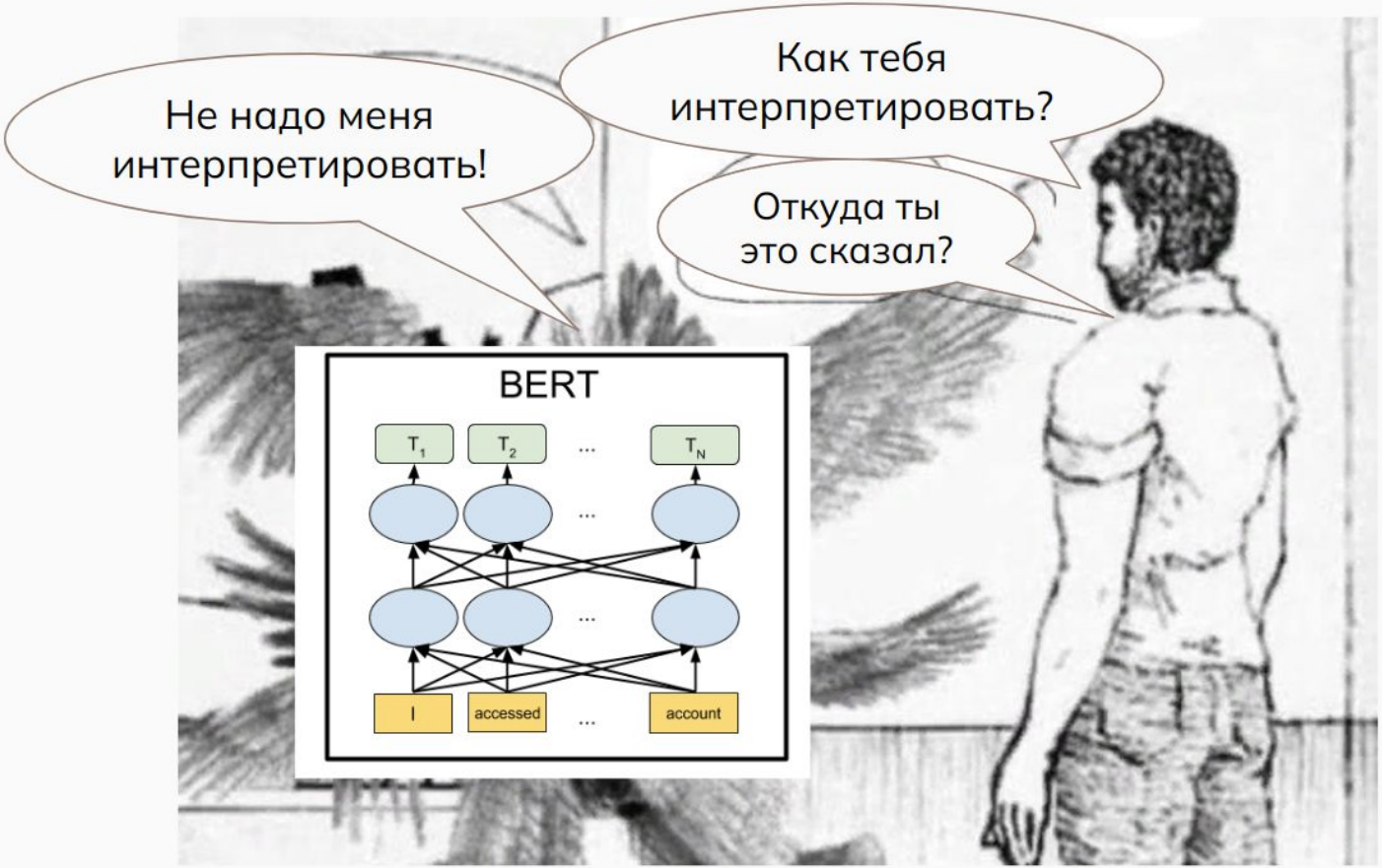
Пробинг: локализация грамматики в нейронных сетях

