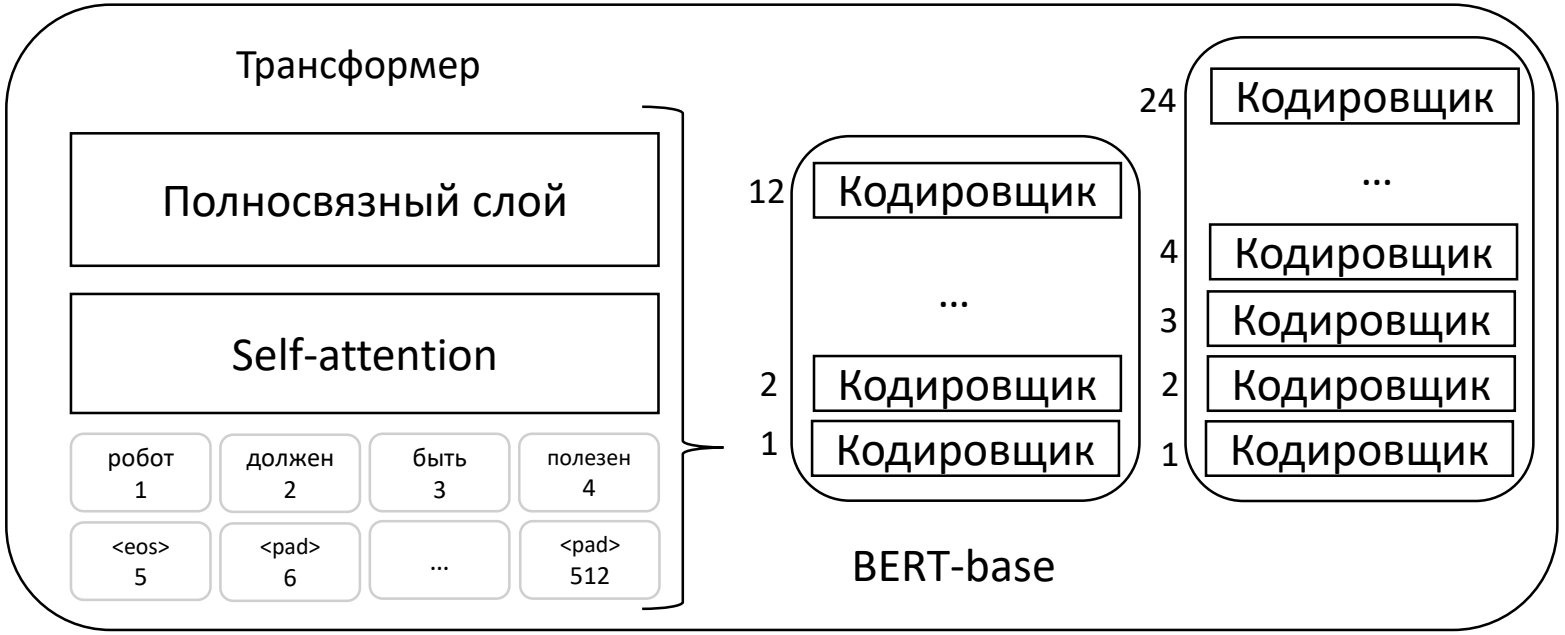
Архитектура Трансформер



https://arxiv.org/pdf/1706.03762.pdf

Модель BERT

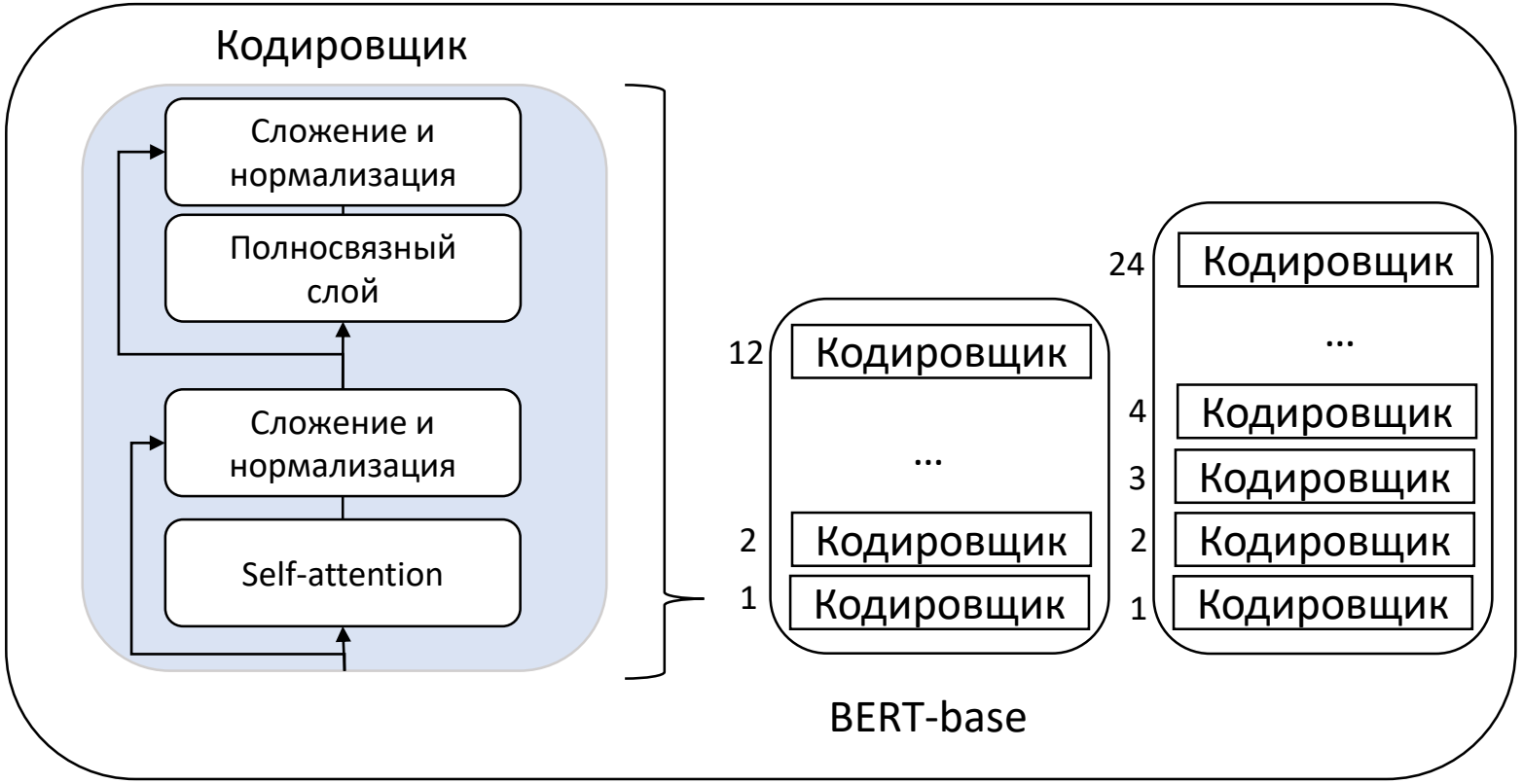
- Модель основана на кодировщике
- Две конфигурации: базовая и расширенная



https://arxiv.org/pdf/1810.04805.pdf

Модель BERT

- Механизм внимания



https://arxiv.org/pdf/1810.04805.pdf

Модель BERT

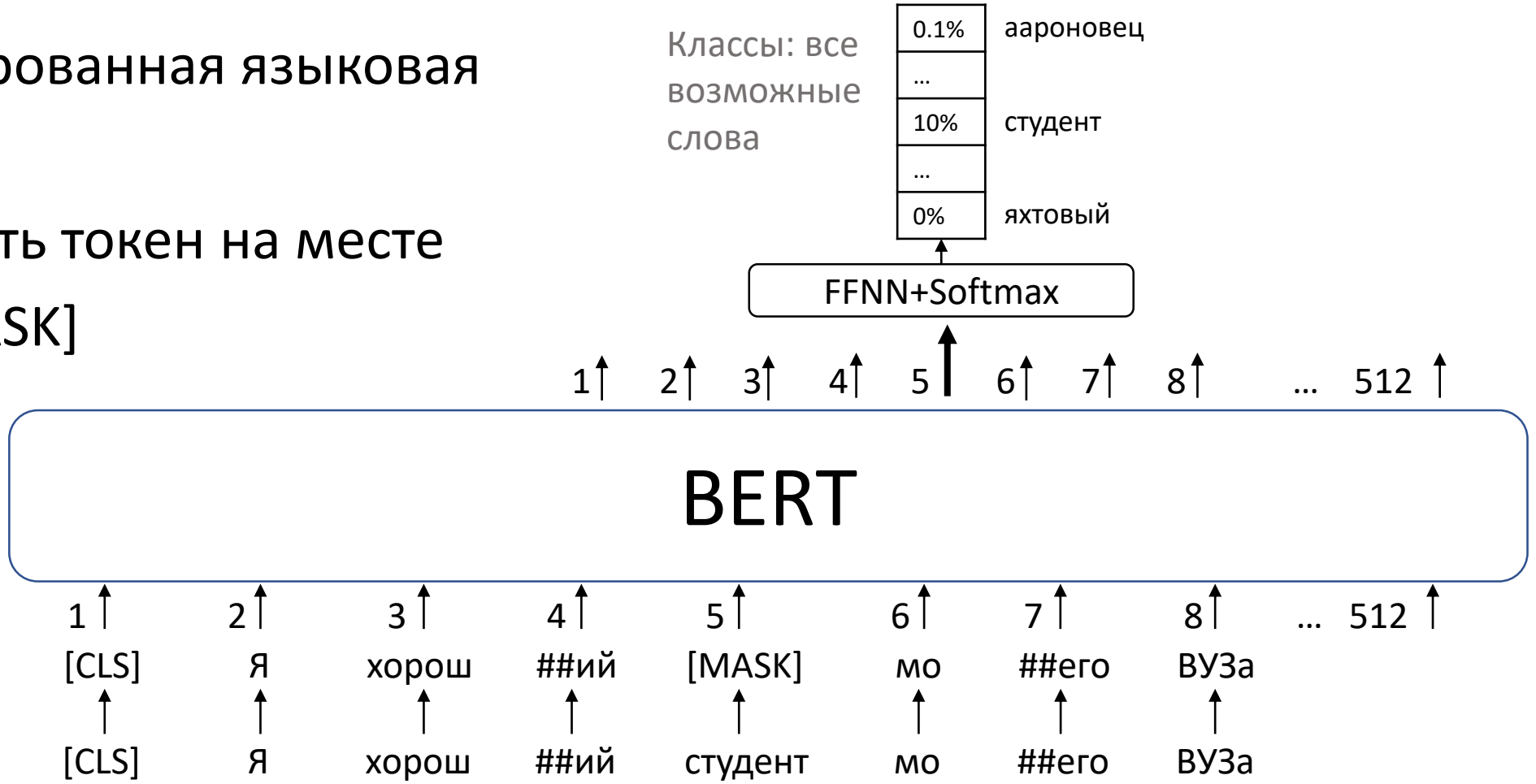
- Два вида токенов: подслова и управляющие токены
- Токен [CLS] всегда находится на первой позиции
- Вектор токена [CLS] – векторное представление предложения
- Остальные токены – векторные представления подслов

