Интерпретируемость NLP-моделей

Дмитрий Колодезев, Промсофт

NLP-митап в Омске

15.10.2022

Как бы план

- NLP и его друзья
- Что и зачем интерпретируем
- Трансформеры
- Fine-tuning
- Few-shot learning
- Чеклист

Однажды на Датафесте

Вопрос из зала: Трансформеры NLP, трансформеры неинтересны для визуализации?

Дмитрий Колодезев: Они интересны для визуализации, просто я не понимаю, какую я пользу могу из визуализации извлечь. У меня есть большая модель. Я смотрю в них и вижу, что как-то она не так визуализирует мои слова, мне не нравится. Что я с этим сделаю?

Если у меня есть маленькая табличная модель, я пошел ее доучил, посмотрел, сравнил.

А языковые модели столь сложны, что по большому счету их уже не столько учат, а в лучшем случае чуть-чуть доучивают. Или вообще подбирают подводки. То есть с языковыми моделями проблема, если вы не живете языковыми моделями, то, наверное, смотреть их лишний раз не надо, чтобы не расстраиваться.

Я не понимаю, какие выводы можно сделать посмотрев Bert, поэтому и не пользуюсь визуализацией, а так-то трансформеры интересны.

Про attention есть история, что attention – это та же самая saliency map в нейронках, то есть он подсвечивает не то, почему модель принимает решение, а то, на что она обращает внимание. Это могут быть просто необычные для набора данных вещи. То есть внезапно что-то странное появилось в наборе данных, модель обратит на это внимание. Необязательно, что она будет учитывать это для принятия решения.

https://kolodezev.ru/interpretable2022.html

Natural Language Processing

- Классификация текста
- Генерация текста
- Разбор текста на кусочки (NER)
- Заполнение пропусков в тексте
- Преобразование текста в текст
- Суммаризация
- …

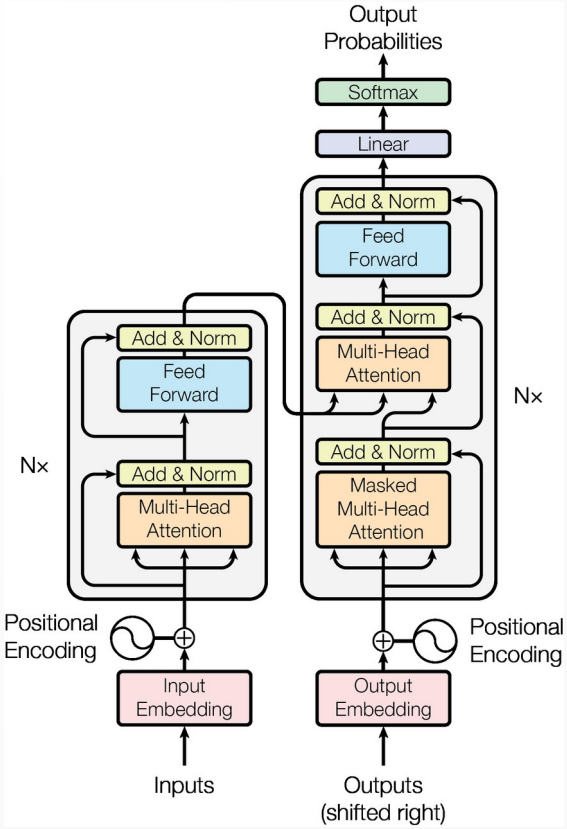
XAI: eXplainable AI

- Имеем право знать
- Отладка модели и данных
- Приемочное тестирование
- Анализ ошибок и разбор полетов
- Выявление уязвимостей
- Проверка очевидных зависимостей
- Социализация модели

https://en.wikipedia.org/wiki/Explainable_artificial_intelligence

В чем, собственно, проблема?

- Преобразования простые
- Но их очень много
- Очень-очень-очень-очень много
  - YaLM — 10^{11} параметров
- Не собираются в голове
- Как будто черный ящик какой-то



Или, говоря другими словами

- «Despite constant advances and seemingly super-human performance on constrained domains, state-of-the-art models for NLP are imperfect. These imperfections, coupled with today's advances being driven by (seemingly black-box) neural models, leave researchers and practitioners scratching their heads asking, why did my model make this prediction?»

