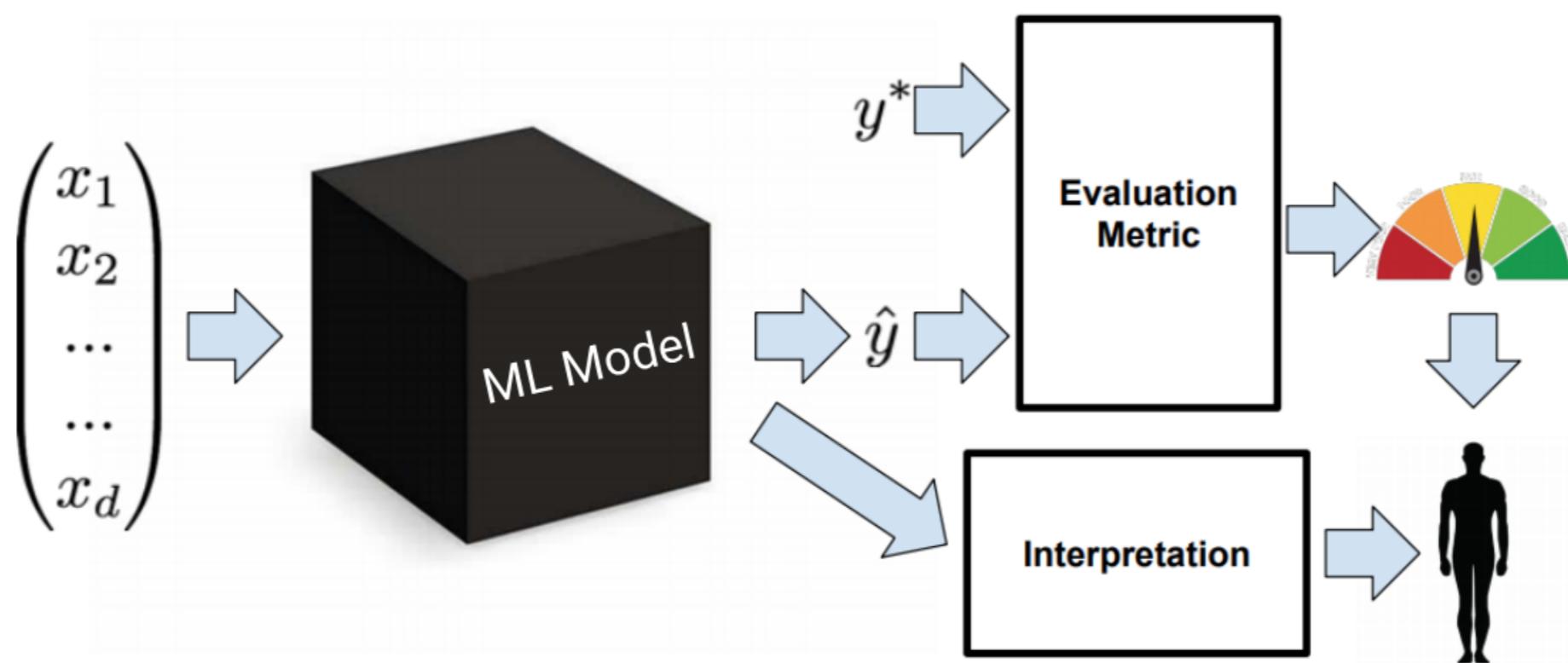


ИНТЕРПРЕТИРУЕМОСТЬ НЕЙРОННЫХ СЕТЕЙ

Камлық Эрик

Интерпретируемость



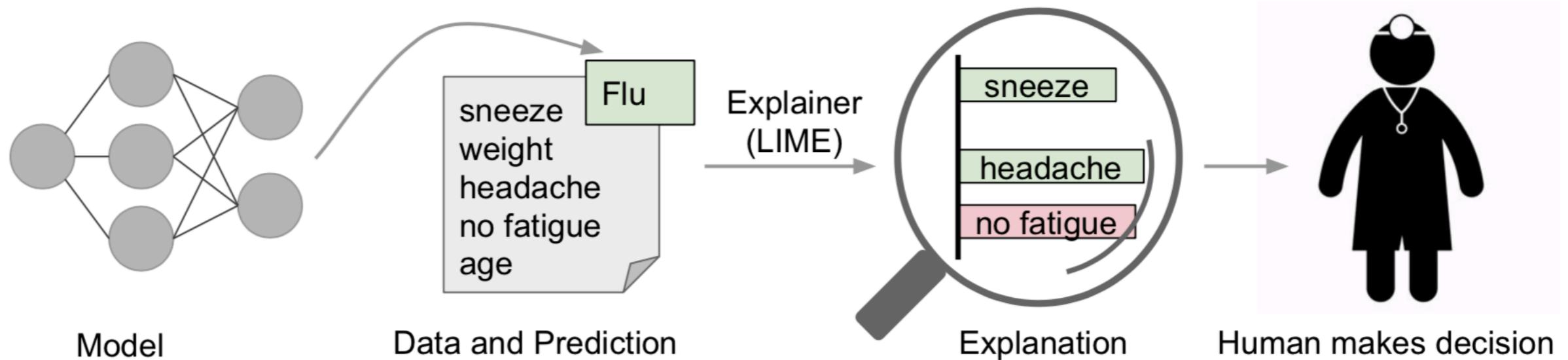
Зачем нужна интерпретируемость

- Юридические ограничения.
- Этичность.
- Безопасность.
- Выбор наилучшей модели.
- Отладка.

LIME

(Local Interpretable Model-Interpretable Explanations)

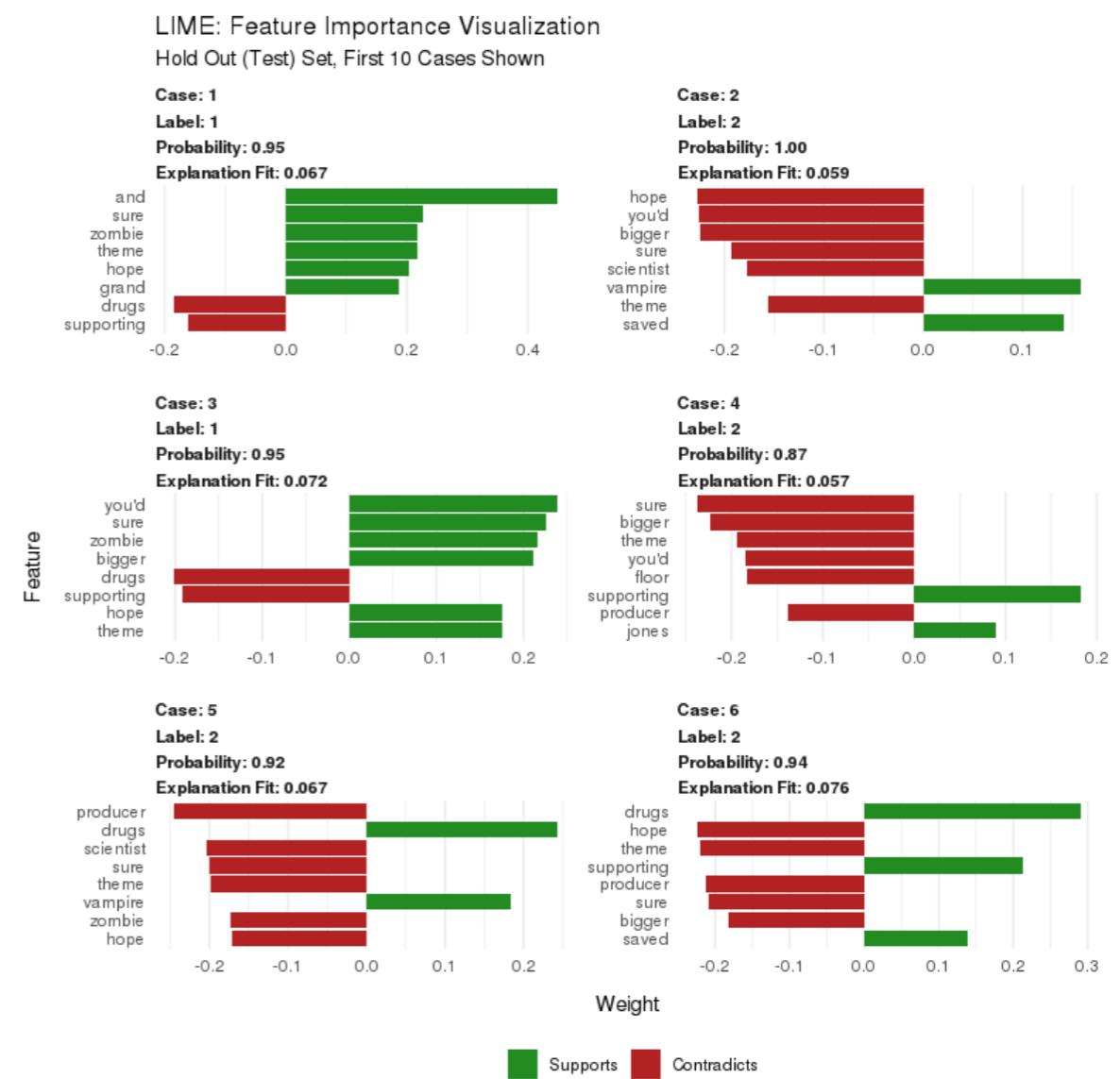
- Легко интерпретируется человеком
- Локально достоверен
- Не зависит от модели
- Даёт глобальное понимание модели