Корнилов Альберт, Степанова Ангелина, Сухарева Мария, Шумакова Лада


Не надо меня интерпретировать!
Как тебя интерпретировать?
Откуда ты это сказал?
BERT
T1
T2
...
TN
I
accessed
account

Проблематика

- большие языковые модели используются в различных отраслях, но имеют низкую объяснимость [Bodria et al. 2023; Rudin 2019]
- важно понимать, как модели принимают решения для обеспечения доверия [Belinkov, Glass 2019; Doshi-Velez, Kim 2017; Lipton 2016]
- также обучение LLM требует большого количества ресурсов, что затрудняет добавление новой информации в модель [Zhang et al. 2024]

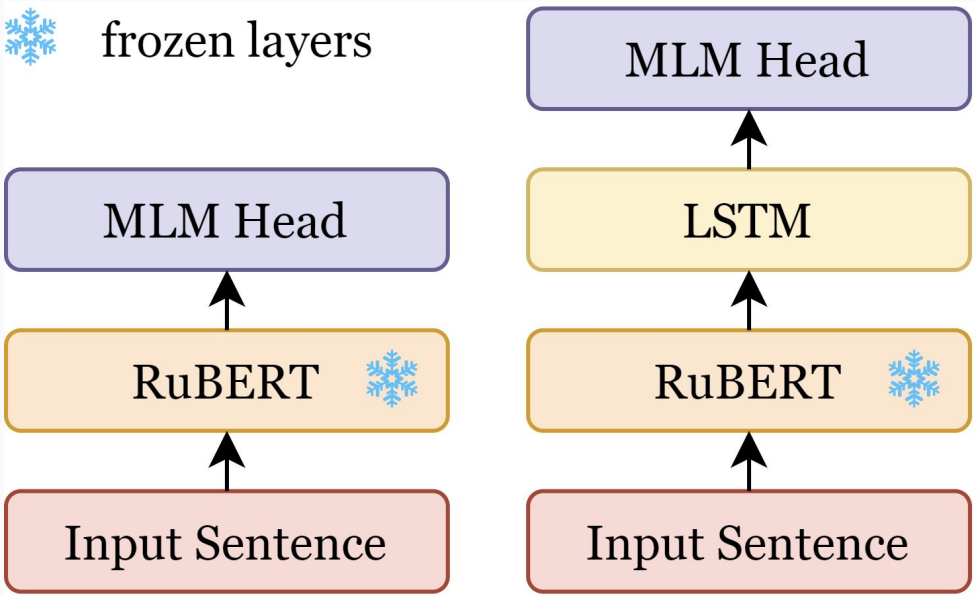
Исследование предыдущей команды [Kudriashov et al. 2024]

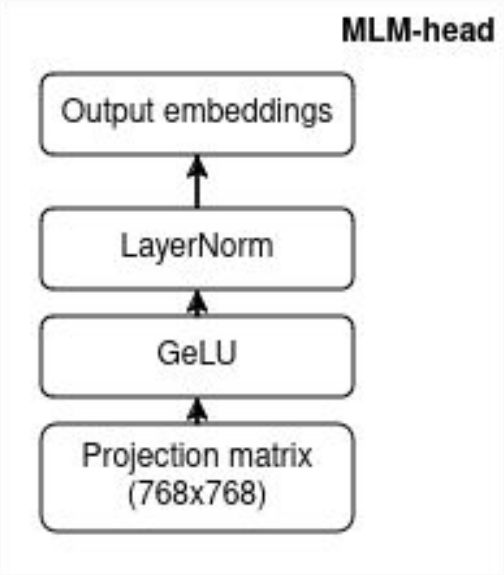
- данные на «поломанном» русском языке: введение полиперсональности