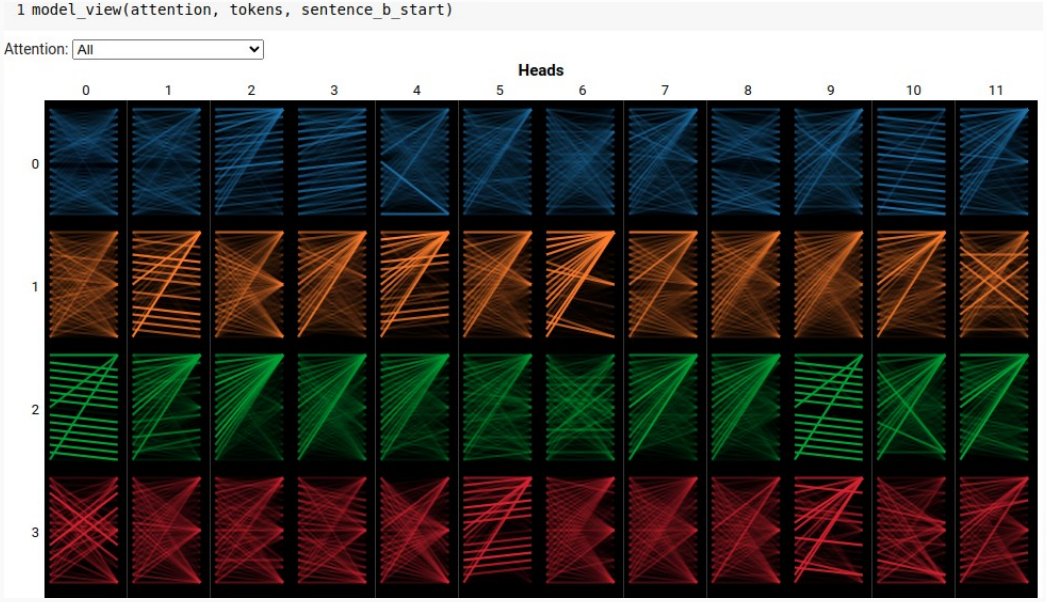
https://demo.allennlp.org/

Как объясняют люди

- Вот по этому слову я сразу узнал в Вас интеллигента
- Он научился этой фразе у одноклассников
- Только не говорите ему про деньги, его это бесит
- Логичным продолжением сказанного будет …
- Это все равно что сказать «я не буду вам помогать»

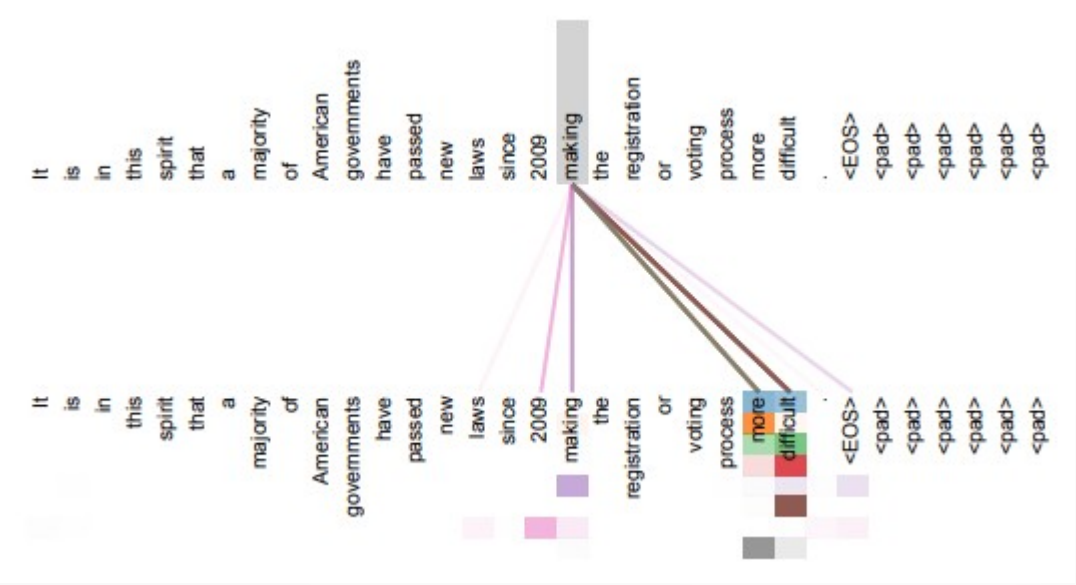
Начнем с ~~трансформаторов~~ трансформеров

- Внимание: главное в трансформерах — внимание!