Что получаем на выходе

Для текста - bag of words размера k.

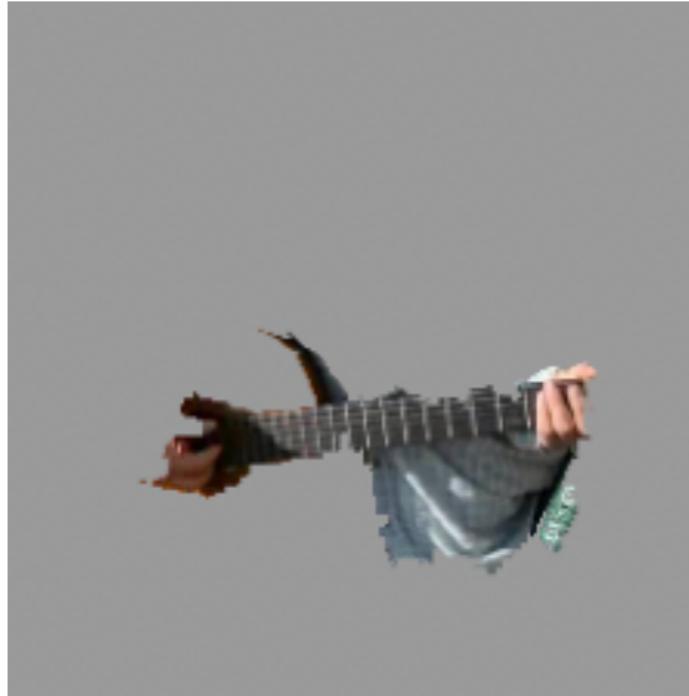
Для изображений -
суперпиксели вместо
слов.



Задача классификации изображений



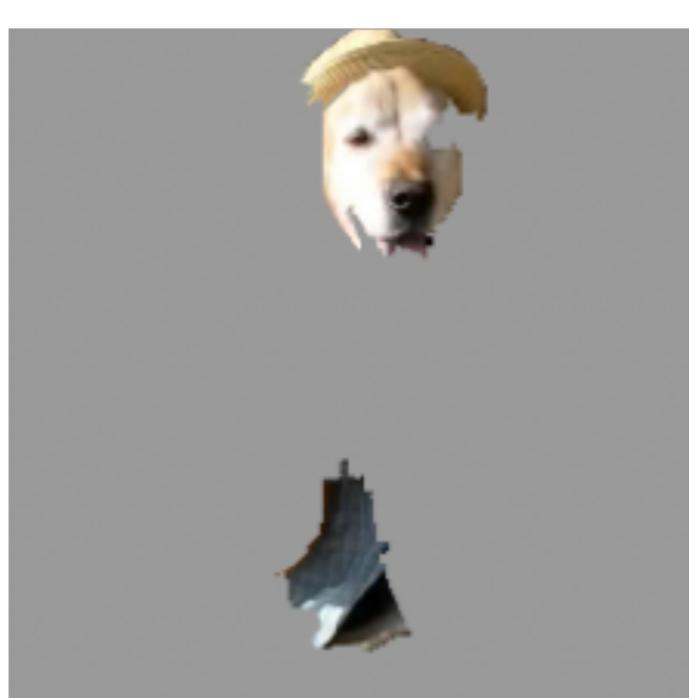
Оригинальное изображение



Электрогитара



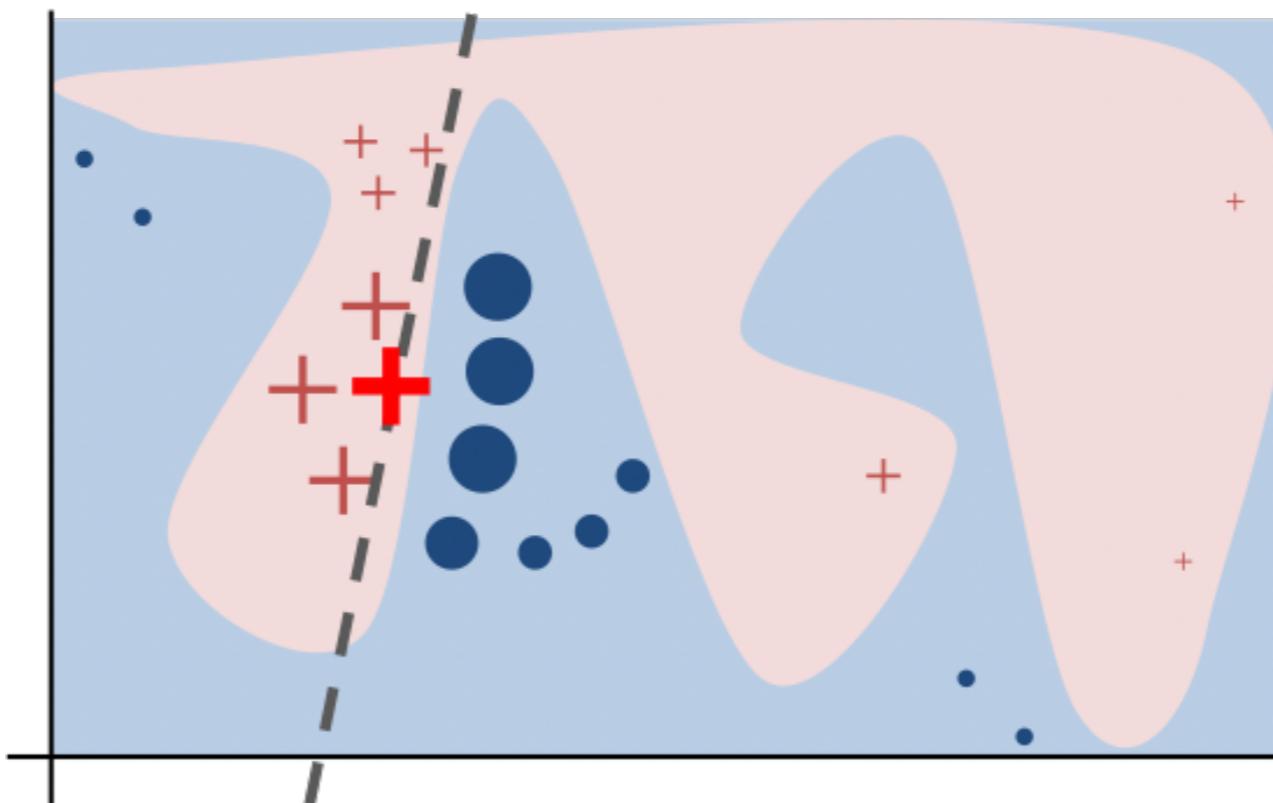
Акустическая гитара



Лабрадор

Алгоритм

1. Выбираем точки z' вокруг текущей x' .
2. Сохраняем $f(z')$ и $\pi(z') = \exp(-D(x, y)^2/\sigma^2)$
3. Алгоритмом K-Lasso выбираем k точек.