https://arxiv.org/pdf/1810.04805.pdf

Обучение BERT

1) Маскированная языковая модель

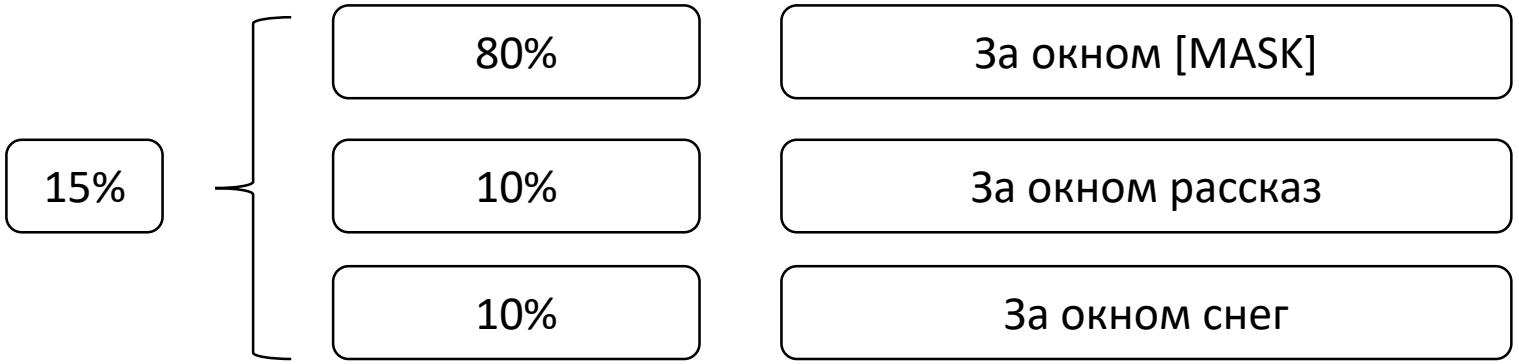
Предсказать токен на месте маски [MASK]



https://arxiv.org/pdf/1810.04805.pdf

Обучение BERT

1) Маскированная языковая модель



https://arxiv.org/pdf/1810.04805.pdf

Обучение BERT

1) Предсказание следующего предложения

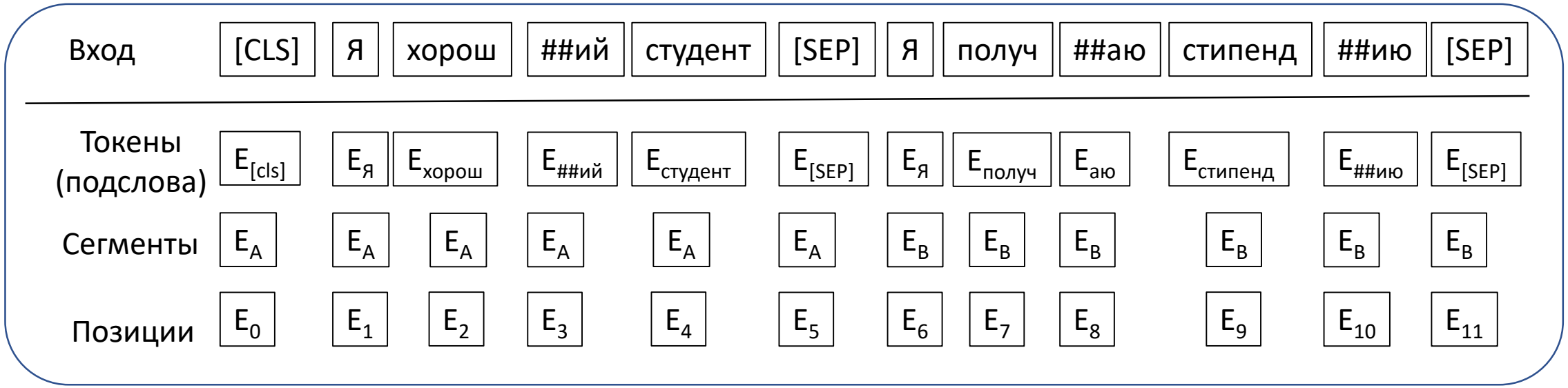
Предсказать, следует ли следующее предложение за другим, предложения соединены токеном [SEP] и разделены на сегменты A и B

https://arxiv.org/pdf/1810.04805.pdf

Обучение BERT: две задачи одновременно

- Вход модели Размерность входа 512 токенов

Управляющие токены



Подслова

- Попарное битовое кодирование (Byte pair encoding, BPE)
- Считаем частоты пар символов
- Склеиваем самую частую пару символов и превращаем её в новый символ
- Продолжаем повторять операцию фиксированное число раз

https://arxiv.org/pdf/1710.02187.pdf

Общая схема

- Обучить языковую модель ->
- Дообучить языковую модель ->
- Обучить классификатор

Как получить вектор предложения из модели BERT?

- [CLS] – вектор токена CLS на последнем слое
- [MEAN] – усреднение векторов слов на последнем слое
- [MAX] – покомпонентный максимум векторов на последнем слое

Сравнение моделей ELMo и BERT

Модель ELMo

- Языковая модель на основе BiLSTM
- Направление чтения текста не взаимодействуют в модели
- Показывает отличные результаты в задачах, требующих понимания синтаксиса и семантики

Модель BERT

- Кодировщик модели Трансформер
- Учитывает оба направления чтения текста
- Маскированная языковая модель
- Обучалась на большом объёме данных
- Применима в большем количестве задач

Демо маскированной языковой модели

demo.allennlp.org

Демо exBERT

exbert.net

Вопрос

Модель BERT построена по аналогии с архитектурой Трансформер, используемой для машинного перевода. BERT это:

а) кодировщик

б) декодировщик

Модель BERT позволяет получить вектора:

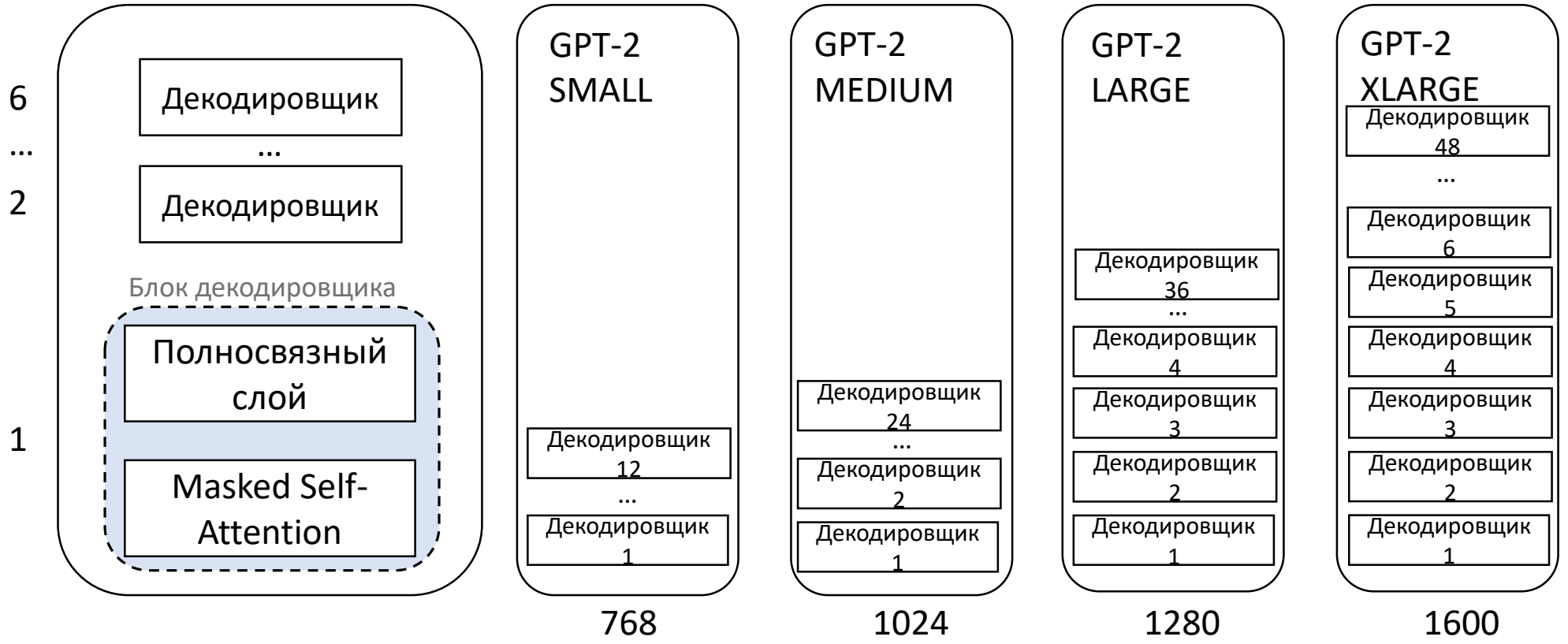
а) только предложений

б) только слов

в)и слов, и предложений

Модель GPT-2

- Модель основана на декодировщике модели Трансформер



https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Модель GPT-2

- Четыре конфигурации: разное количество слоёв и разные размерности слоёв.
- Используется маскированный механизм внимания

BERT	**GPT-2**
Механизм внимания смотрит на все токены слева и справа от текущего	Механизм внимания смотрит только на токены слева от текущего