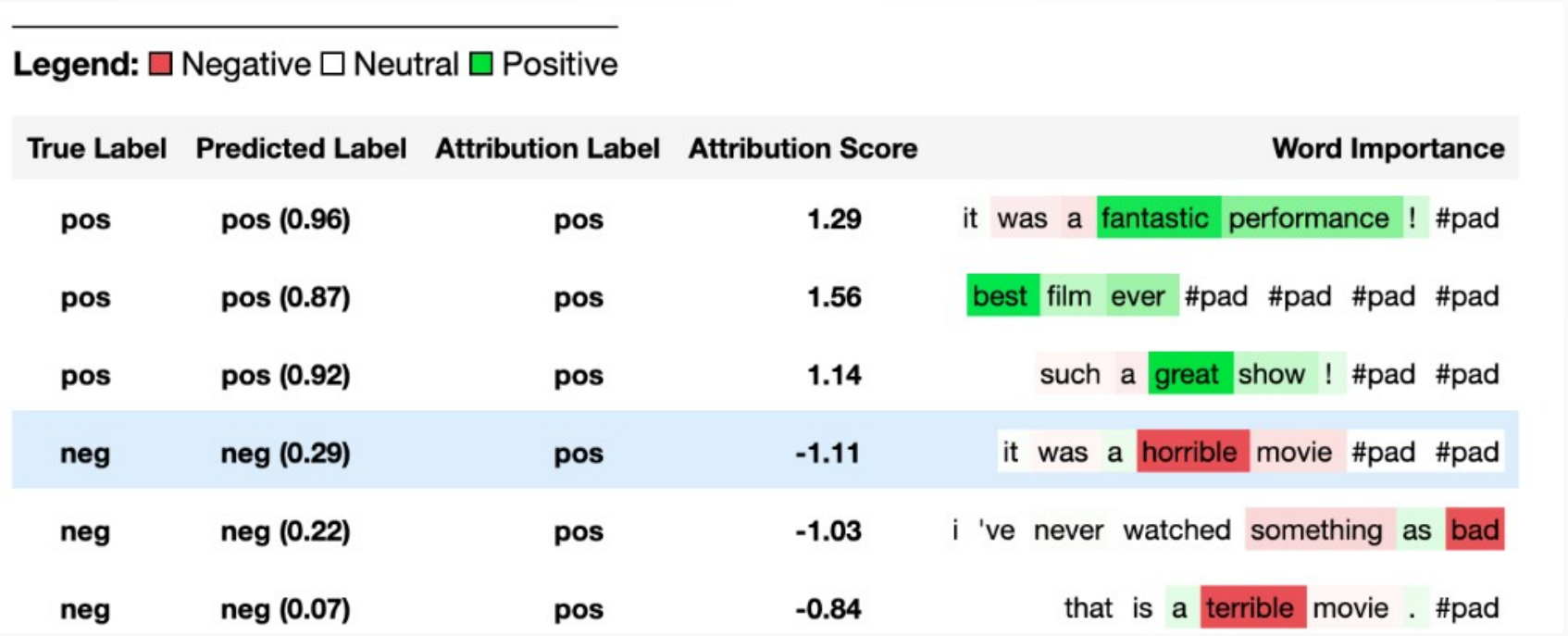
https://github.com/jessevig/bertviz

Attention Is All You've Got



Attention Is All You Need

Может, градиент нарисуем?

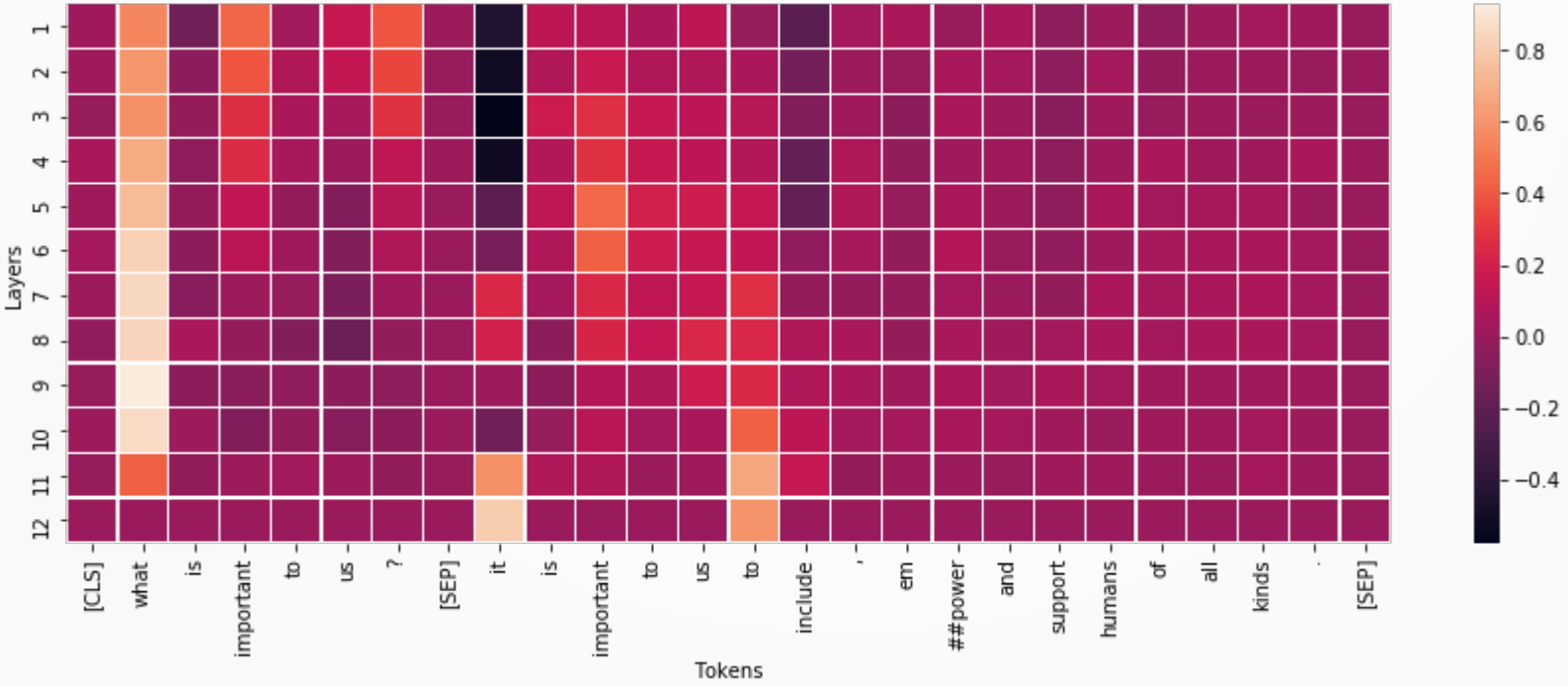


Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
pos	pos (0.96)	pos	1.29	it was a fantastic performance ! #pad
pos	pos (0.87)	pos	1.56	best film ever #pad #pad #pad #pad
pos	pos (0.92)	pos	1.14	such a great show ! #pad #pad
neg	neg (0.29)	pos	-1.11	it was a horrible movie #pad #pad
neg	neg (0.22)	pos	-1.03	i 've never watched something as bad
neg	neg (0.07)	pos	-0.84	that is a terrible movie . #pad