Автоматический выбор экземпляров

- Не более В экземпляров
- Репрезентативная выборка объяснений

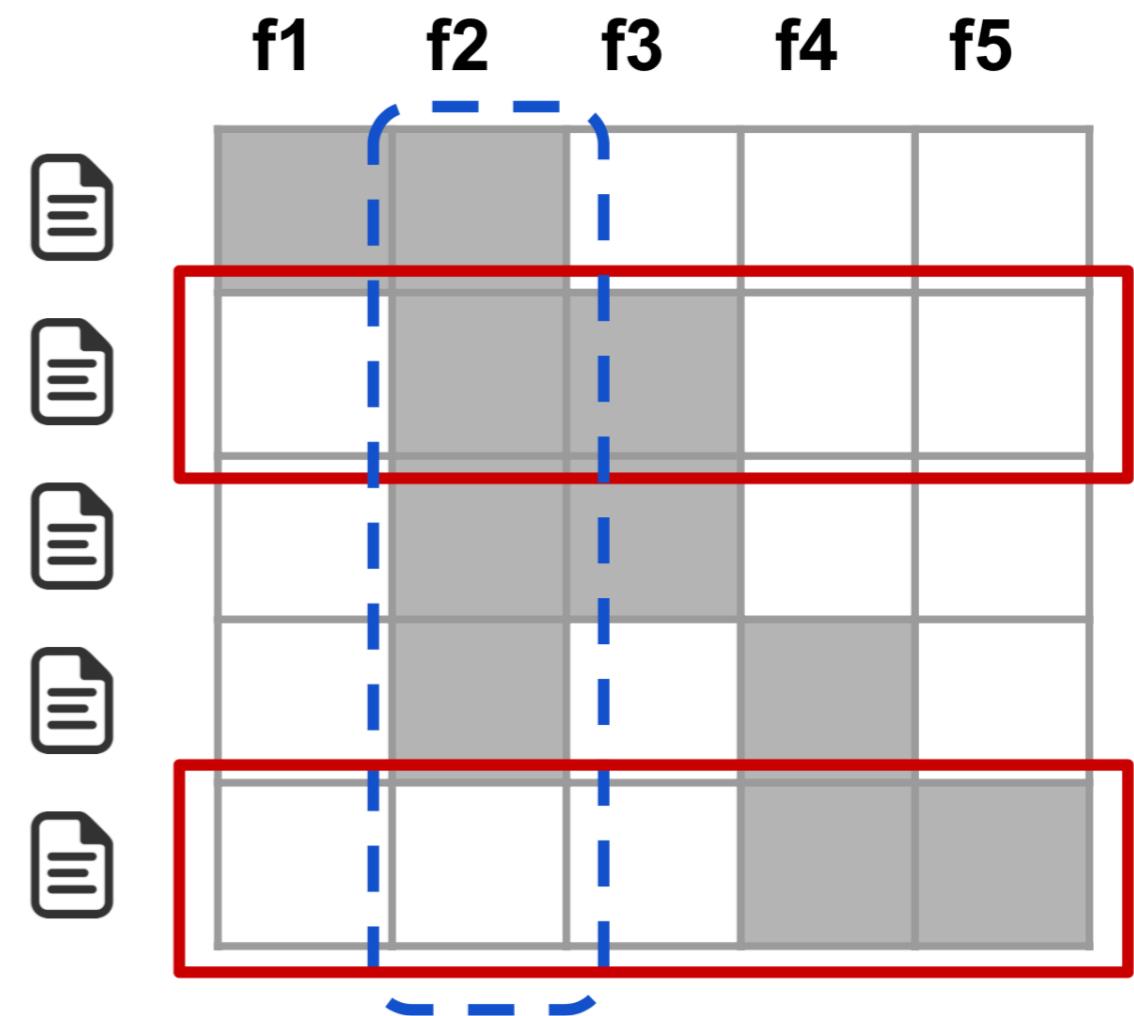
1. Матрица объяснений W

2. Значимость признаков

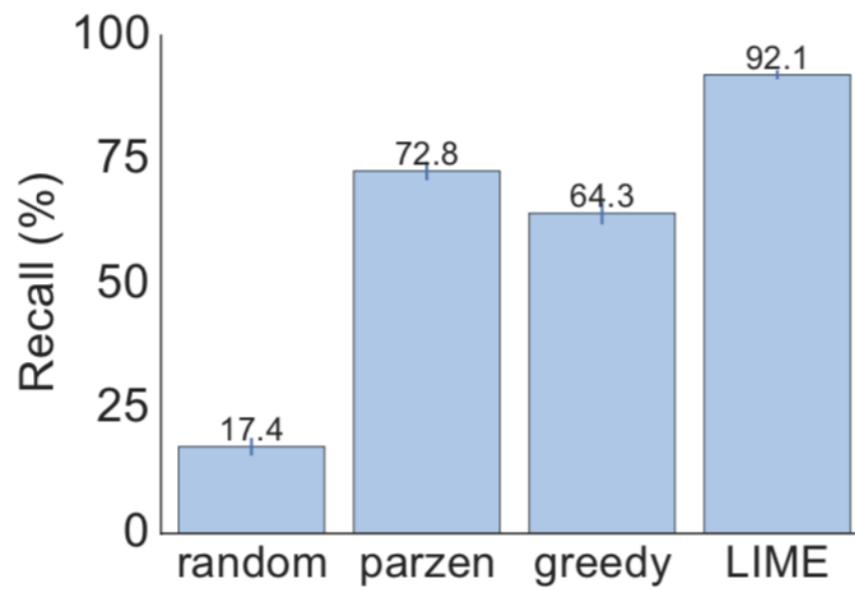
$$I_j = \sqrt{\sum_{i=1}^n |W_{ij}|}$$

3. $V = V \cup argmax_i c(V \cup i, W, I)$

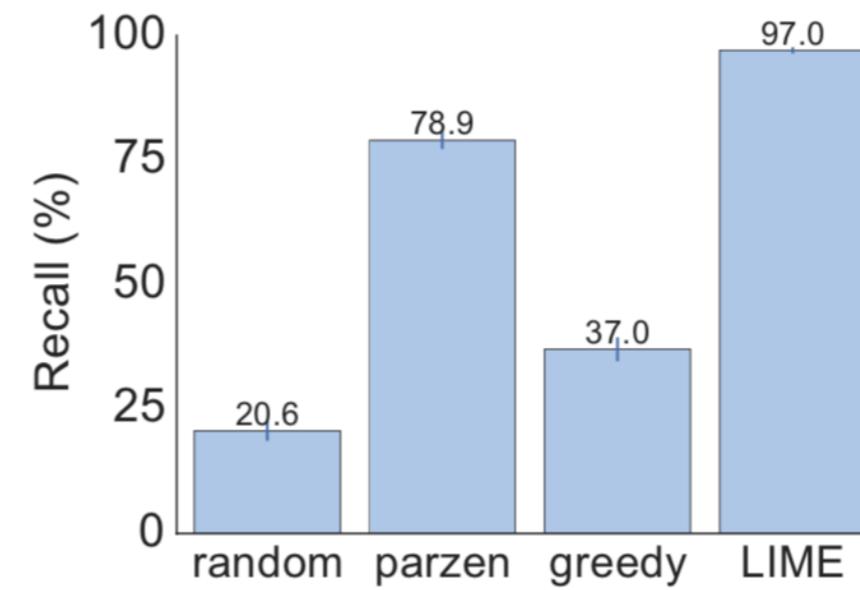
$$c(V, W, I) = \sum_{j=1}^{d'} Ind_{\exists i \in V: W_{ij} > 0} I_j$$