Она делалает кашу

- цель – локализовать полиперсональность в модели
- «замораживание» модели

Метод локализации знаний в модели

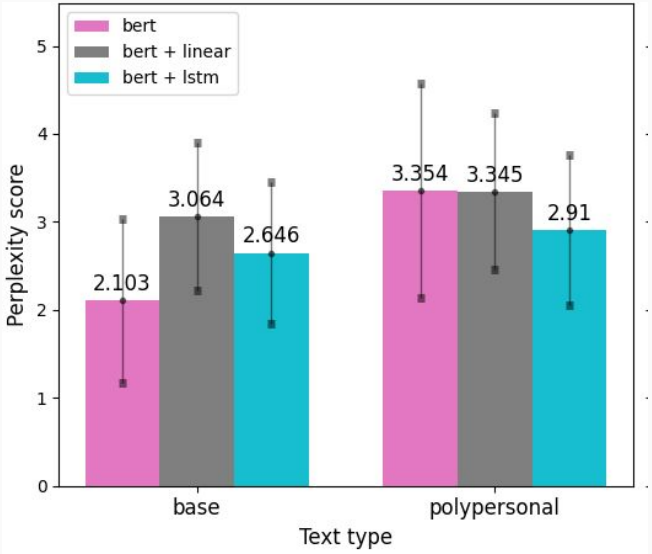


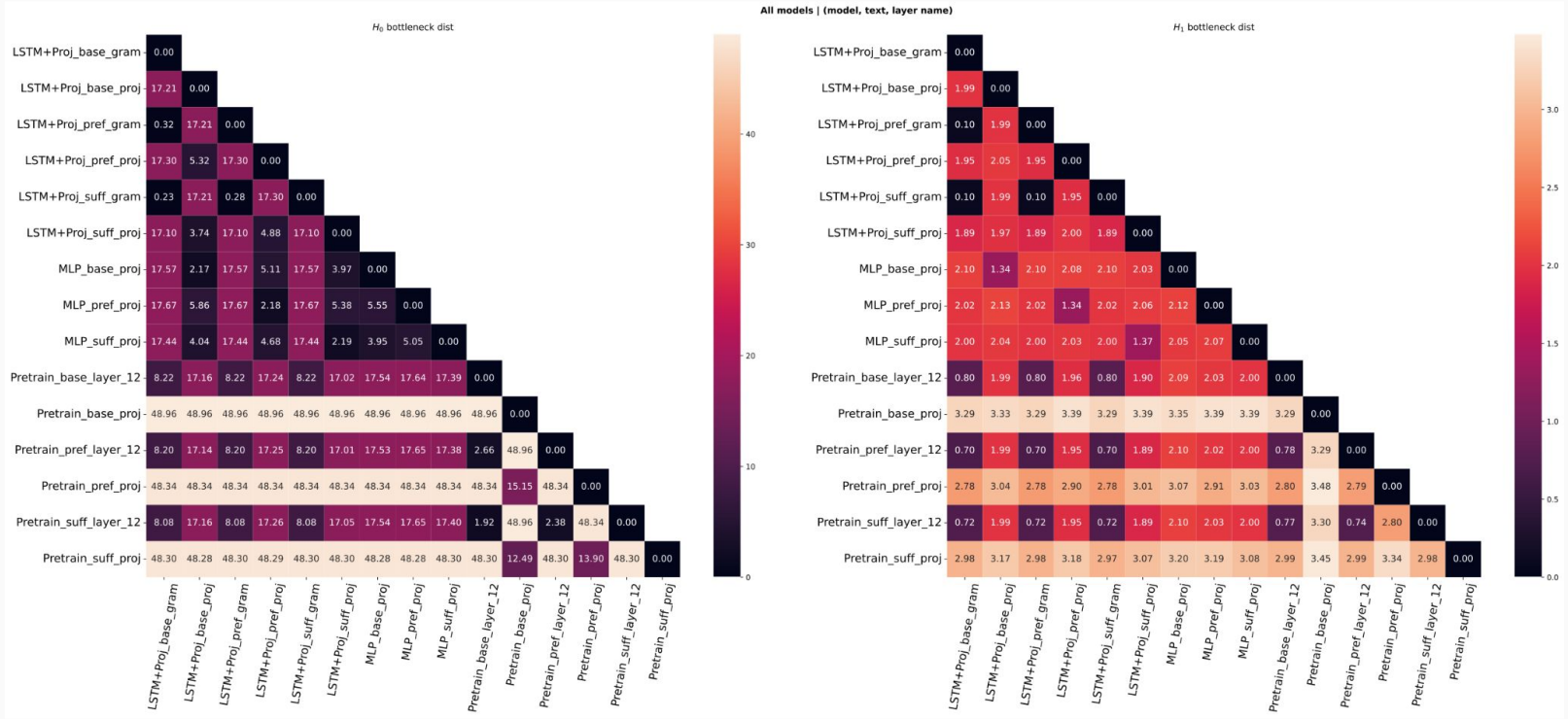


Дообучение RuBERT на стандартных текстах корпуса было проведено Сергеем Кудряшовым

Оценка перплексии

Показатели перплексии обученных без флагов моделях




All models | (model, text, layer name)
H_0 bottleneck dist
H_1 bottleneck dist
LSTM+Proj_base_gram
LSTM+Proj_base_proj
LSTM+Proj_pref_gram
LSTM+Proj_pref_proj
LSTM+Proj_suff_gram
LSTM+Proj_suff_proj
MLP_base_proj
MLP_pref_proj
MLP_suff_proj
Pretrain_base_layer_12
Pretrain_base_proj
Pretrain_pref_layer_12
Pretrain_pref_proj
Pretrain_suff_layer_12
Pretrain_suff_proj