https://captum.ai/tutorials/IMDB_TorchText_Interpret

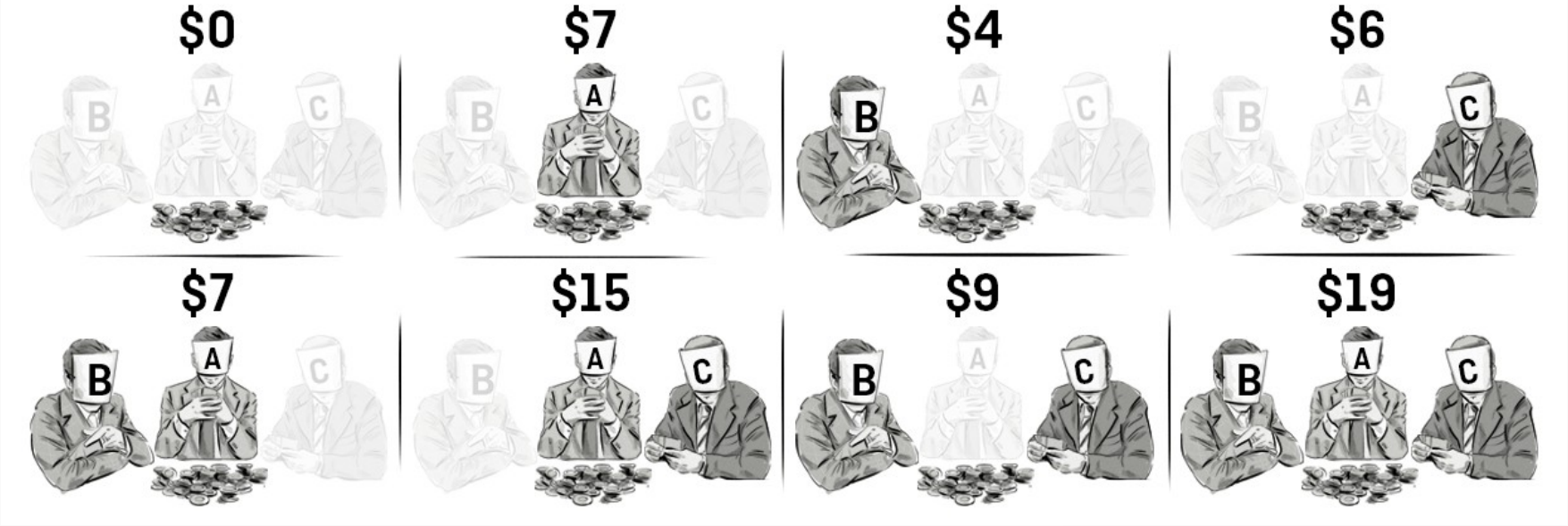
Норма вектора имеет значение



Attention is Not Only a Weight: Analyzing Transformers with Vector Norms

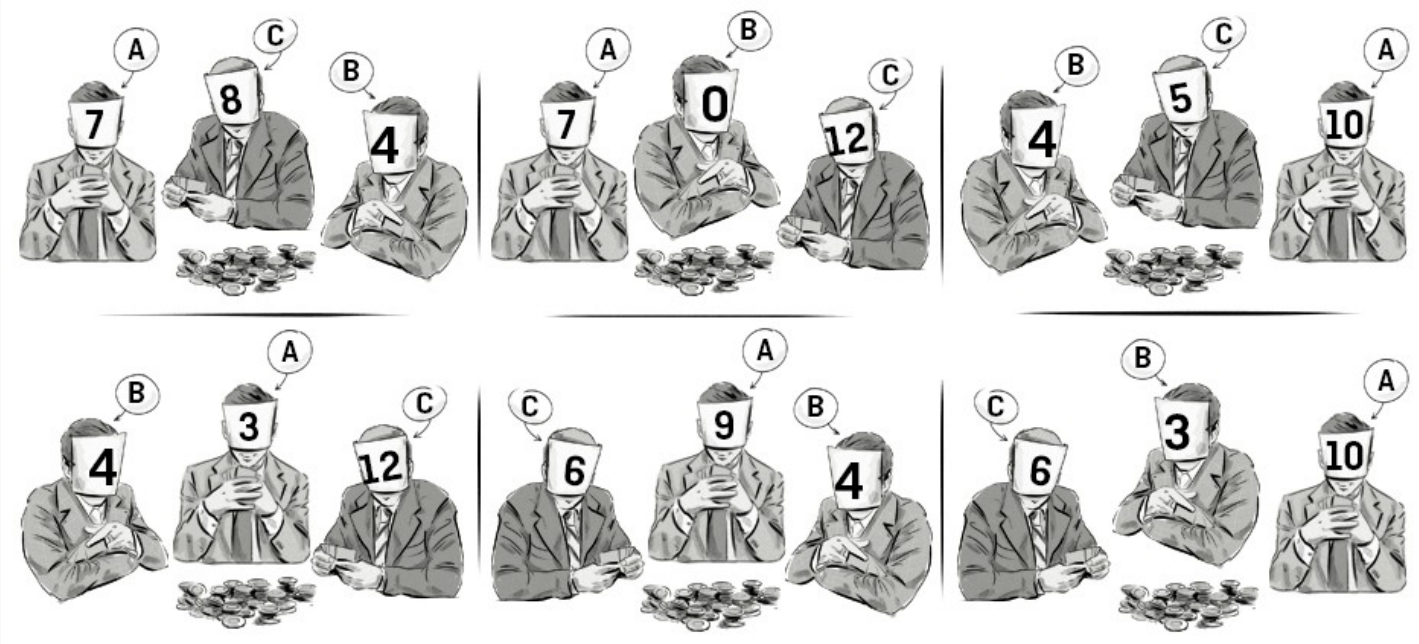
https://captum.ai/tutorials/Bert_SQUAD_Interpret2