Результаты



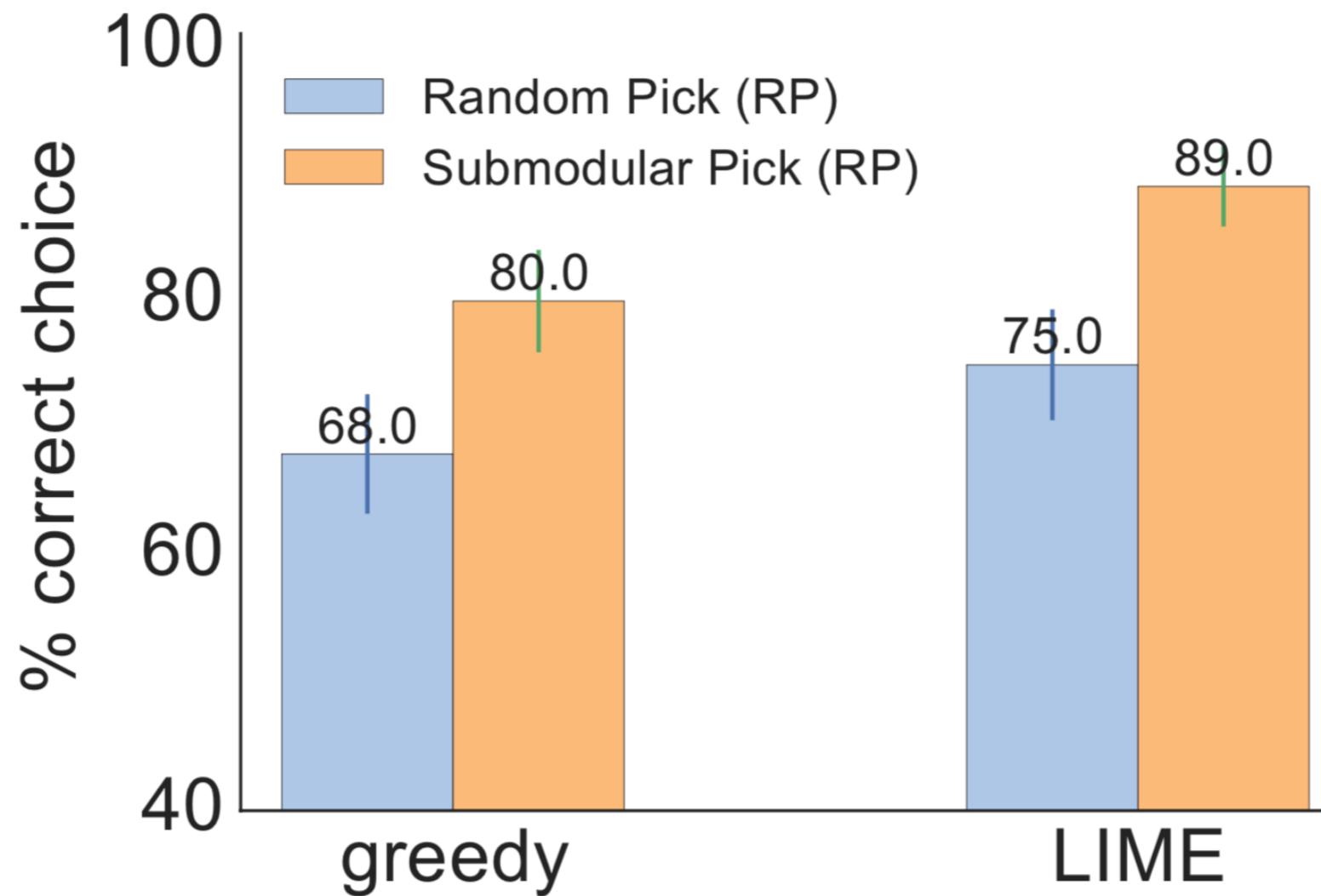
(a) Sparse LR



(b) Decision Tree

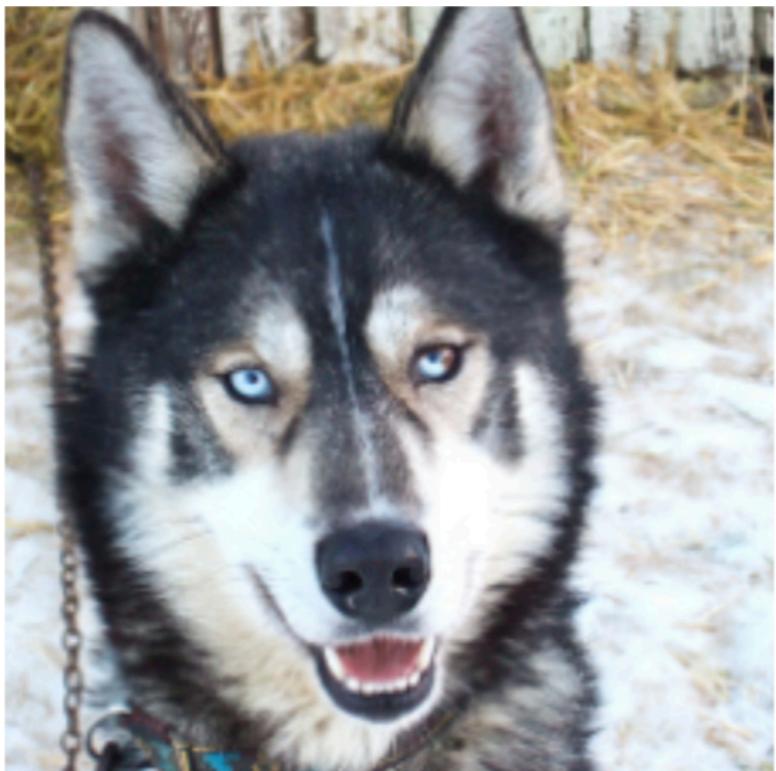
Recall важных признаков

Результаты

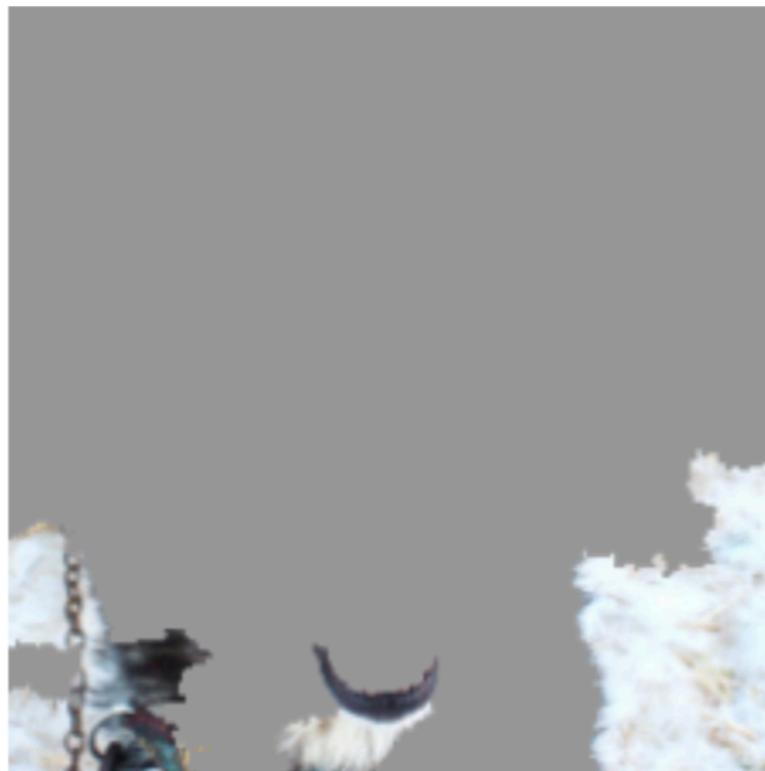


Средняя точность выбора модели

Применение для поиска ошибок в модели/данных



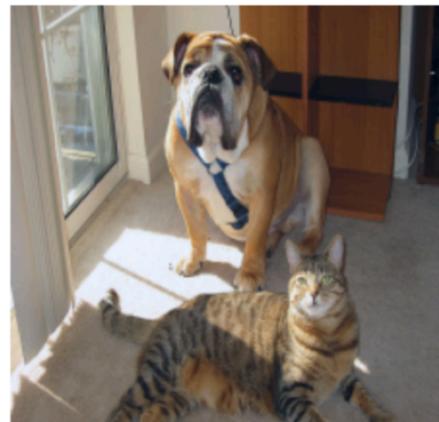
(a) Husky classified as wolf



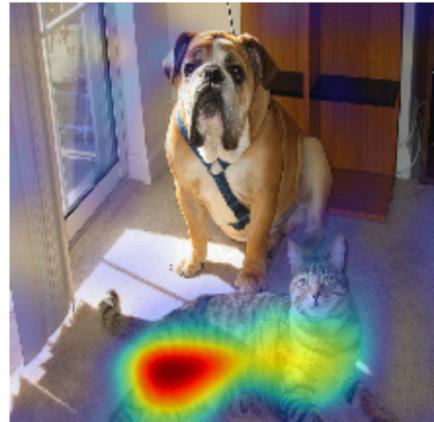
(b) Explanation

Grad-CAM (Gradient-weighted Class Activation Mapping)