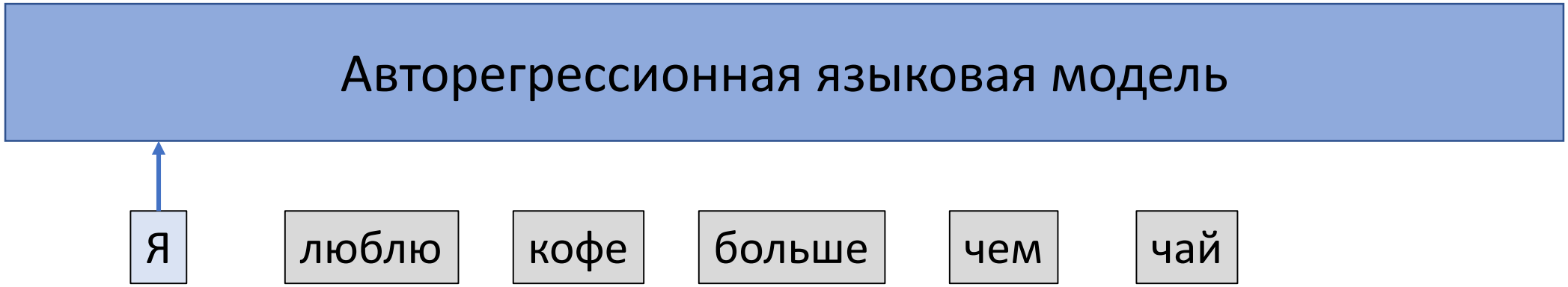
https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Демо GPT-2

demo.allennlp.org

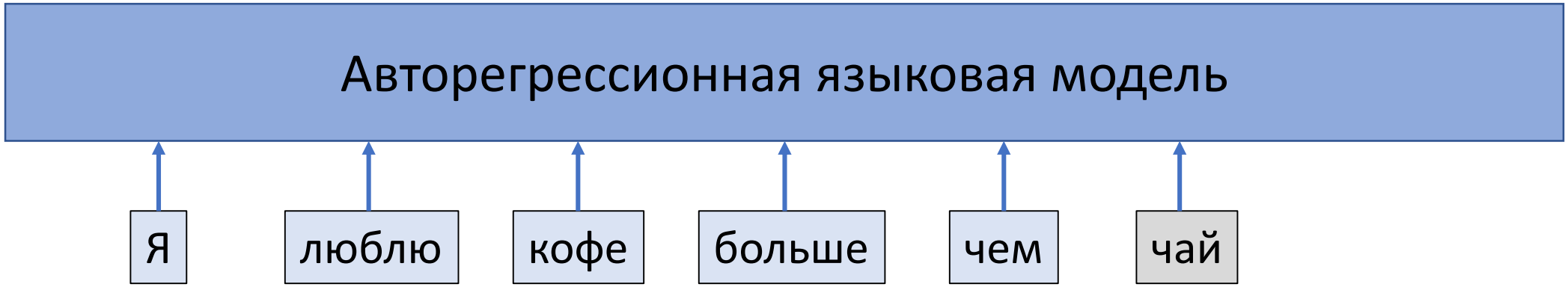
Задача языковой модели

- Так работает GPT-2



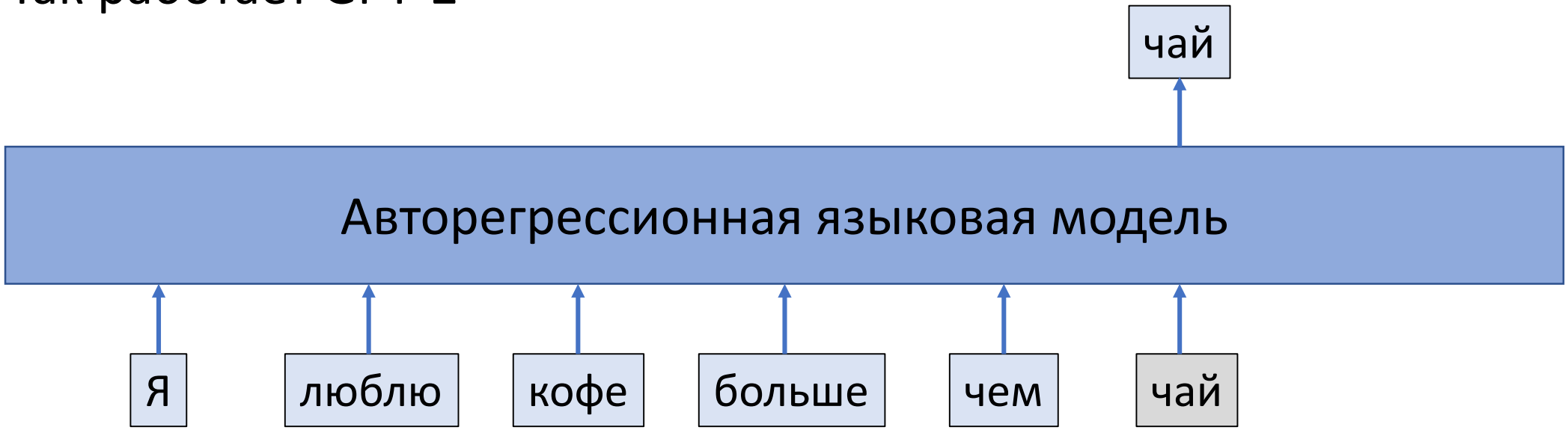
Задача языковой модели

- Так работает GPT-2



Задача языковой модели

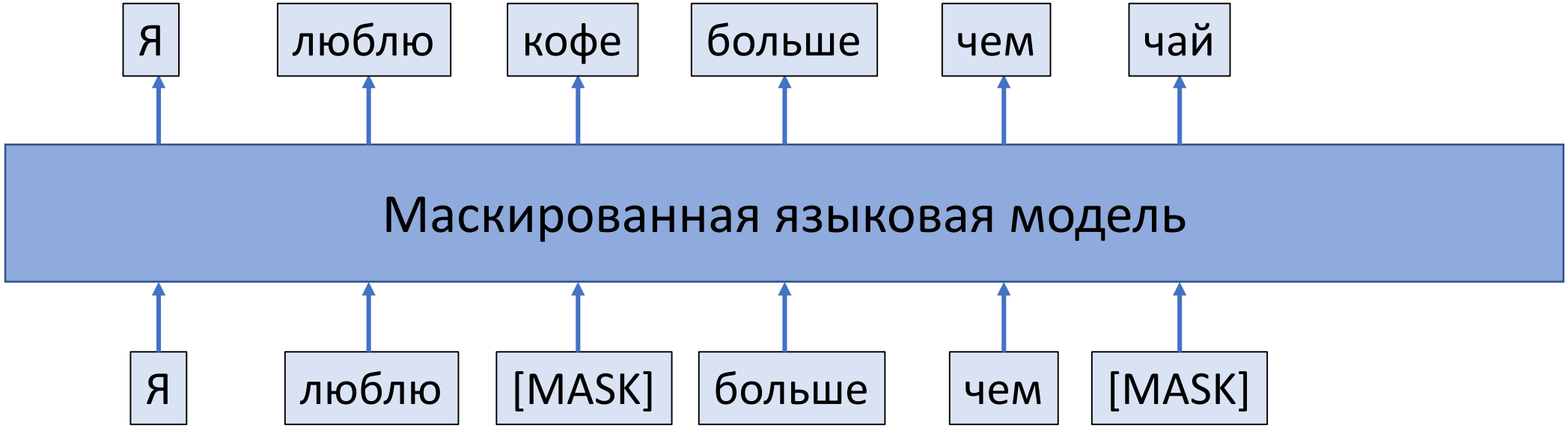
- Так работает GPT-2



Недостатки авторегрессионных моделей

- Не учитывается правый контекст
- Медленно работает

Так работает модель BERT

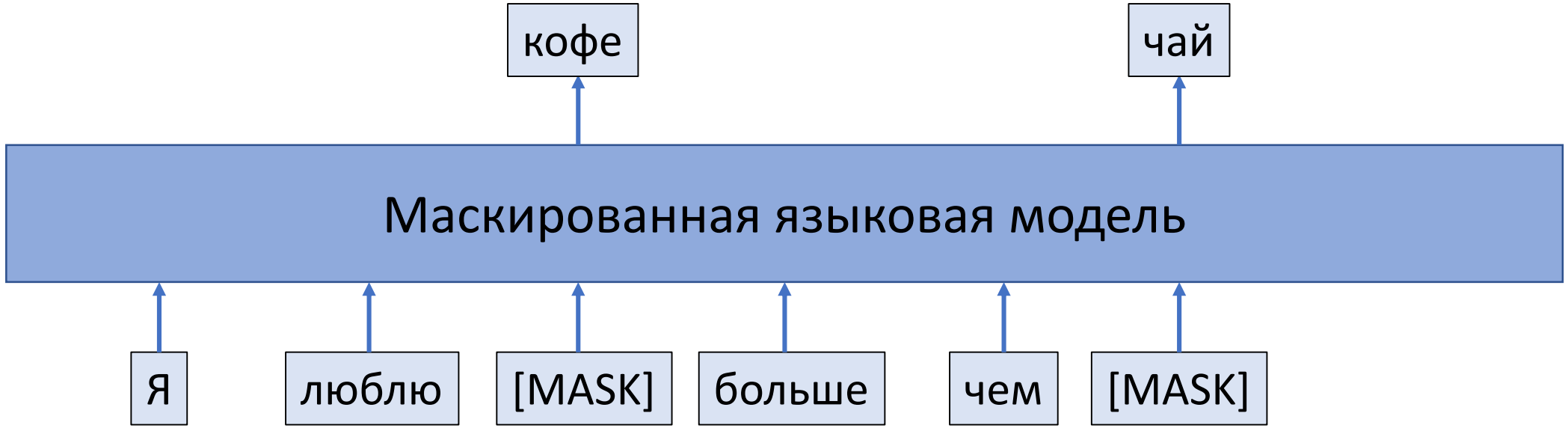


Недостатки маскированных моделей

- BERT не умеет обрабатывать очень длинные последовательности
- Управляющий токен [MASK] используется только во время обучения, но не используется во время дообучения или тестирования
  - Обучающее множество искажено по сравнению с тестовым
- BERT генерирует независимые предсказания