Shapley Values



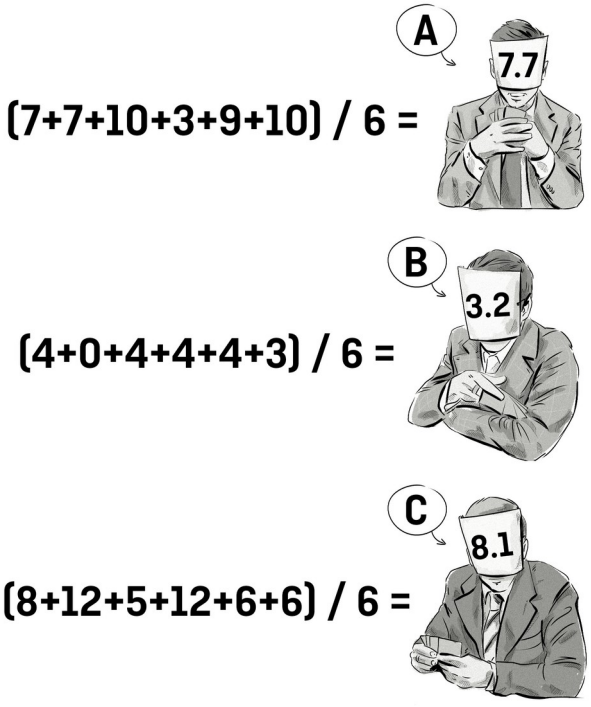
https://clearcode.cc/blog/game-theory-attribution/

Shapley Values



https://clearcode.cc/blog/game-theory-attribution/

Shapley Values vs SHAP



A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

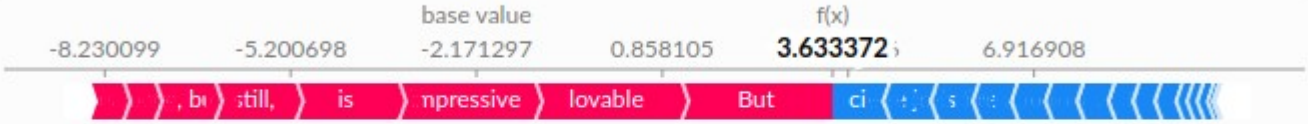
Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

https://clearcode.cc/blog/game-theory-attribution/ https://arxiv.org/pdf/1705.07874.pdf

Нормально объясним

```
[10]: # plot the first sentence's explanation
      shap.plots.text(shap_values[3])
```



This is easily the most underrated film inn the Brooks cannon. Sure, its flawed. It does not give a realistic view of homelessness (unlike, say, how Citizen Kane gave a realistic view of lounge singers, or Titanic gave a realistic view of Italians YOU IDIOTS). Many of the jokes fall flat. But still, this film is very lovable in a way many comedies are not, and to pull that off in a story about some of the most traditionally reviled members of society is truly impressive. Its not The Fisher King, but its not crap, either. My only complaint is that Brooks should have cast someone else in the lead (I love Mel as a Director and Writer, not so much as a lead).

https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/text.html#

Понятно покажем

- In this picture , there are four persons :
  my father , my mother ,
  my brother and my sister .
- En este cuadro , hay cuatro personas :
  mi padre , mi madre ,
  mi hermano y mi hermana .

https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/text.html#

```
shap.plots.text(shap_values)
```

0th instance:

Visualization Type: Saliency Plot

Saliency Plot
x-axis: Output Text
y-axis: Input Text

	In this picture	,	there are four persons	: my father ,	my mother ,	my brother and my sister .	
En	7.191	0.141	0.137	0.122	-0.064	-0.055	0.0
este	1.878	-0.031	0.48	0.524	0.106	-0.279	0.0
cuadro	7.078	0.146	0.348	0.358	-0.122	-0.202	0.0
,	0.149	1.93	-0.544	-0.293	-0.073	-0.118	0.0
hay	-0.434	0.101	6.947	-0.51	-0.341	-0.165	0.0
cuatro	-0.624	0.201	7.434	0.106	-0.099	0.081	0.0
personas	0.017	0.128	5.455	0.033	0.009	0.088	0.0
:	0.349	0.37	1.496	2.519	0.02	0.197	0.0
mi	0.504	-0.074	0.358	5.669	2.105	2.861	0.0
padre	-0.253	-0.034	-0.179	6.934	-0.832	-0.823	0.0
,	0.259	0.123	0.293	1.729	1.075	0.42	0.0
mi	-0.102	-0.309	-0.387	-1.664	3.499	2.038	0.0
madre	-0.222	-0.01	-0.87	0.979	5.393	-1.941	0.0
,	0.023	0.115	0.036	0.371	1.394	0.802	0.0
mi	-0.085	-0.037	-0.454	-0.857	-0.624	4.571	0.0
hermano	-0.131	-0.015	0.094	-0.173	0.159	5.536	0.0
y	0.128	0.106	0.057	0.138	0.377	3.904	0.0
mi	-0.067	0.017	-0.121	-0.174	-0.026	2.961	0.0
hermana	-0.03	-0.036	-0.119	0.093	0.229	5.962	0.0
.	0.059	-0.066	-0.011	-0.15	-0.006	2.303	0.0

Проблема post hoc объяснений

- Люди верят красивым картинкам Interpreting Interpretability
- «Эффект Расемон» Leo Breiman
- Еще и патологии эти ваши

SQuAD

Context	In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	did
Confidence	0.78 → 0.91

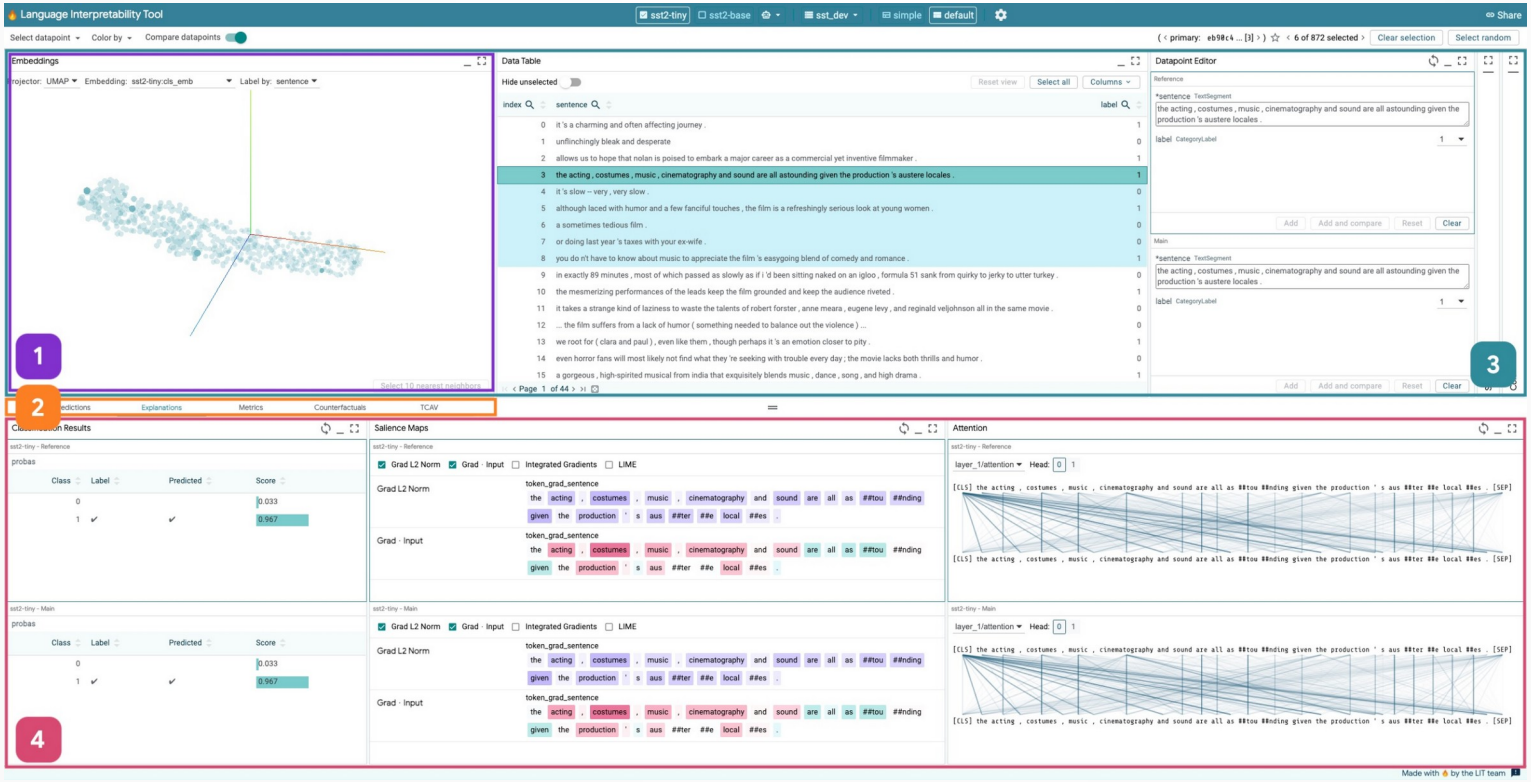
SNLI

Premise	Well dressed man and woman dancing in the street
Original	Two man is dancing on the street
Reduced	dancing
Answer	Contradiction
Confidence	0.977 → 0.706