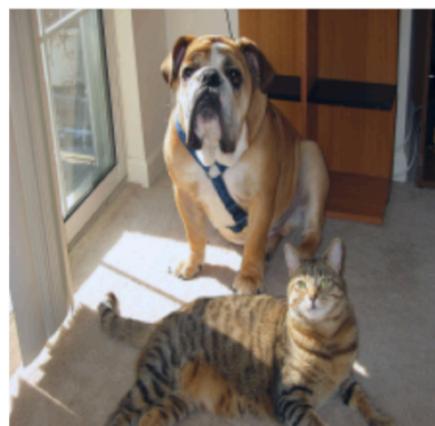
- Применяется для сверхточных сетей
- Выделяет области, наиболее значимые для предсказания



(a) Original Image



(c) Grad-CAM ‘Cat’

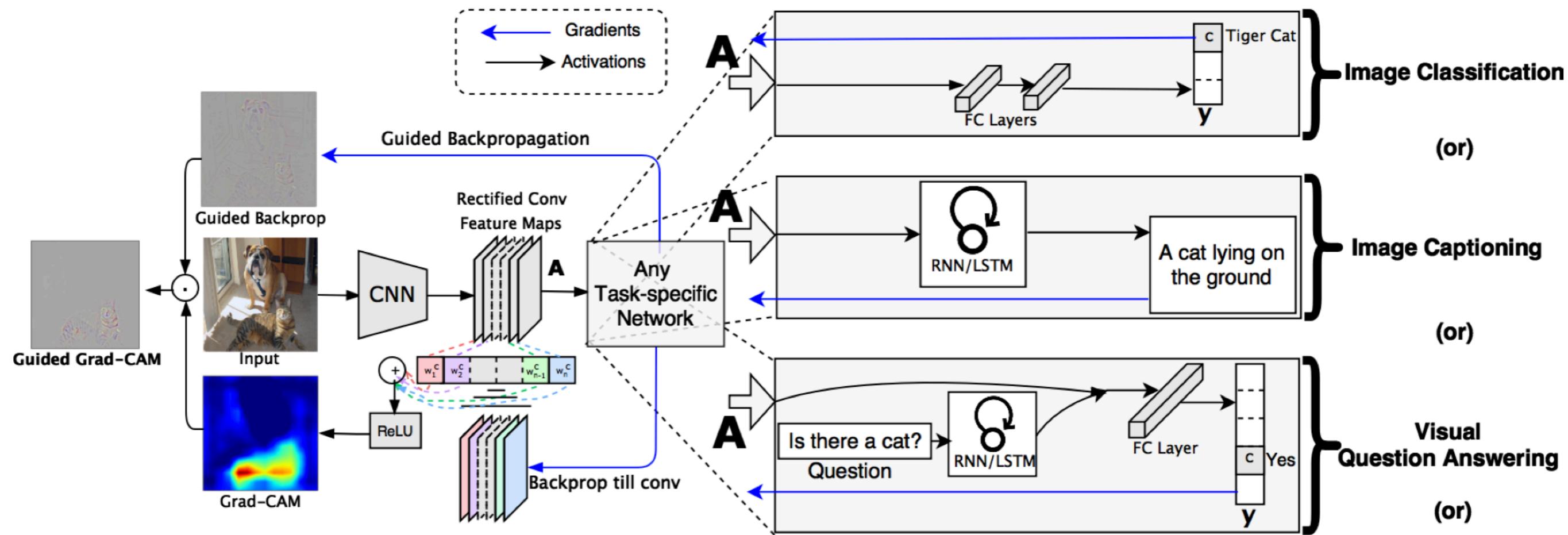


(g) Original Image



(i) Grad-CAM ‘Dog’

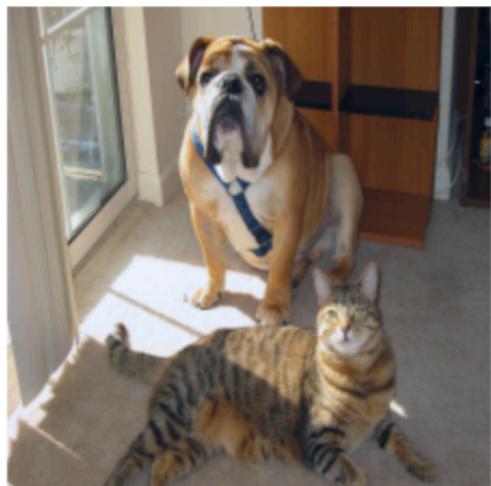
Алгоритм



$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

$$L_{GradCAM}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

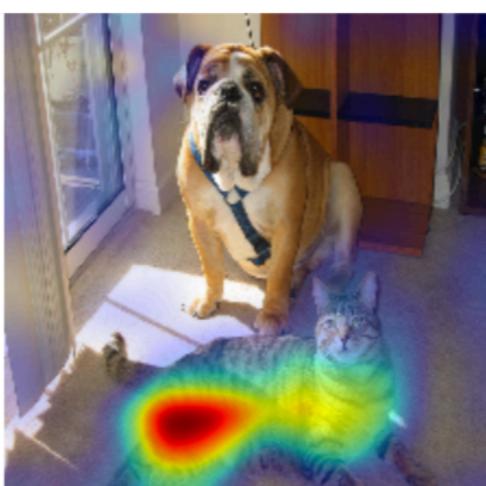
Сравнение алгоритмов



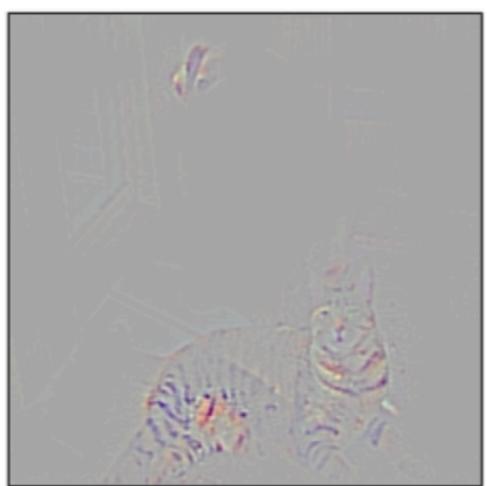
(a) Original Image



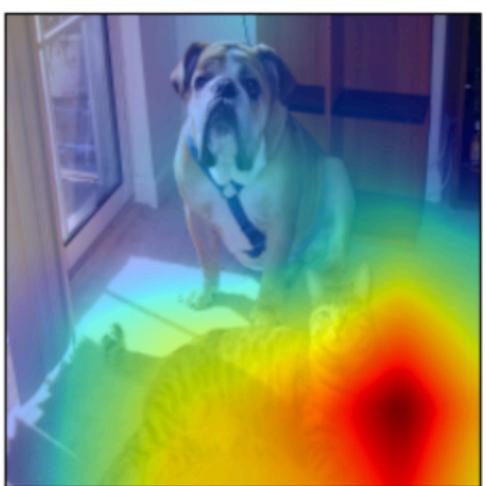
(b) Guided Backprop ‘Cat’



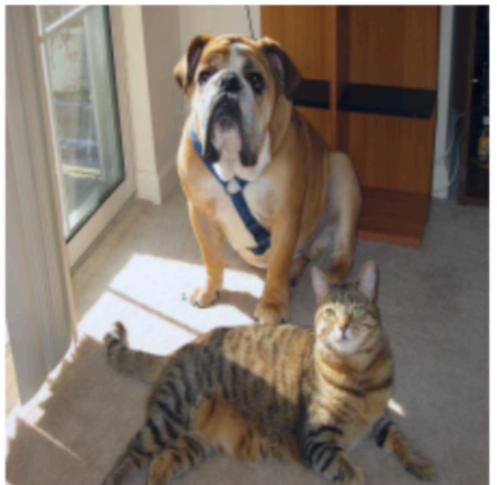
(c) Grad-CAM ‘Cat’



(d) Guided Grad-CAM ‘Cat’



(f) ResNet Grad-CAM ‘Cat’