Пробинг на стандартном RuBERT

На выходах слоёв обучили классификаторы, которые должны были по вектору распознать, есть ли показатель полиперсональности в предложении или нет – **диагностический пробинг.**
Пробинг показал, что на выходах моделей, не обученных на полиперсональных текстах, примеры разделяются на три класса с высокой accuracy (0,94), в то время как показатель accuracy на случайных значениях равен 0,5.

Пробинг на стандартном RuBERT

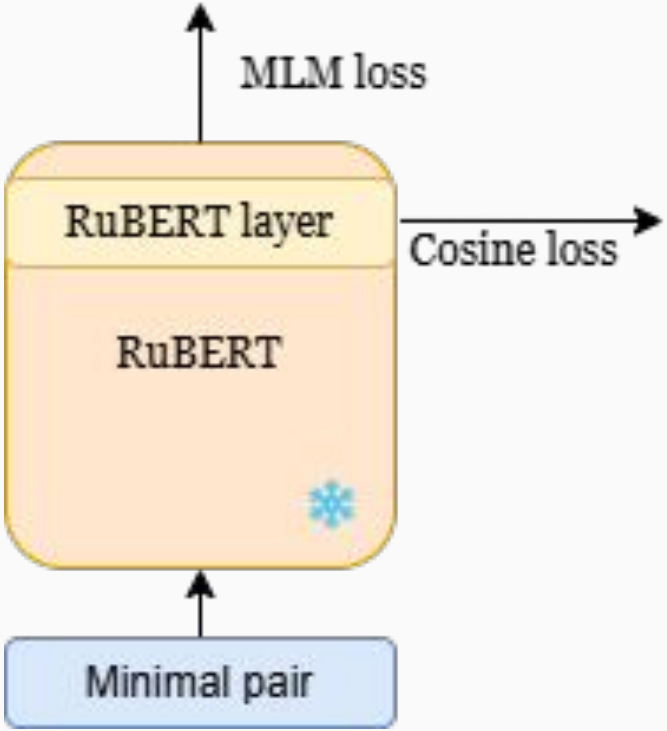
К диагностическому пробингу есть вопросы:

- насколько корректно оценивать знания модели по результатам работы классификатора?
- действительно ли вся заложенная в векторах информация, на которую опираются линейные классификаторы, используется в моделях?