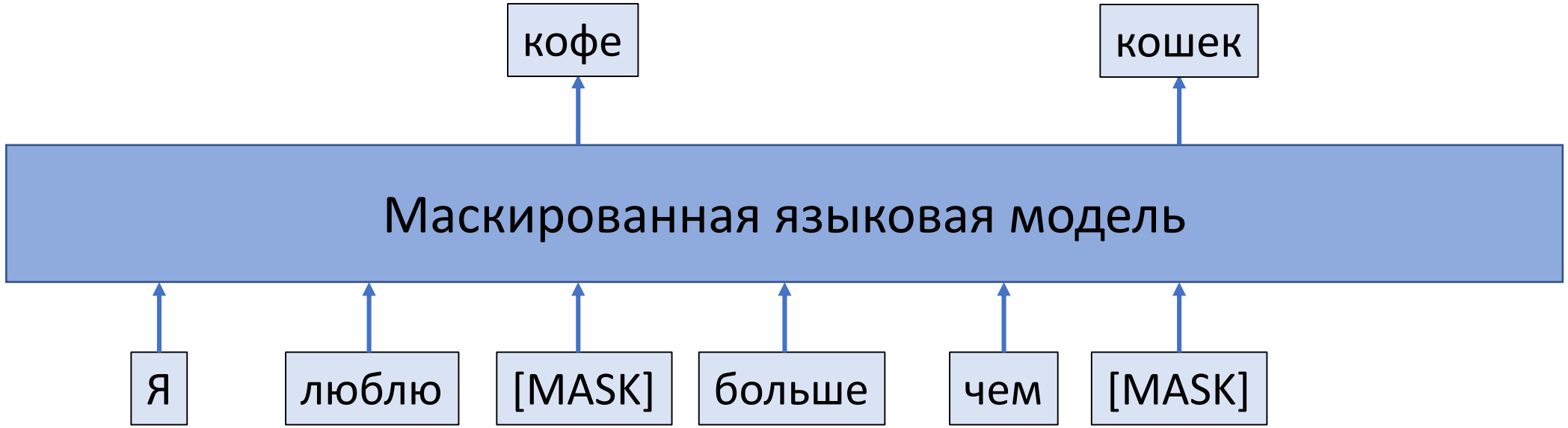
Задача языковой модели

BERT генерирует некорректное предложение



Задача языковой модели

BERT генерирует некорректное предложение



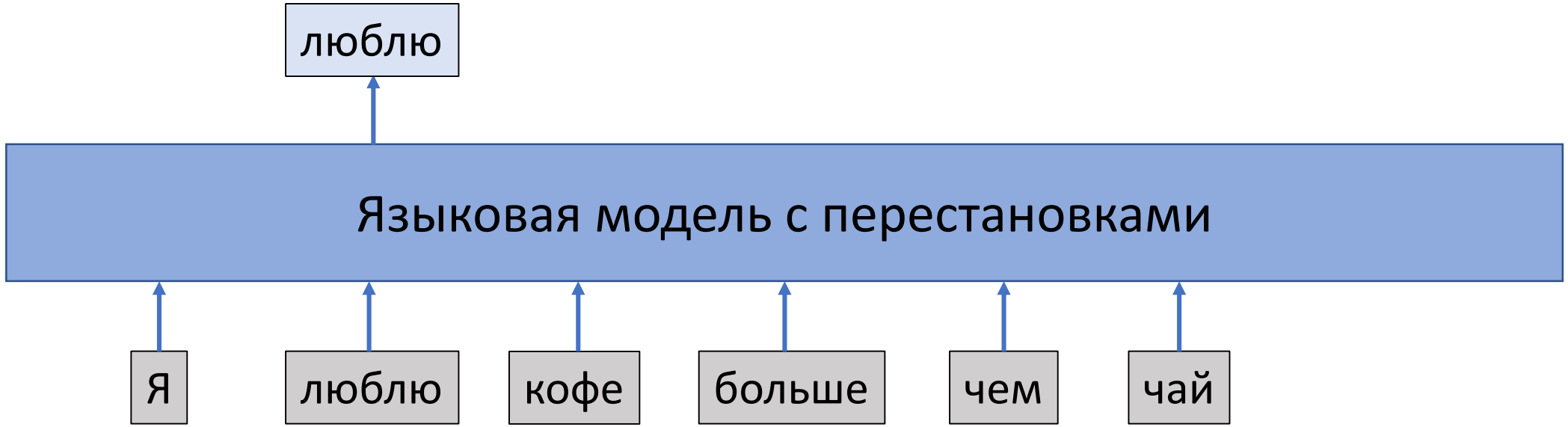
XLNet преодолевает проблемы BERT

- Языковая модель с перестановками

https://arxiv.org/pdf/1906.08237.pdf

Задача языковой модели

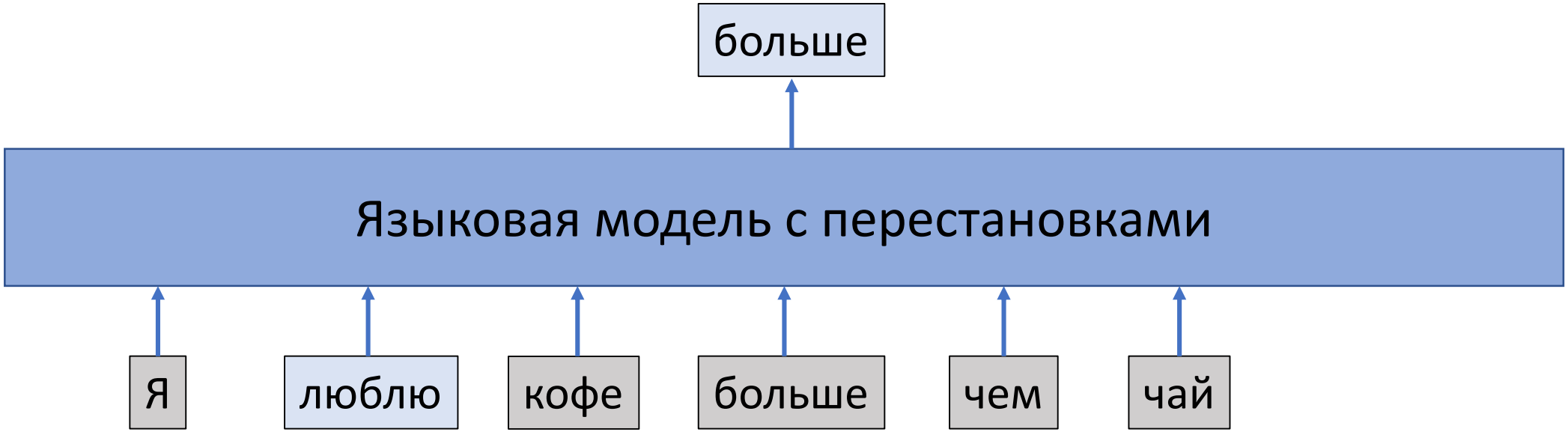
- Так работает модель XLNet



https://arxiv.org/pdf/1906.08237.pdf

Задача языковой модели

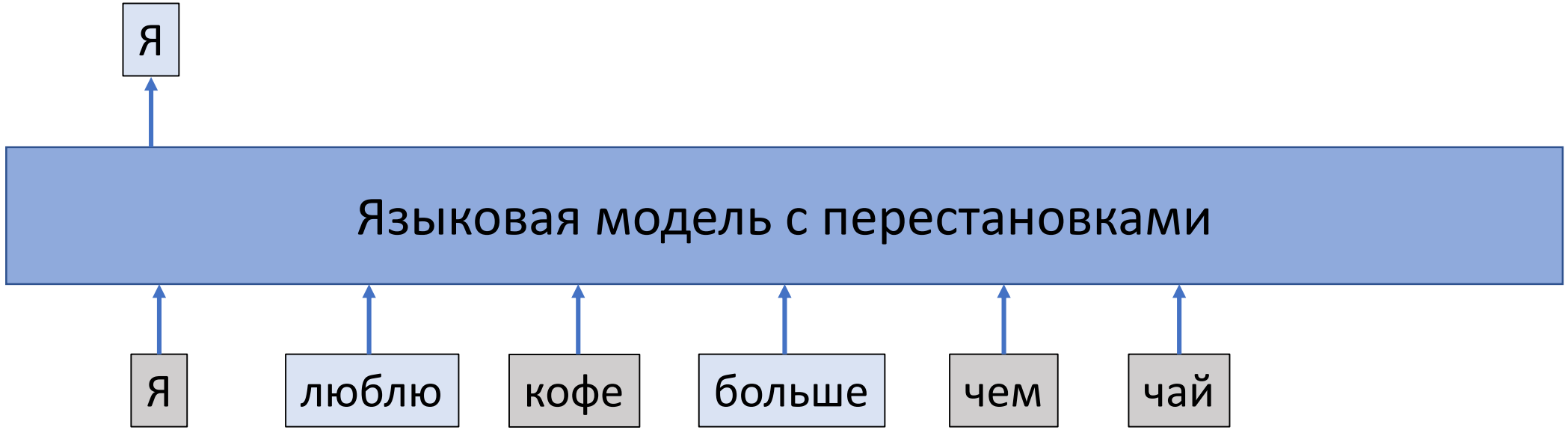
- Так работает модель XLNet



https://arxiv.org/pdf/1906.08237.pdf

Задача языковой модели

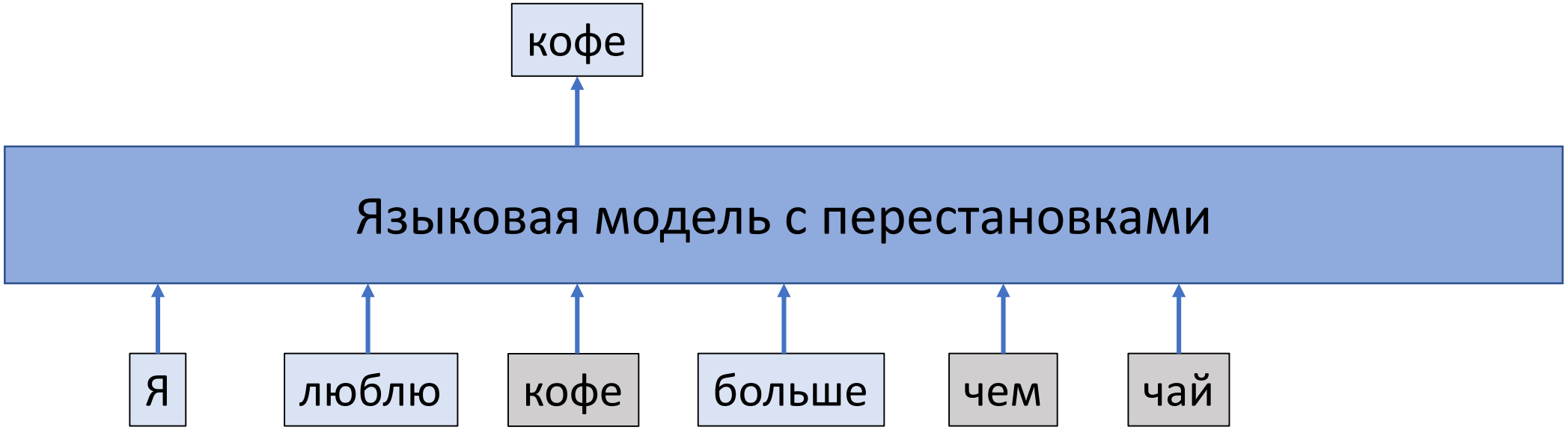
- Так работает модель XLNet



https://arxiv.org/pdf/1906.08237.pdf

Задача языковой модели

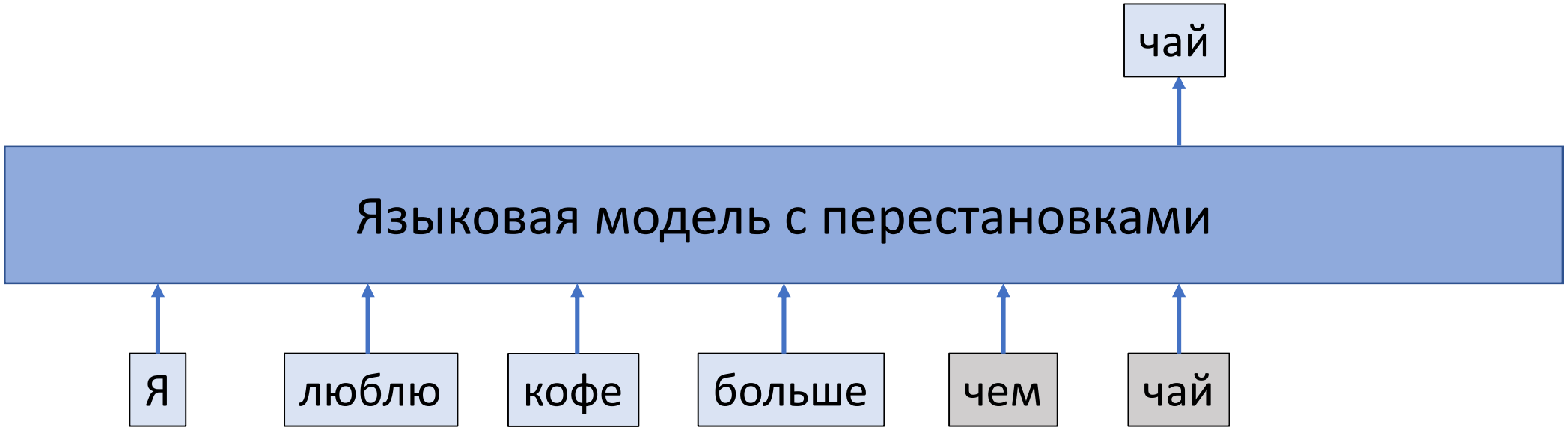
- Так работает модель XLNet



https://arxiv.org/pdf/1906.08237.pdf

Задача языковой модели

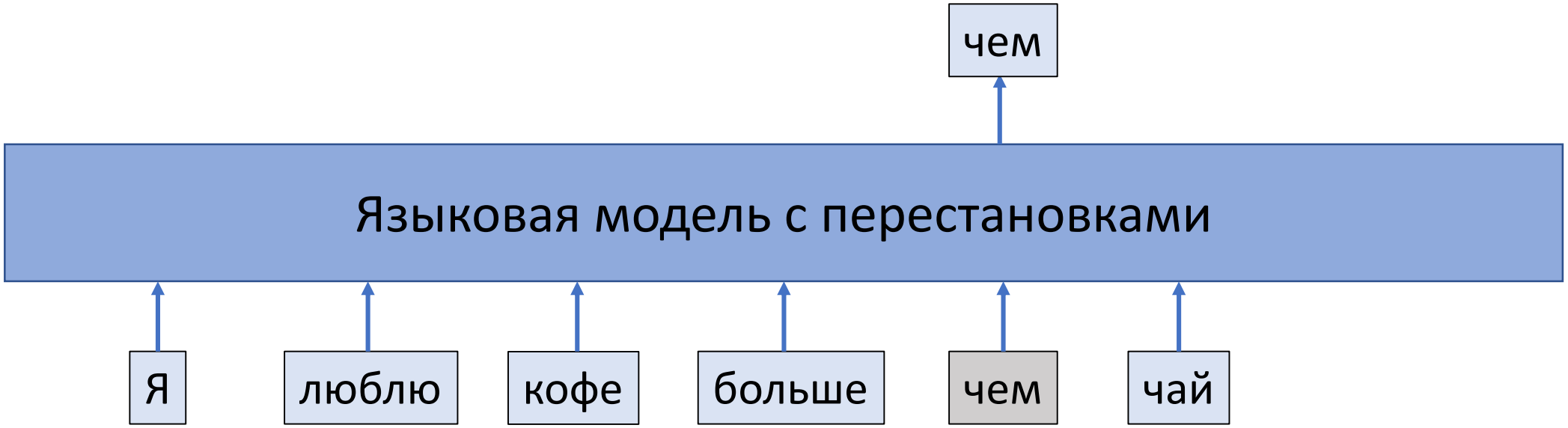
- Так работает модель XLNet



https://arxiv.org/pdf/1906.08237.pdf

Задача языковой модели

- Так работает модель XLNet