VQA

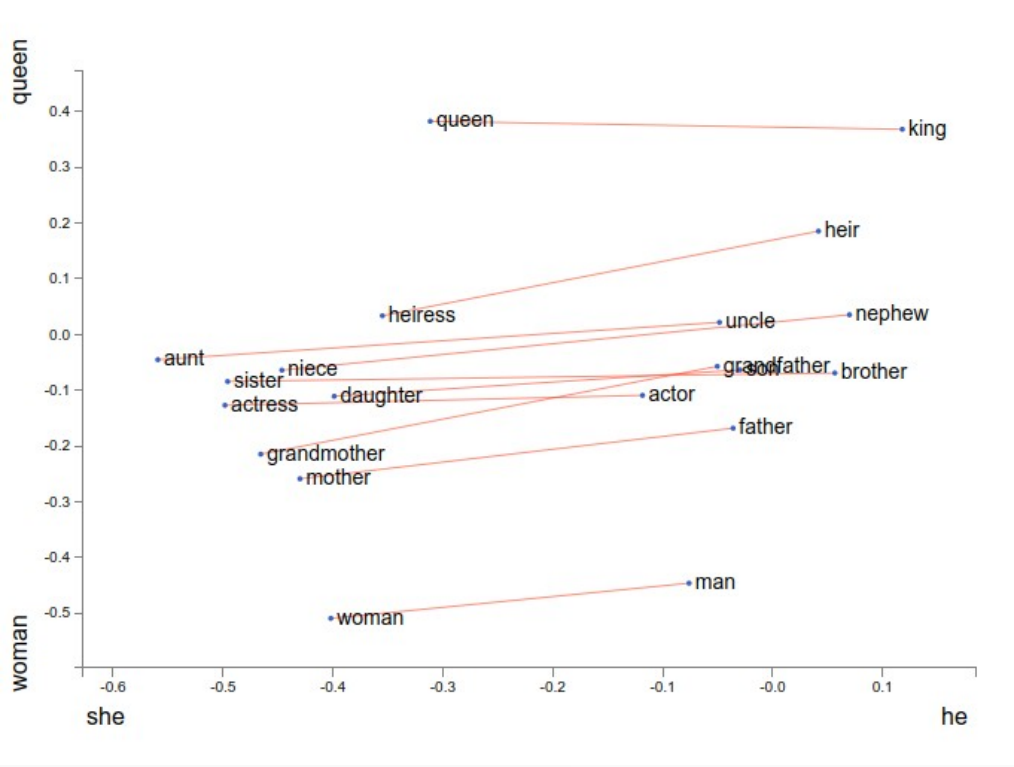
Original	What color is the flower ?
Reduced	flower ?
Answer	yellow
Confidence	0.827 → 0.819

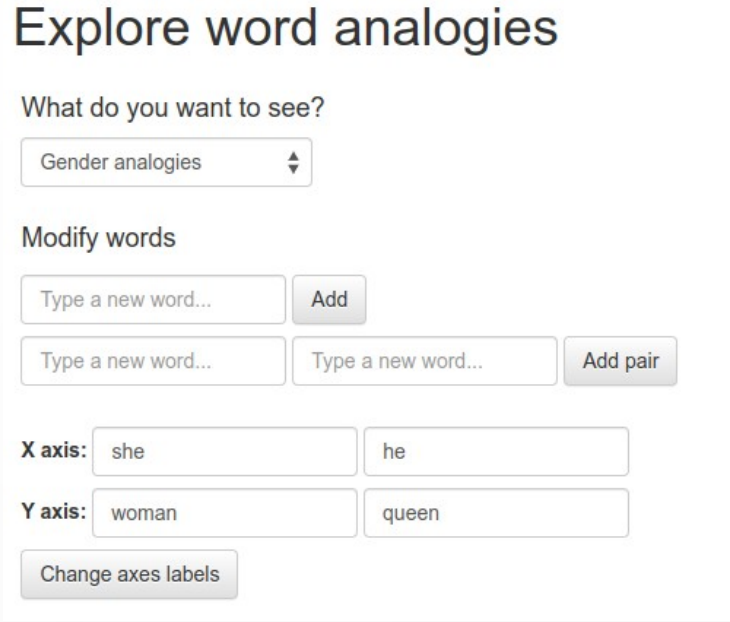
Сесть и во всем разобраться



Language Interpretability Tool https://github.com/PAIR-code/lit

Король + Королева — Арифметика = ?