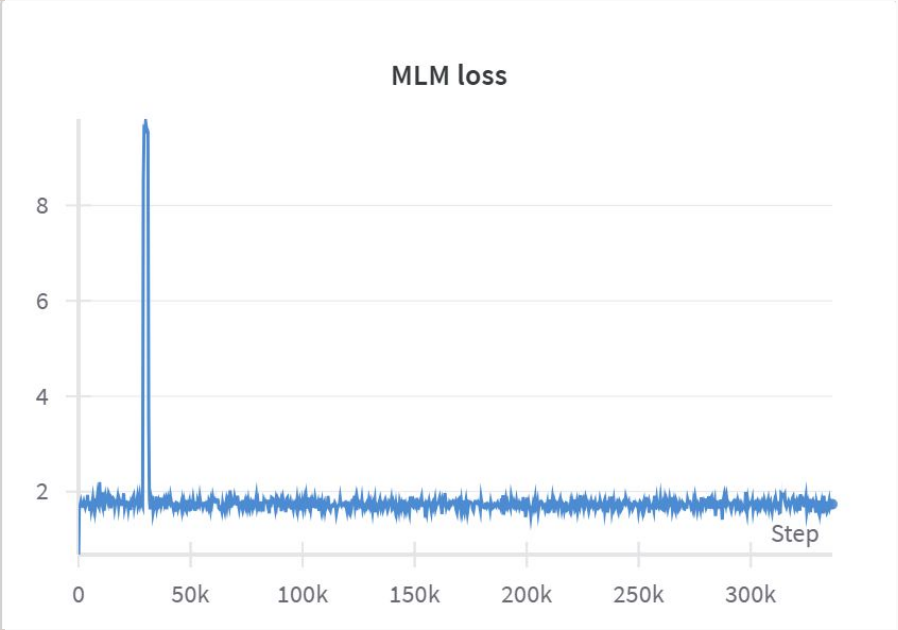
Итог: нужна дальнейшая интерпретация моделей

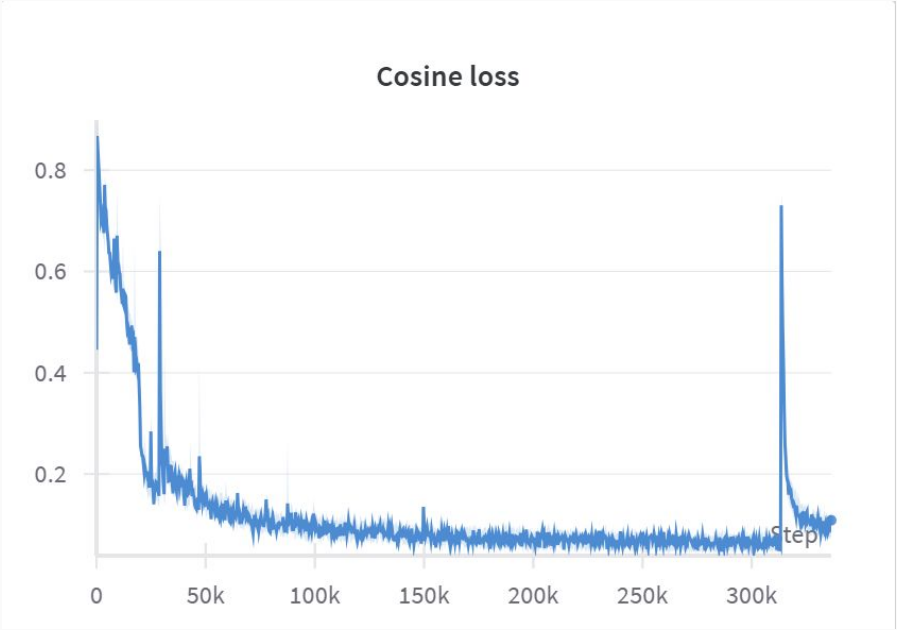
Новая архитектура

- Минимальная пара – предложение в стандартной и полиперсональной версиях
- Обучается только один слой посреди модели, считается cosine loss
- На общем выходе считается MLM loss



Новая архитектура





Что удалось? [Kudriashov et al. 2024]

- подтвердить гипотезу об отличающихся выходах модели
- снизить перплексию на искуственно сгенерированных данных благодаря предложенным методам обучения моделей → успешная локализация информации об одном грамматическом явлении
- обучить модели, которые не стали значительно хуже справляться со стандартным языком → не начали забывать стандартный язык;

А что не удалось?