(g) Original Image



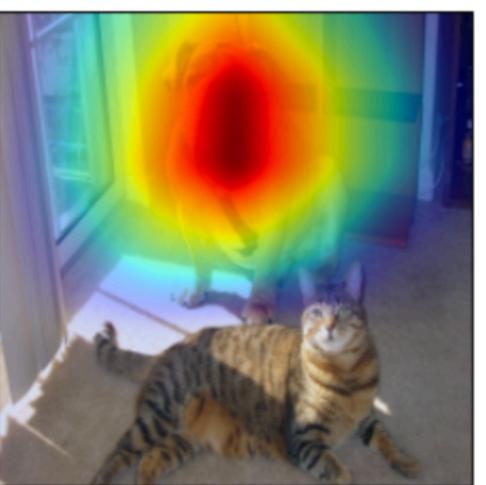
(h) Guided Backprop ‘Dog’



(i) Grad-CAM ‘Dog’



(j) Guided Grad-CAM ‘Dog’



(l) ResNet Grad-CAM ‘Dog’

Другие задачи

Grad-CAM



A group of people flying kites on a beach

Grad-CAM



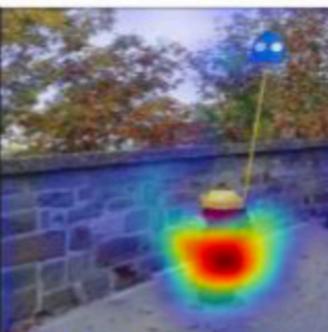
A man is sitting at a table with a pizza

Guided Backprop



What color is the firehydrant?

Grad-CAM



Guided Grad-CAM

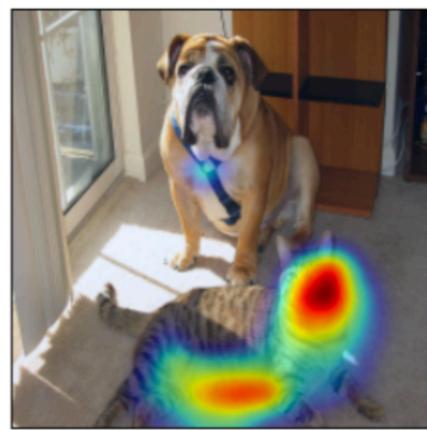
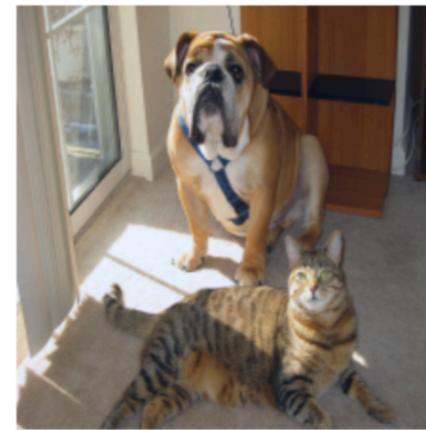
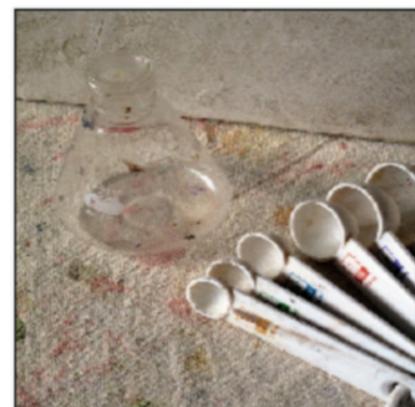
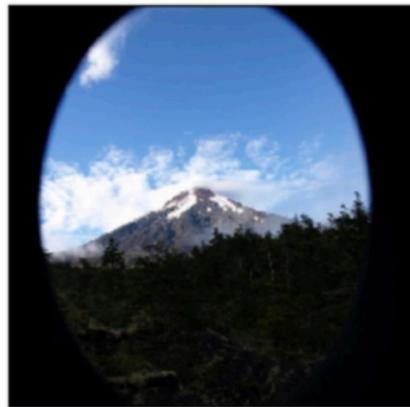


red

yellow

yellow and red

Применение для поиска ошибок в модели/данных



(a) Original Image

(b) Cat Counterfactual exp (c) Dog Counterfactual exp



Ground truth: volcano



Ground truth: beaker



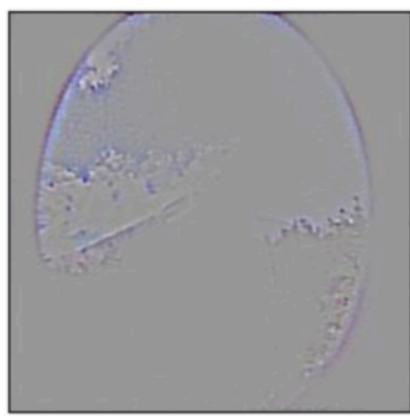
Ground-Truth: Doctor



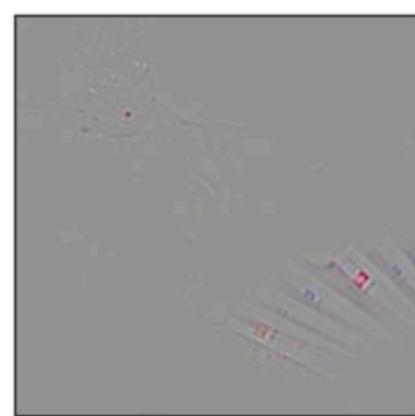
Predicted: Nurse



Predicted: Doctor



Predicted: car mirror



Predicted: syringe

Результаты

- Для задачи локализации GradCAM показывает лучшие результаты, чем другие методы (ошибка 56.51 против 57.20 для CAM и 61.12 для Backprop)
- Показывает большую корреляцию с методом затемнения частей изображения (0.254 против 0.208 для CAM и 0.168 для Guided Backprop)

SHAP (SHapley Additive exPlanations)

- Попытка объединить несколько существующих методов на основании результатов теории игр.

Методы с аддитивным влиянием признаков

- имеют модель объяснений вида

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i$$

где $z' \in \{0,1\}^M$

Примеры:

1. LIME
2. DeepLIFT
3. Layer-Wise Relevance Propagation

Вектор Шепли

- принцип оптимальности распределения выигрыша между игроками в задачах теории кооперативных игр

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$$

- можно доказать, что единственная модель с аддитивным влиянием признаков, которая удовлетворяет условиям локальной точности, условию отсутствия $(x'_i = 0 \Rightarrow \phi_i = 0)$ и условию согласованности (

$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \quad \text{then } \phi_i(f', x) \geq \phi_i(f, x)$$

вычисляет векторы Шепли

Kernel SHAP

- LIME + векторы Шелли
- LIME не вычисляет векторы Шелли при эвристическом выборе параметров, поэтому его точность или согласованность нарушается
- поэтому будем использовать следующие параметры

$$\pi_{x'}(z') = \frac{(M - 1)}{\binom{M}{|z'|} |z'| (M - |z'|)},$$
$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} [f(h_x^{-1}(z')) - g(z')]^2 \pi_{x'}(z'),$$