https://lamyiowce.github.io/word2viz/

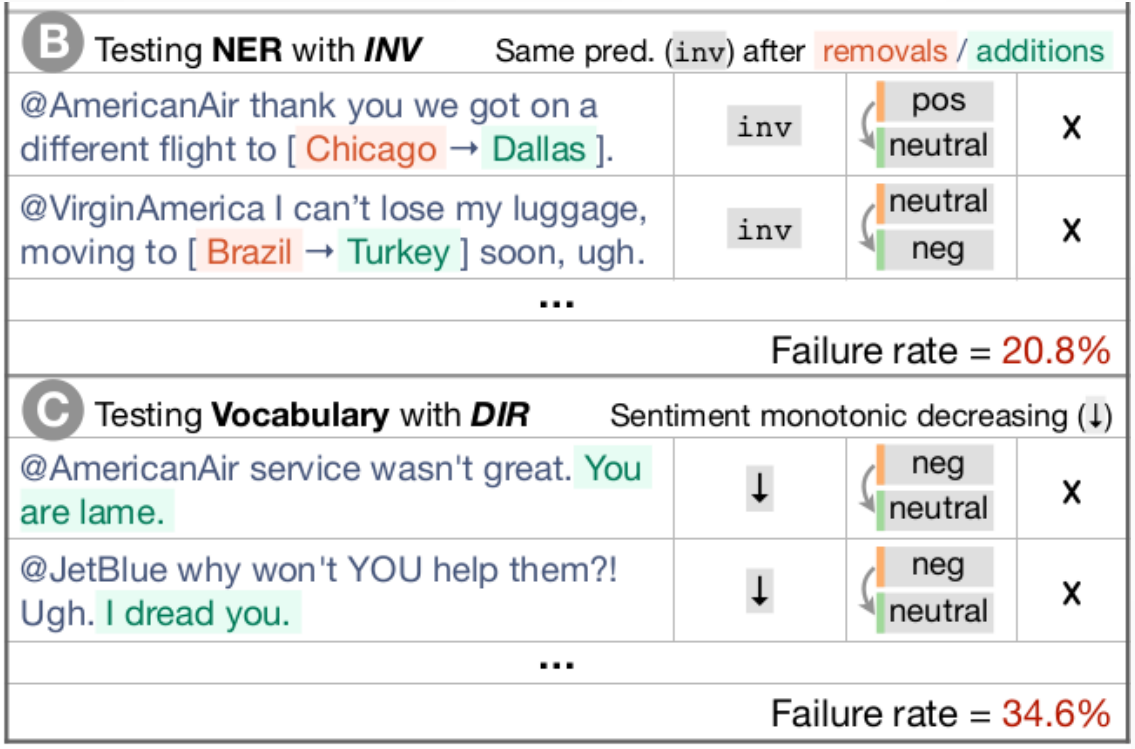
Чеклист

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT***	Labels: negative, positive, neutral		
`Template: I {NEGATION} {POS_VERB} the {THING}.`			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
		Failure rate = 76.4%	

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Чеклист



Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

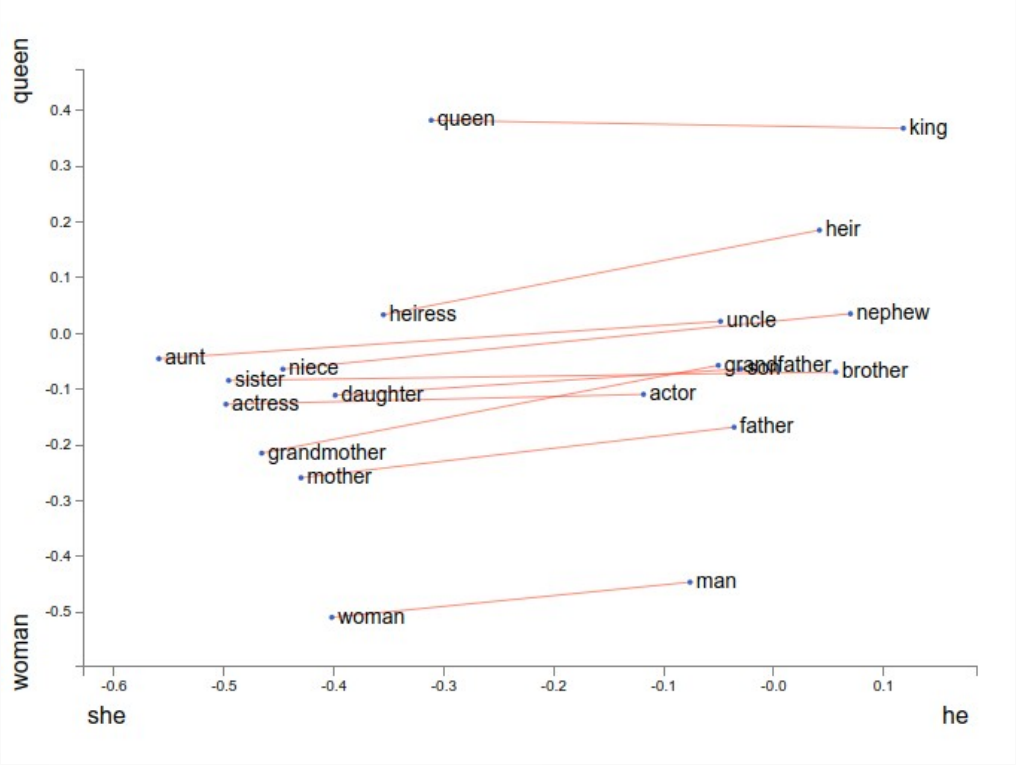
Секретный слайд №1

- Fine-tuning
- Изменение весов
- Интерполяция весов

  Экстраполяция весов

Секретный слайд №2

- Опорные фразы
- Граф



Вопросы

Слайды тут

dkolodezev

dmitry_kolodezev

promsoft

d_key

https://kolodezev.ru/download/omsk_nlp_2022.pdf