Параметры оказались линейно разделимыми на другом фрагменте модели → нет четкой уверенности в том, что локализация действительно удалась

Цели текущей работы

Модели предыдущей команды + новые методы локализации, что позволит:

- повысить интерпретируемость языковых моделей
- сделать добавление новых знаний в модель менее ресурсозатратным
- попробовать избежать проблем с линейно разделимыми параметрами
- возможно, поработать с естественным языком без искусственных категорий

Как понять, что цель достигнута?

Модель должна:

- Не быть сильно хуже модели без локализации
- Показывать предсказуемые результаты при fine-grained пробинге:
  - отдел с локализацией должен быть активен при обработке релевантных категорий токенов
  - остальная модель не должна знать о категории (модель без локализованной части должна работать хуже на примерах, содержащих категорию)

Возможные методы

1. A Bayesian Framework for Information-Theoretic Probing [Pimentel, Cotterell 2021]

- пробинг как аппроксимация Mutual Information
- Bayesian Mutual Information
- идея: работать не с косинусом и линейно-алгебраическими преобразованиями, а с теоретико-информационными функционалами в латентном пространстве, задавая их для различающихся токенов

Возможные методы

Взаимная информация (Mutual Information):
$I(X, Y) = H(X) - H(X|Y)$.
(Claude E. Shannon, "A Mathematical Theory of Communication", 1948)

Свойства теории информации, которые не соотносятся с реальностью ML

1) новые данные не добавляют информацию: $H(X \mid D_N = d_N) = H(X)$
   (т. к. предположение, что они из того же распределения)
2) условное распределение уменьшает энтропию: $H(X \mid Y) \leq H(X)$
3) процессинг данных не увеличивает информативность:
   $I(X; f(Y)) \leq I(X; Y)$

Возможные методы

Решение: байесовская взаимная информация

~~$I(X, Y) = H(X) - H(X|Y)$~~

$I_\theta(Y \to X \mid d_N) = H_\theta(X \mid d_N) - H_\theta(X \mid Y, d_N)$

главное отличие от исходной взаимной информации:

реальное распределение ≠ распределение модели

Возможные методы

Решение: байесовская взаимная информация

~~$I(X, Y) = H(X) - H(X|Y)$~~

$I_\theta(Y \to X \mid d_N) = H_\theta(X \mid d_N) - H_\theta(X \mid Y, d_N)$

1) новые данные добавляют информацию: ~~$H(X \mid D_N = d_N) = H(X)$~~
2) условное распределение уменьшает энтропию: ~~$H(X \mid Y) \leq H(X)$~~
3) процессинг данных может увеличивать информативность:

   ~~$I(X; f(Y)) \leq I(X; Y)$~~

   $I_\theta(f(Y) \to X \mid d_N) \; ? \; I_\theta(Y \to X \mid d_N)$

“data can add information, processing can help, and information can hurt”

Возможные методы

2. Sliced Mutual Information: A Scalable Measure of Statistical Dependence [Goldfeld, Greenewald 2021]

- еще один вариант аппроксимации
- Sliced Mutual Information

Возможные методы

3. Sparse Autoencoders Find Highly Interpretable Features In Language Models [Cunningham et al. 2023]

- чтобы сделать модели более интерпретируемыми, можно использовать разреженные автоэнкодеры
- для них не требуются размеченные данные (=> можно использовать естественный язык)

Ожидаемые результаты

Ожидаемые результаты

Место для ваших вопросов

Спасибо за внимание!