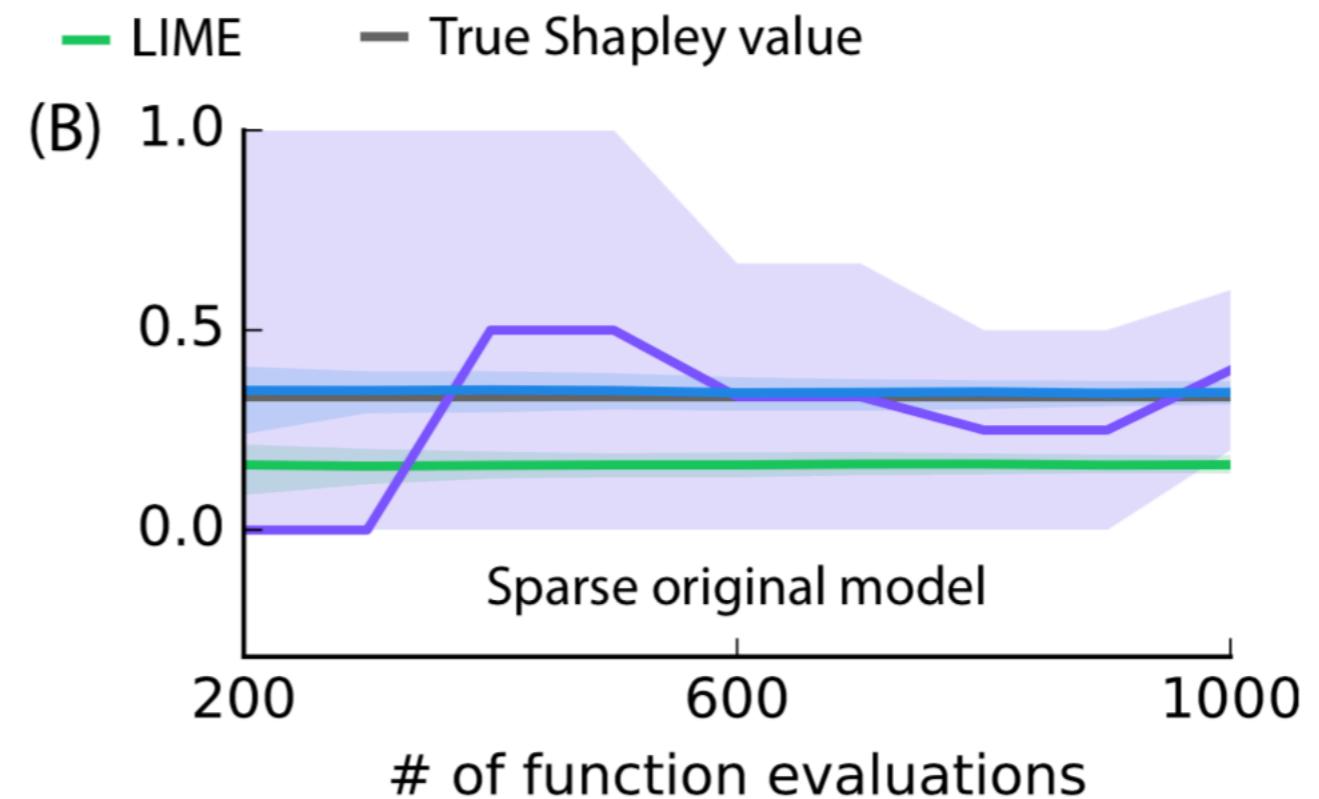
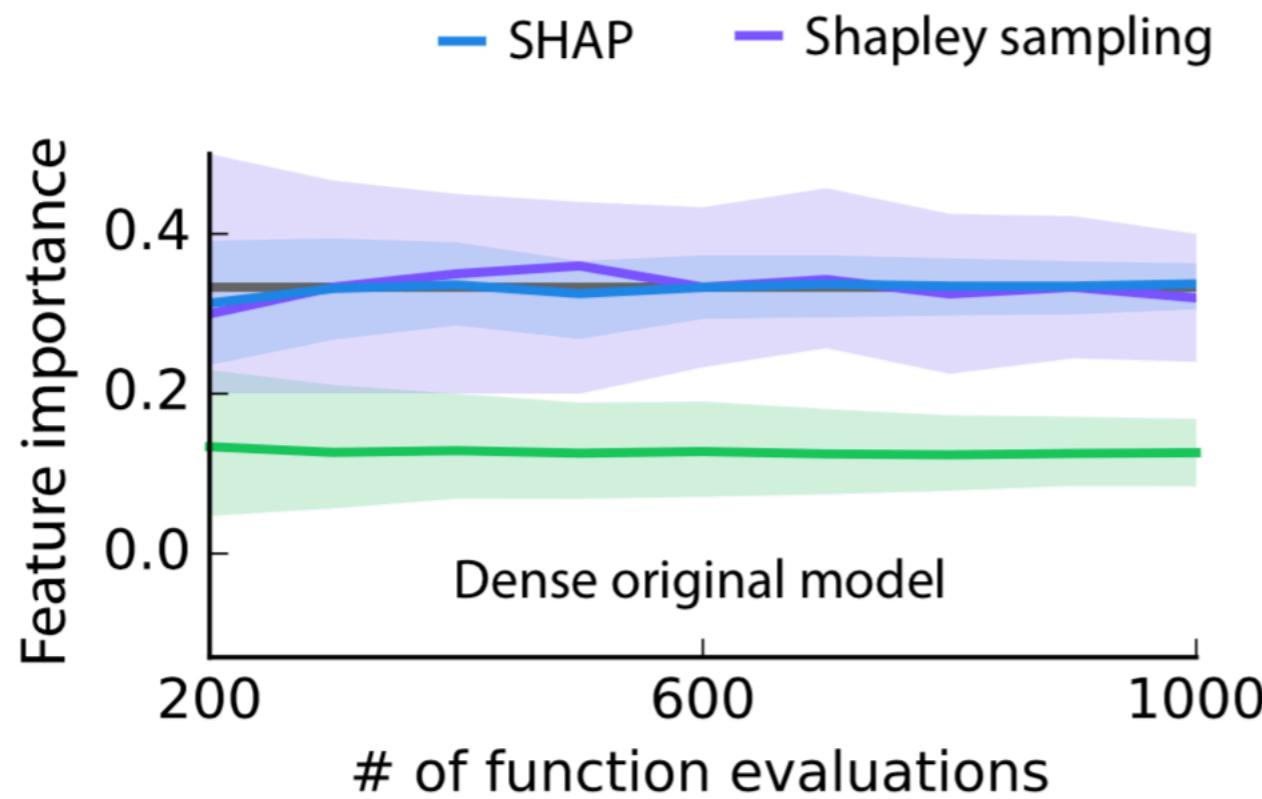
Другие методы

Специфичные для модели:

- Linear SHAP
- Low-Order SHAP
- Max SHAP
- Deep SHAP

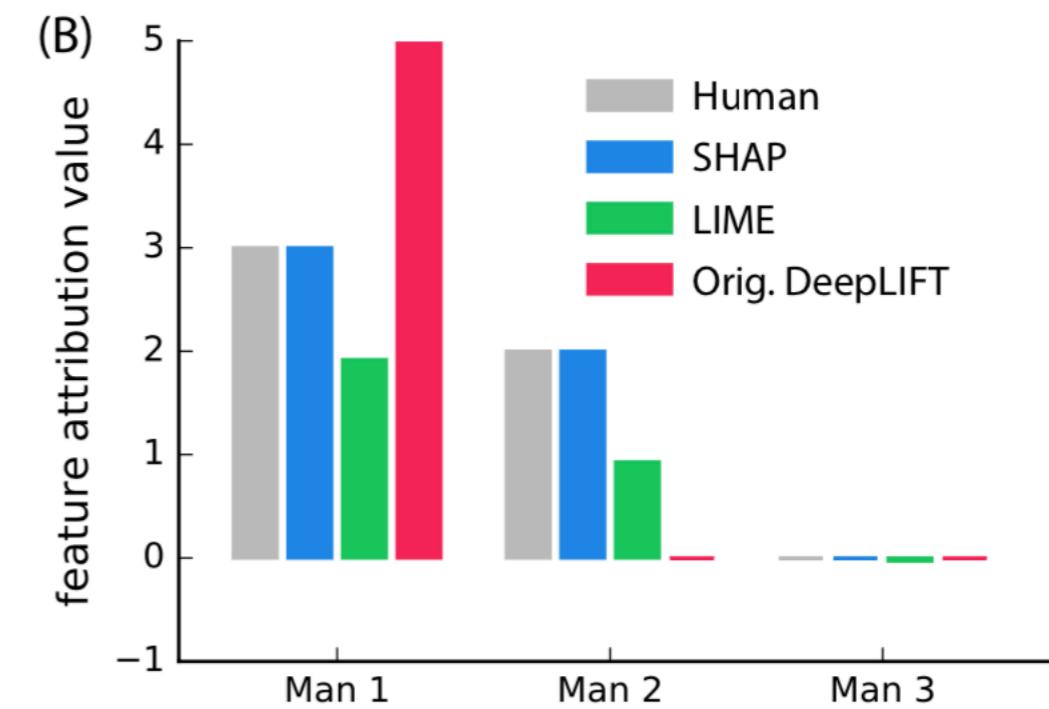
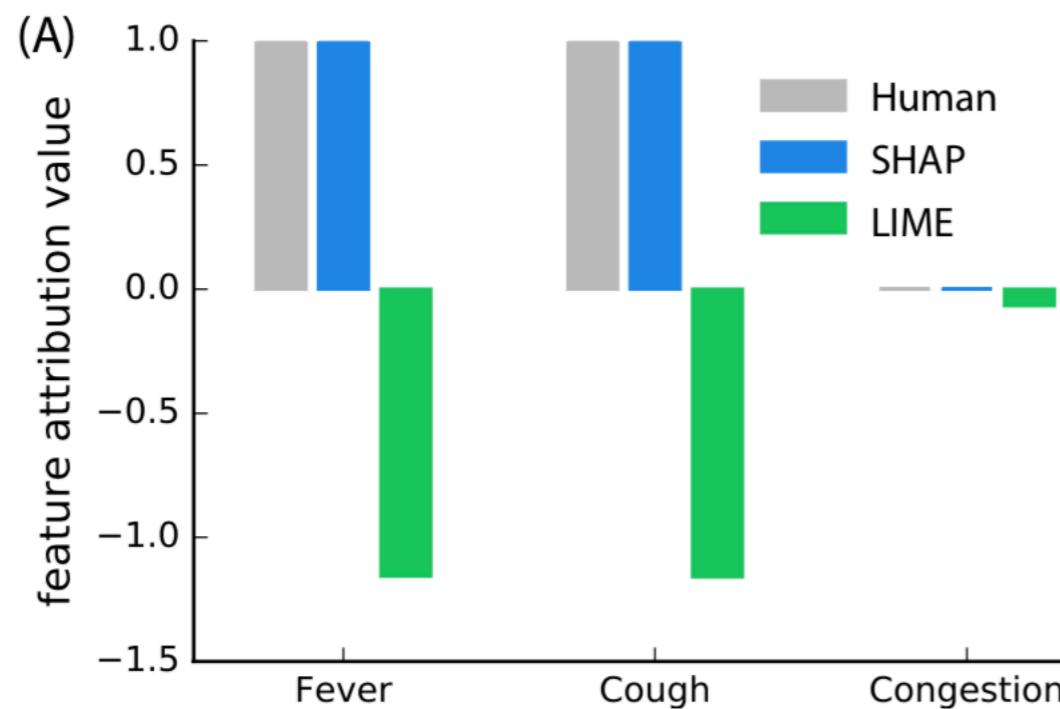
Результаты