https://arxiv.org/pdf/1906.08237.pdf

XLNet преодолевает проблемы BERT

- XLNet – языковая модель с перестановками
- В модели XLNet есть память

https://arxiv.org/pdf/1906.08237.pdf

Память в модели XLNet

- Механизм внимания в модели BERT:

  $Self\text{-}attention(q=h_i, k=h_i, v=h_i)$

- Механизм внимания в модели XLNet, использующий память:

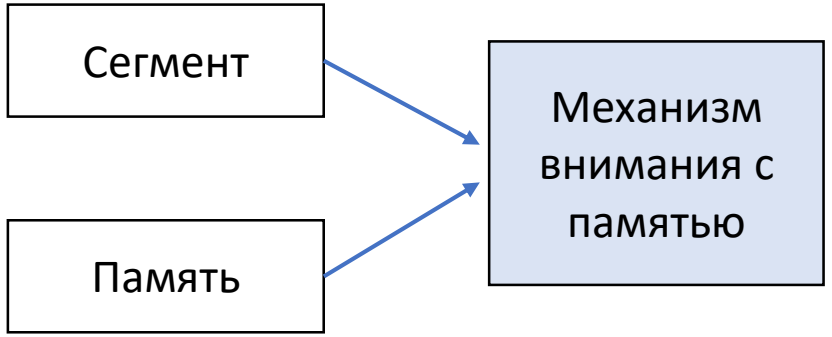
  $Self\text{-}attention(q=h_i, k=[h_i, m], v=[h_i, m])$

- Кэширование памяти: $m=h$

Я пошла в магазин.
Я купила коту корм.

сегменты

1. Я пошла в магазин.
2. Я купила коту корм.



https://arxiv.org/pdf/1906.08237.pdf

Память в модели XLNet

- Механизм внимания в модели BERT:

  $Self\text{-}attention(q=h_i, k=h_i, v=h_i)$

- Механизм внимания в модели XLNet, использующий память:

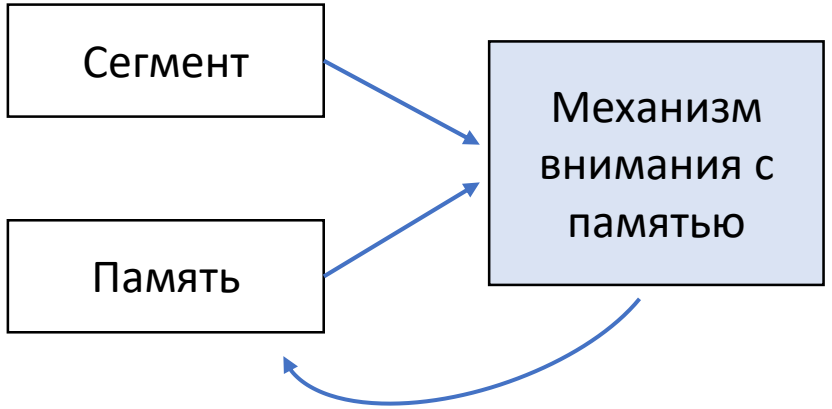
  $Self\text{-}attention(q=h_i, k=[h_i, m], v=[h_i, m])$

- Кэширование памяти: $m=h$

Я пошла в магазин.
Я купила коту корм.

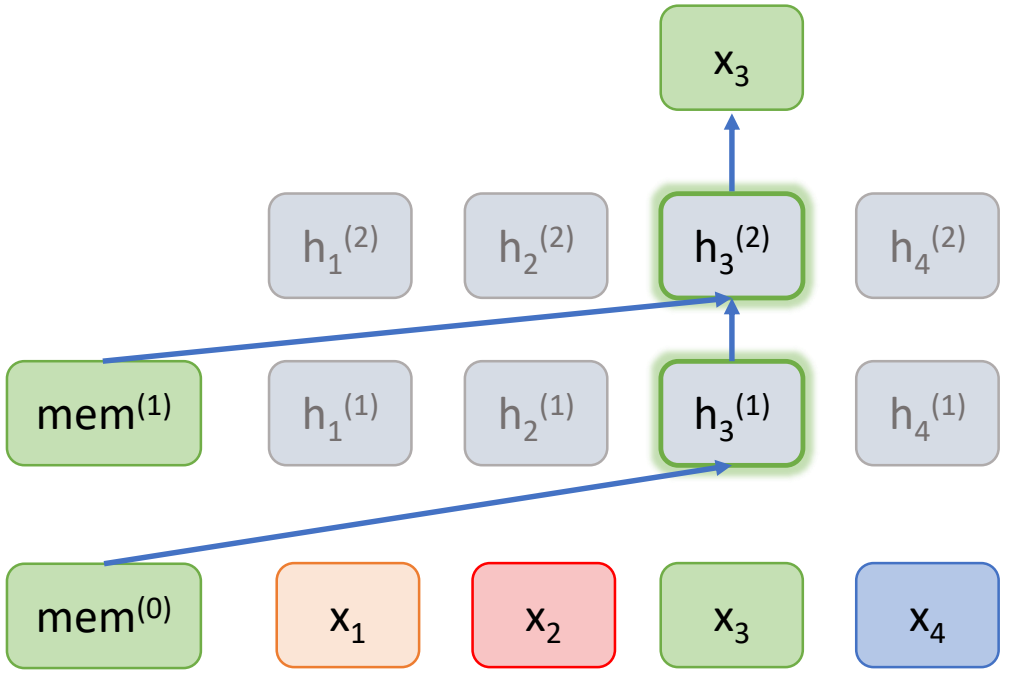
сегменты

1. Я пошла в магазин.
2. Я купила коту корм.



https://arxiv.org/pdf/1906.08237.pdf

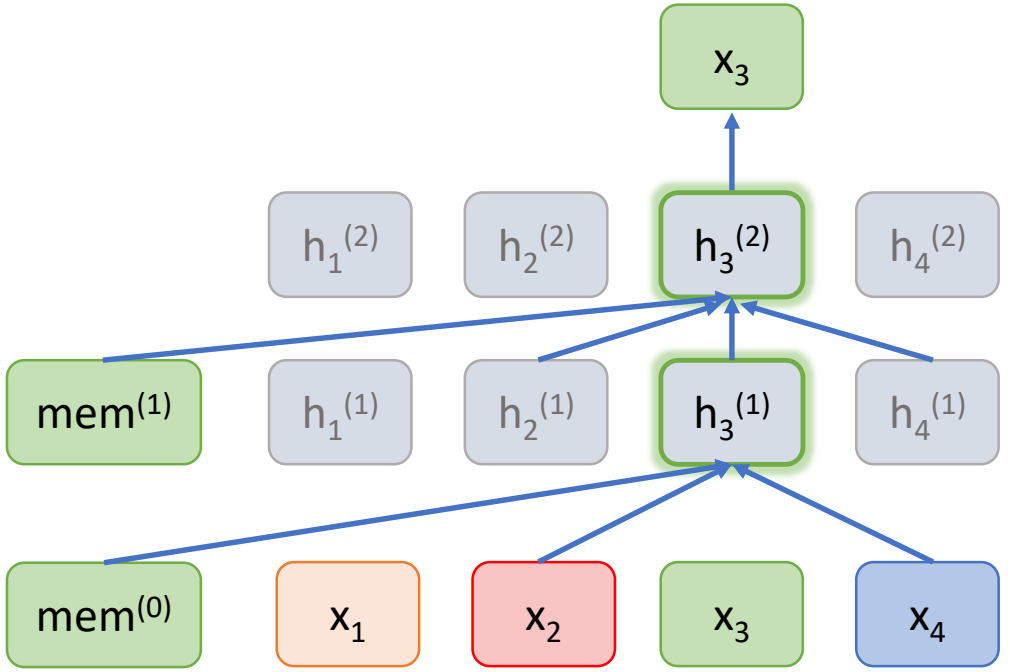
Память в модели XLNet



Порядок факторизации 3->2->4->1

https://arxiv.org/pdf/1906.08237.pdf

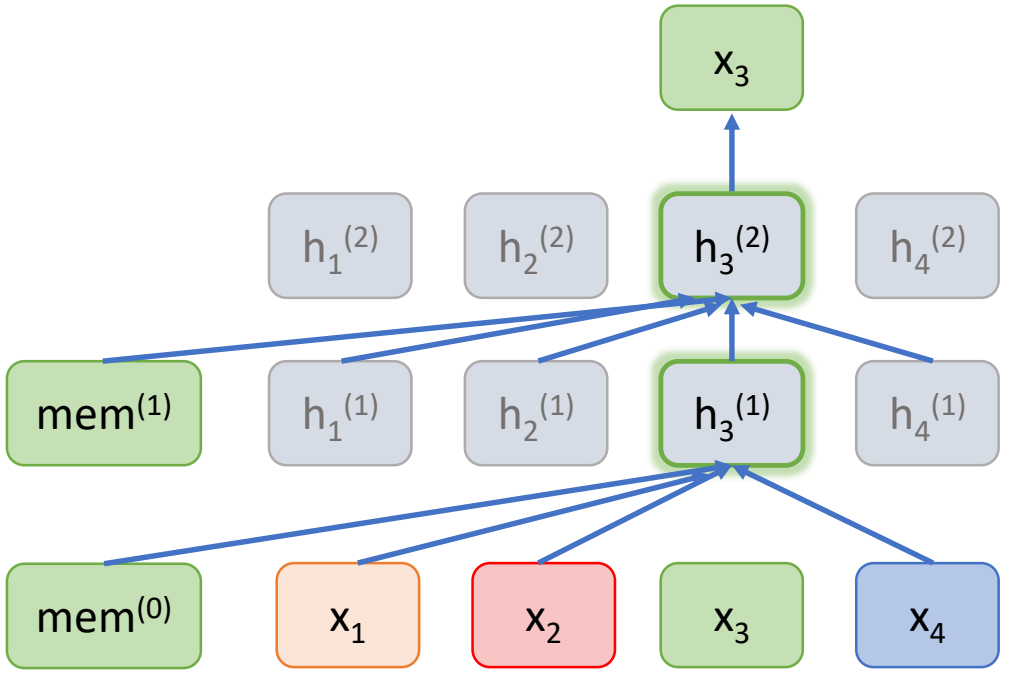
Память в модели XLNet



Порядок факторизации 2->4->3->1

https://arxiv.org/pdf/1906.08237.pdf

Память в модели XLNet



Порядок факторизации 1->4->2->3

https://arxiv.org/pdf/1906.08237.pdf

Относительные позиционные кодировки

	я	иду	в	магазин
я	0	1	2	3
иду				
в				
магазин				

https://arxiv.org/pdf/1906.08237.pdf

Относительные позиционные кодировки

	я	иду	в	магазин
я	0	1	2	3
иду	-1	0	1	2
в				
магазин				

https://arxiv.org/pdf/1906.08237.pdf

Относительные позиционные кодировки

	я	иду	в	магазин
я	0	1	2	3
иду	-1	0	1	2
в	-2	-1	0	1
магазин	-3	-2	-1	0

https://arxiv.org/pdf/1906.08237.pdf

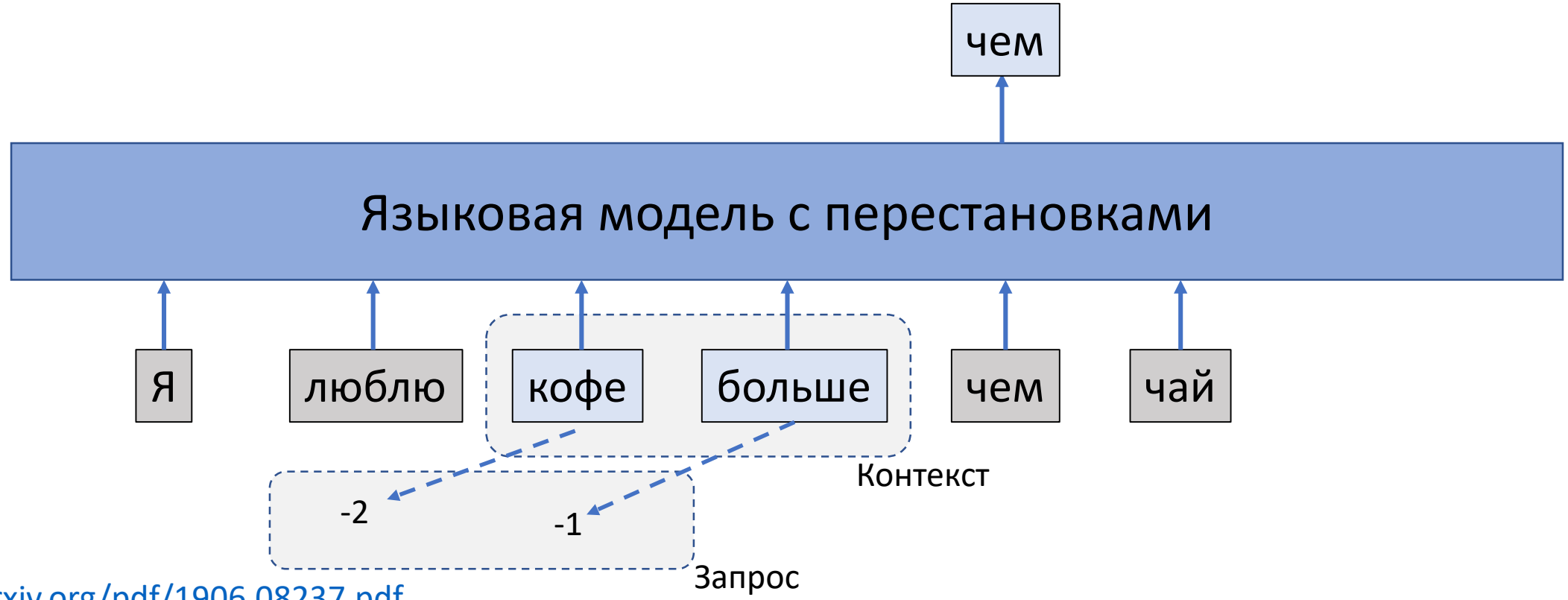
Двухпотоковый механизм самовнимания

- Two-Stream Self-Attention
  - Поток контекста $Attention(q=c_{z_t}, k=c_{z_{<=t}}, v=c_{z_{<=t}})$
  - Поток запроса $Attention(q=pos_{z_t}, k=c_{z_{<=t}}, v=c_{z_{<t}})$

  z_t – перестановка

Двухпотоковый механизм самовнимания

- Two-Stream Self-Attention



https://arxiv.org/pdf/1906.08237.pdf

Сравнение модели XLNet и BERT

	BERT	XLNet
Размер (млн. параметров)	Base: 100 Large: 340	Base: 110 Large: 340
Качество	Лучшее на 2019 год	+2-15% относительно BERT
Данные	16 Гб	Base: 16 Гб (корпус BERT) Large: +97 Гб
Обучение	Base: 8*V100 – 12 дней Large: 64 TPU – 4 дня	Large: 512 TPU – 3 дня
Целевая задача	Маскированная языковая модель + предсказание следующего предложения	Языковая модель с перестановками

Вопрос

Какой тип языковой модели используется в модели XLNet?