Эффективность вычислений



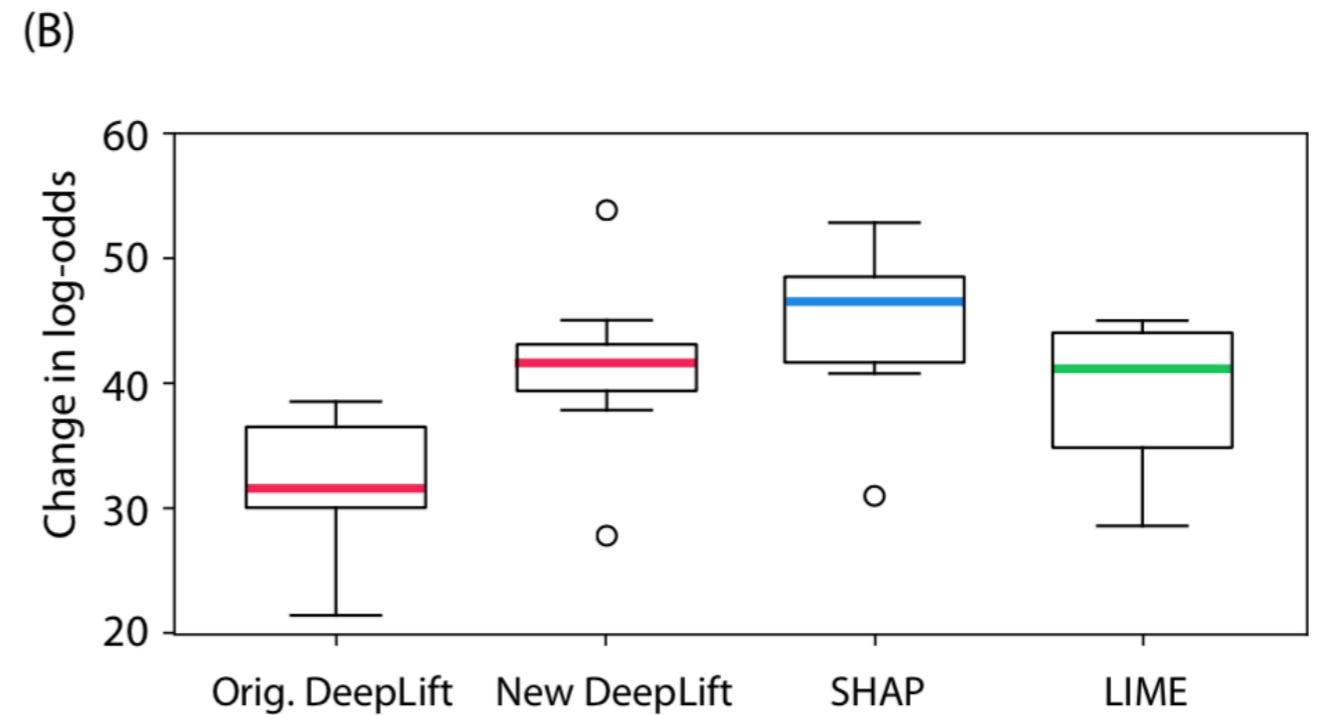
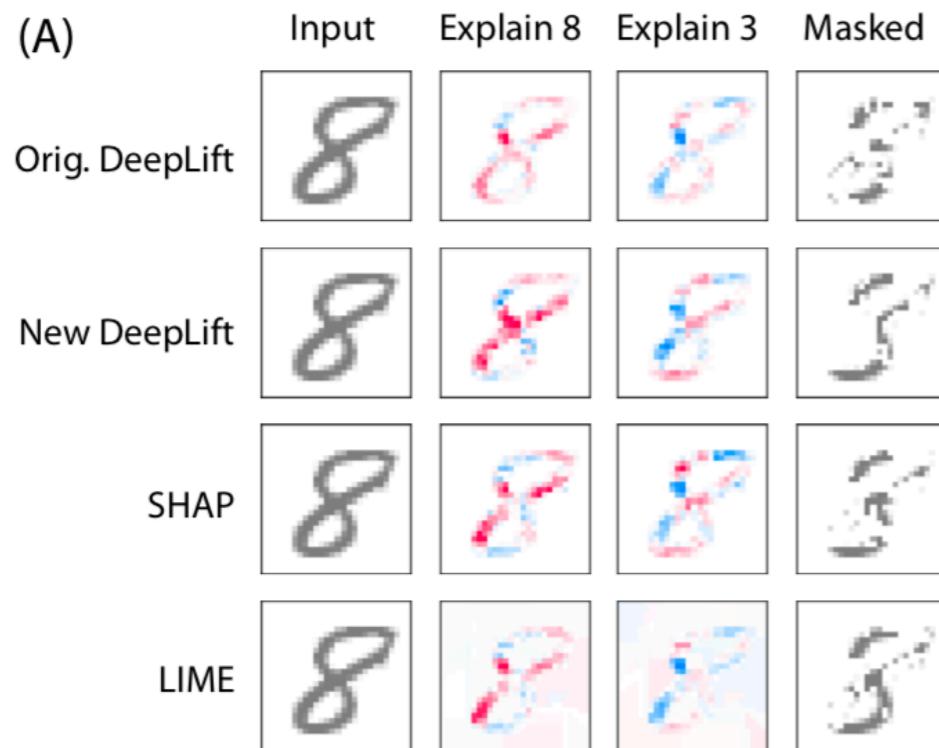
Результаты

Близость с человеческим пониманием



Результаты

Качество объяснения моделей классификации



Вопросы

- 1) Опишите алгоритм, используемый методом LIME.
- 2) Приведите примеры того, как интерпретация нейросетей помогает улучшить модель.
- 3) Какой метод позволяет одновременно хорошо локализовать релевантную область картинки и сохранить высокую чёткость для свёрточных нейросетей? В чём его суть?

Источники

1. LIME <https://arxiv.org/pdf/1602.04938.pdf>
2. SHAP <https://arxiv.org/abs/1705.07874>
3. Grad-CAM <https://arxiv.org/pdf/1610.02391.pdf>
4. Interpretability of machine learning models <https://www.kdnuggets.com/2019/05/interpretability-machine-learning-models.html>