а) авторегрессионная

б) маскированная

в) безусловная

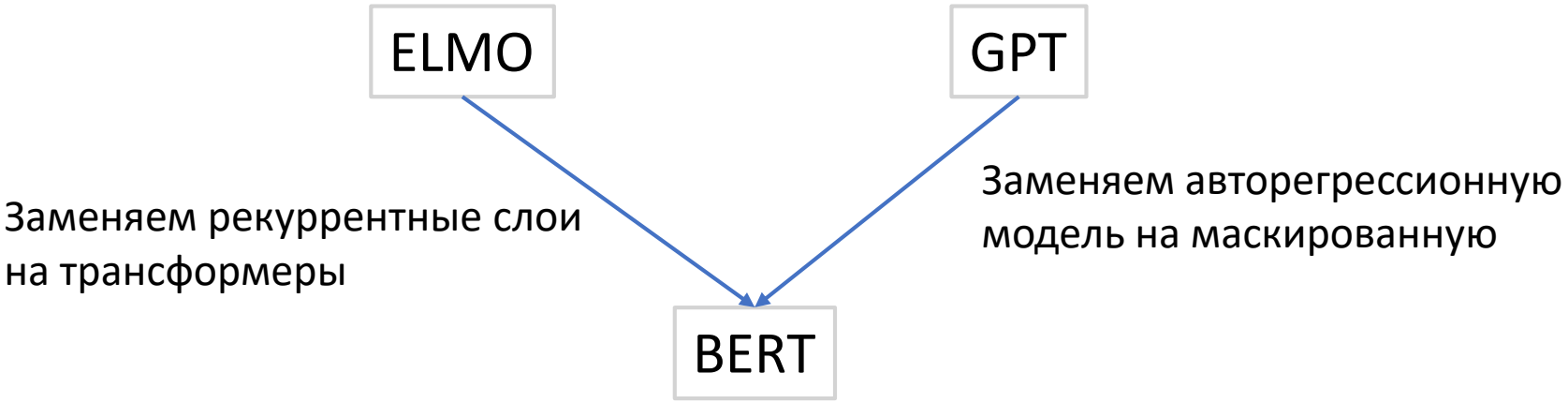
г) языковая модель с перестановками

BERT-like models

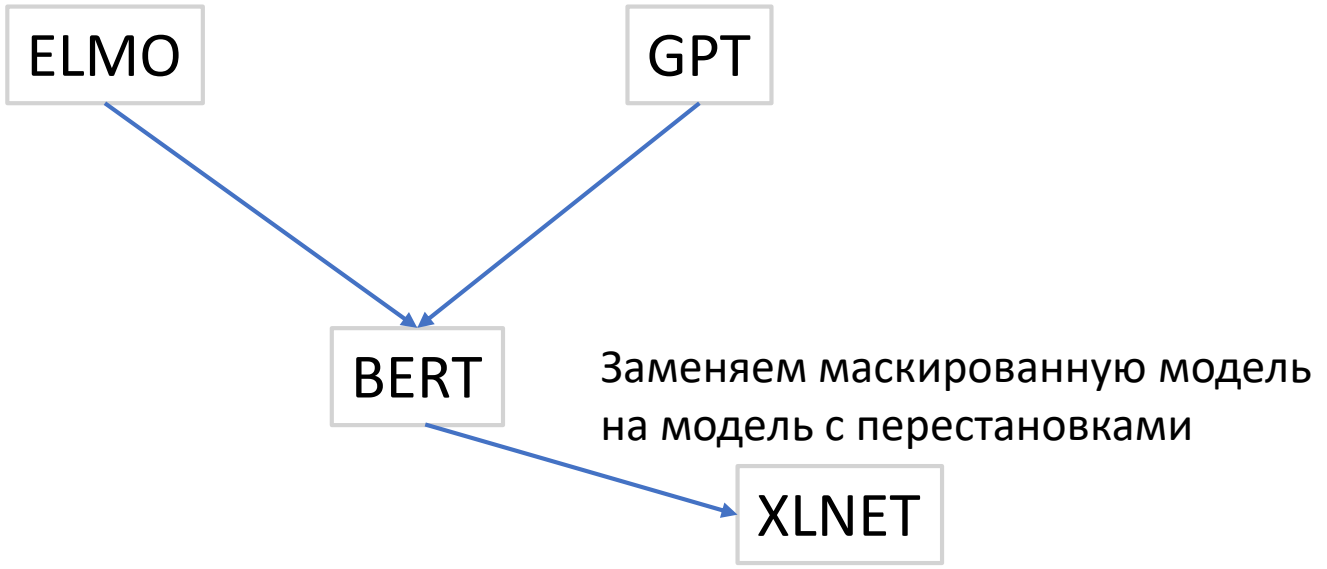
Рекуррентные нейронные сети ELMO

GPT Авторегрессионная языковая модель

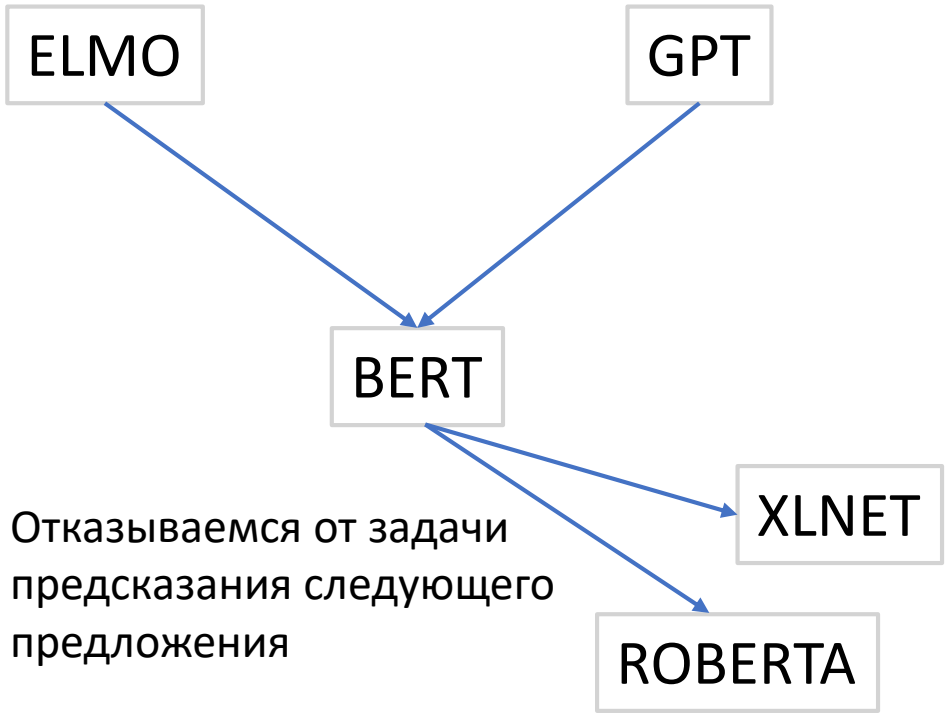
BERT-like models



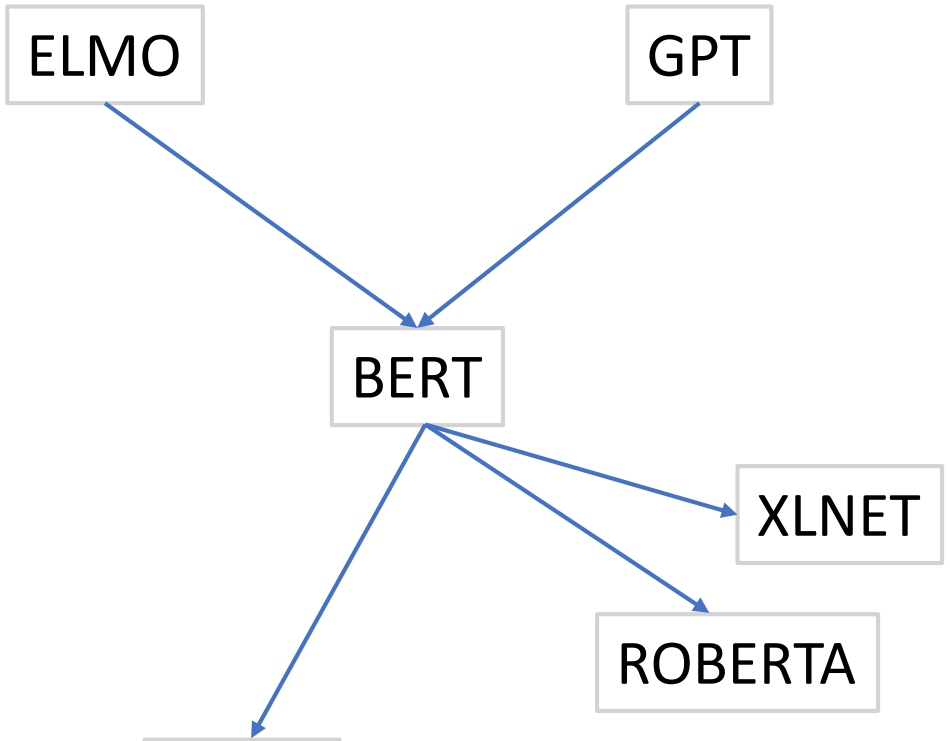
BERT-like models



BERT-like models

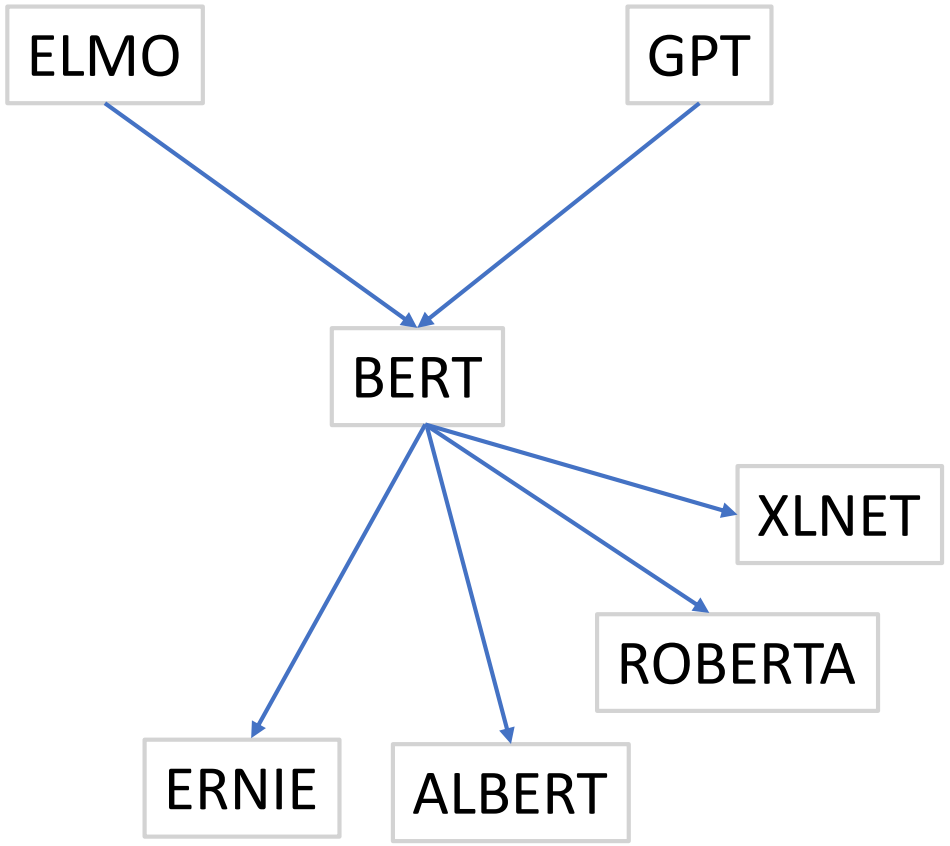


BERT-like models



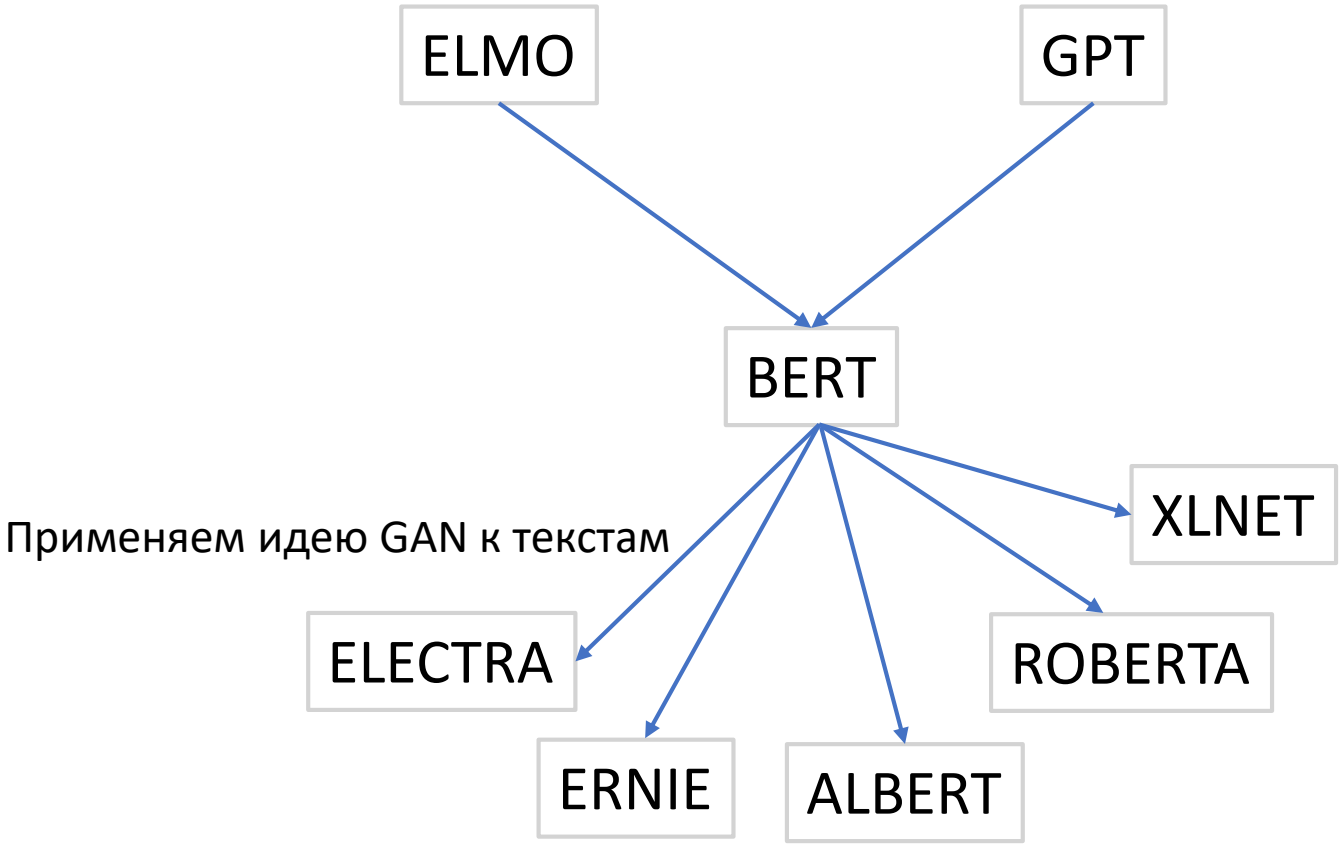
Маскируем именованные сущности

BERT-like models



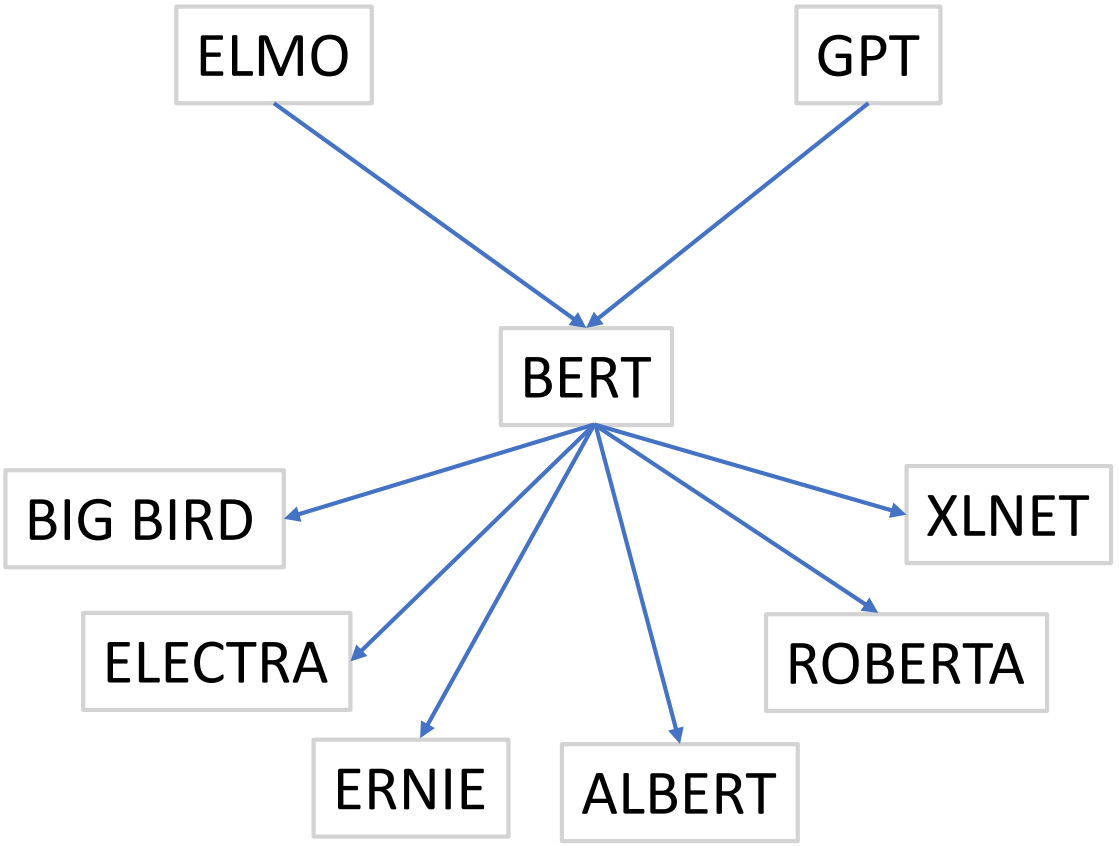
Требуем совпадения весов на всех слоях

BERT-like models



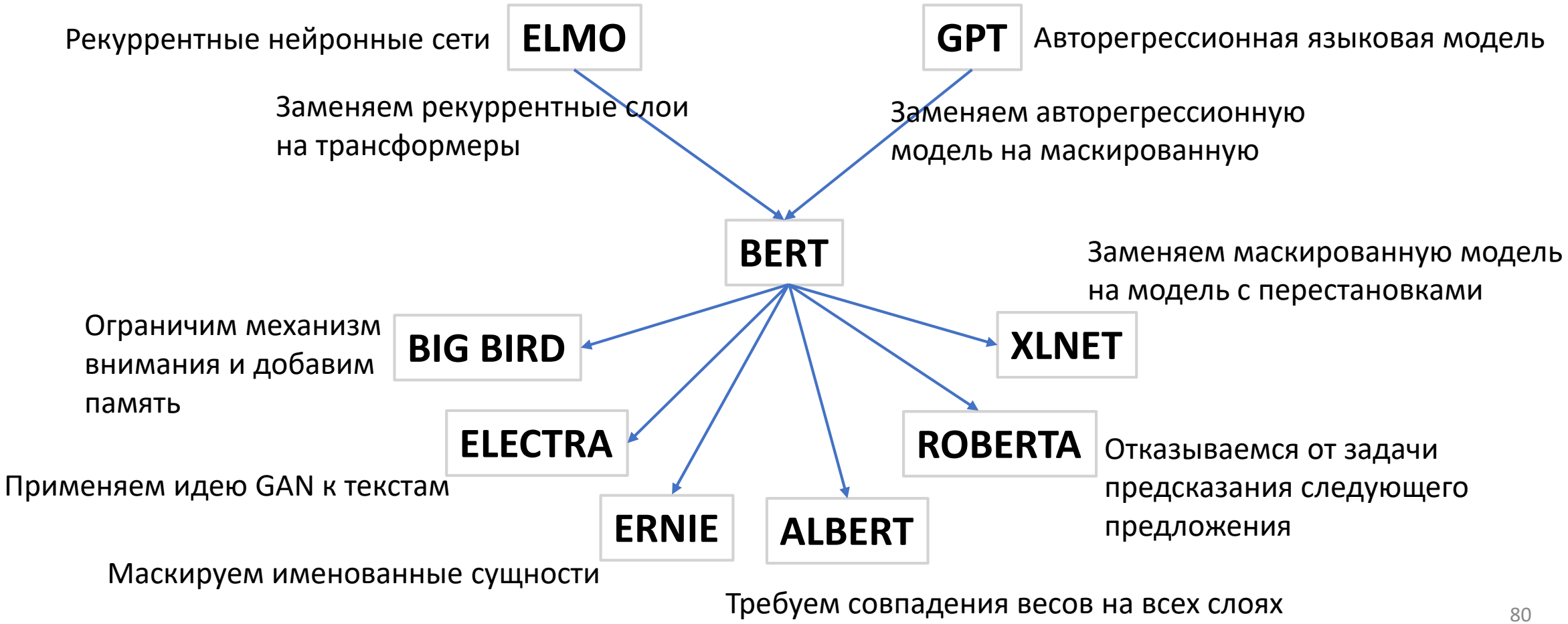
Применяем идею GAN к текстам

BERT-like models



Ограничим механизм внимания и добавим память

BERT-like models



RoBERTa

- Динамическое маскирование
- Отказ от задачи предсказания следующего предложения
- В 10 раз больше данных для обучения
- Необходимо много вычислителей

ERNIE

- Случайным образом убираем именованные сущности и какое-то количество случайных слов или словосочетаний
- Необходимо перед обучением извлечь именованные сущности
- Работает лучше BERTa на задачах извлечения информации

ALBERT

- Уменьшение количества параметров
- На каждом слое одинаковые веса
- Засчёт того, что веса одинаковые, можно сделать больше слоёв и углубить модель
- В самой большой конфигурации 233 млн параметров

ELECTRA

- Слова заменяются не на токен [MASK], а на другие случайные слова
- Для этого нужен отдельный генератор предложений с заменами
- Используем не очень хорошо обученный BERT для генерации
- Модель учится определять, было ли слово заменено
- Очень быстрое обучение
- Сильно обыгрывает BERT, если обучать на том же количестве параметров и слоёв

BIG BIRD

- Добавляем к токенам BERT-а дополнительный управляющий токен, имеющий доступ ко всей последовательности слов
- Механизм внимания имеет доступ к избранным словам
- Избранные слова выбираются на основе весов из механизма внимания
- Можно подать большой текст на вход и не потратить много памяти на внимание

Вопросы

- Для обучения моделей требуется много ресурсов
- Это не круто для экологии
- Transformer-as-a-service
- Методология оценивания языковых моделей

Сжатие моделей

- Квантизация
- Удаление весов (прунинг)
- Дистилляция

Квантизация

- Смена формата весов с float32 на int8
- Сокращает затраты памяти в 4 раза
- Качество падает в пределах 1% (Q8bert)
- Если применять на специальных аппаратных средствах, ускорение в 3.7 раза

Удаление весов (прунинг)

- Удалять можно слои, а также модули внимания
- Удаляем слои и облегчаем и ускоряем модель, потеряв около 2% качества
- Почти никакие из модулей внимания не работают хорошо
- Удаляем те, которые плохо работают и получаем ускорение модели на 17%
- Надо знать, какие из них плохие

Дистилляция

- Обучаем маленькую модель воспроизводить результаты большой предобученной модели
- distilBERT – параметров в 2 раза меньше, качество на 5% ниже
- Можно использовать для разработки инфраструктуры и отладки, потому что входные и выходные форматы такие же, как у BERT

Сравнение методов сжатия моделей

- Все уменьшают размер модели
- Все ускоряют работу модели
- Только с использованием дистилляции можно сохранить свойство переноса обучения

Оценка качества моделей

- GLUE: https://gluebenchmark.com/leaderboard

10 задач, связанных с общим пониманием структуры языка

Есть значение результатов от обычных людей

Слишком простой для моделей

- SuperGLUE: https://super.gluebenchmark.com/leaderboard

8 усложнённых задач для понимания структуры языка

Есть значение результатов от обычных людей

Вопросы к бенчмаркам

- Задачи в бенчмарках не применимы на практике
- Чтобы модель хорошо выступила, необходимо потратить колоссальное число ресурсов

Вопросы?

Телеграм @flerchy

Вк vk.com/flerchy/