Интерпретируемость нейронных сетей

Чёлушкин Максим БПМИ172

Что такое интерпретируемость?

Интерпретируемость – возможность объяснить или показать на понятном для человека языке.

Зачем объяснять предсказания?

- Необходимость доверия пользователя к результатам модели

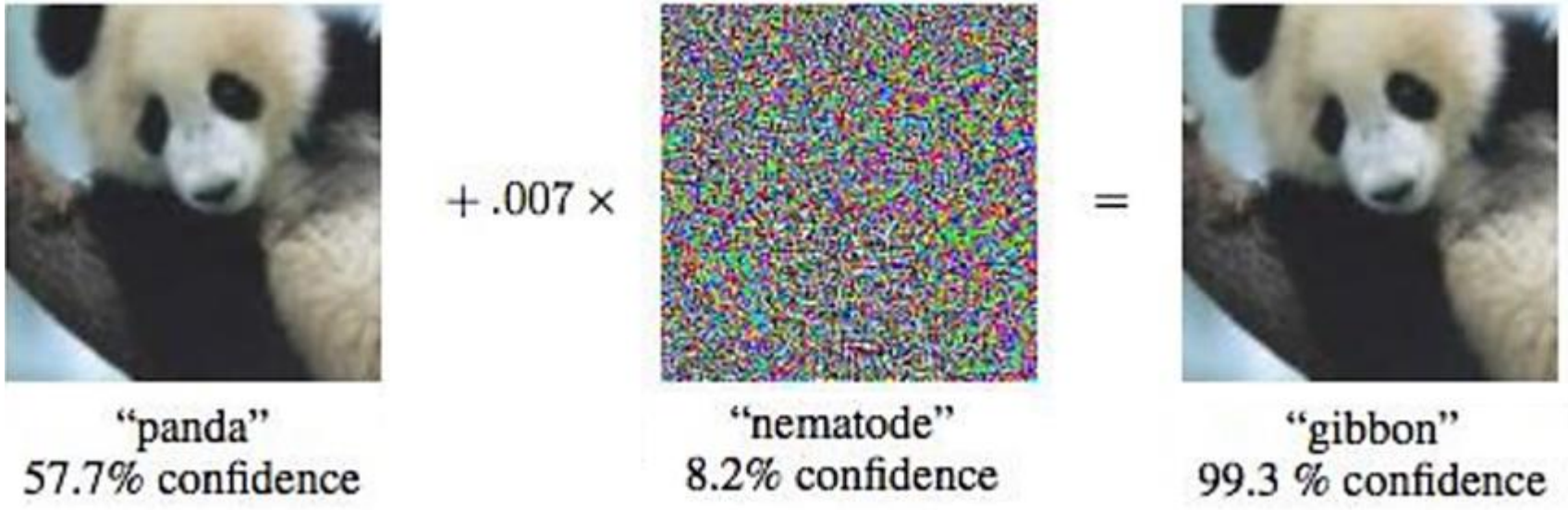
  Целевой метрики не хватает

- Получение рекомендаций относительно того, как улучшить работу модели

  Почему модель ошибается?

Зачем объяснять предсказания?

- Наличие состязательных примеров



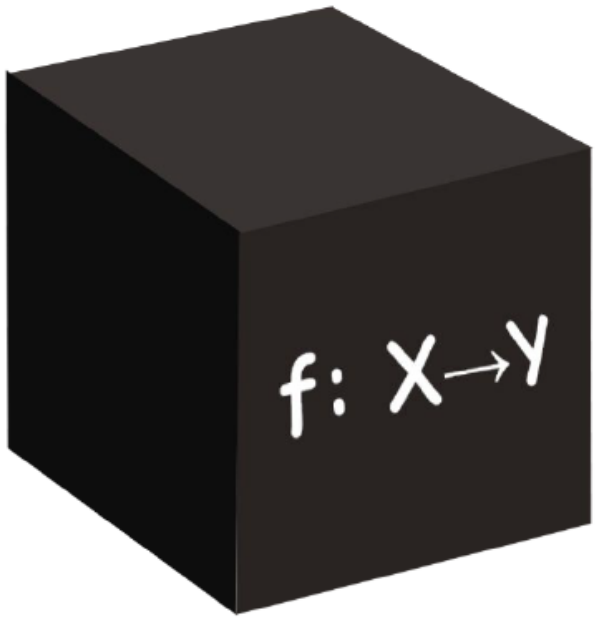
Зачем объяснять предсказания?

- Исследование модели, получение новых знаний из модели

  Что человек может узнать нового из модели?

- Этические и правовые причины

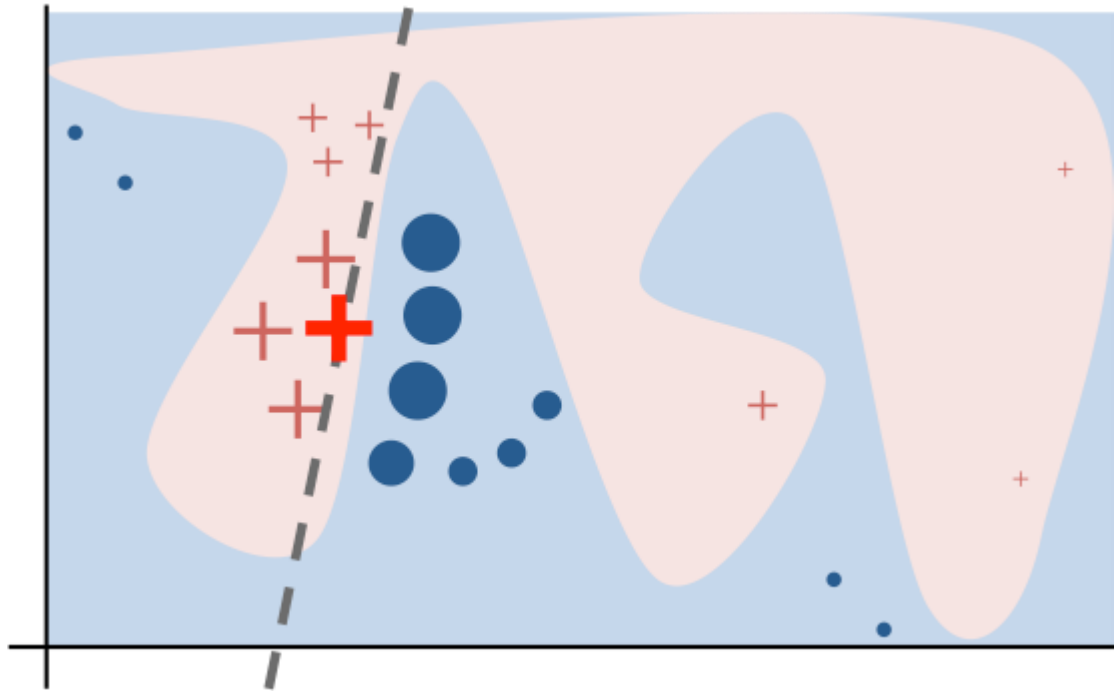
Чёрная коробка

признаки → 

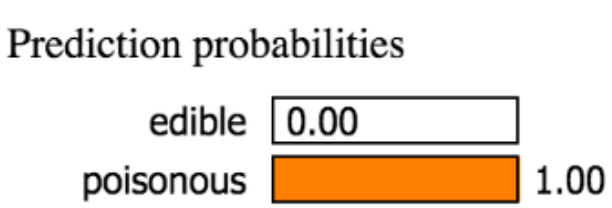
→ **предсказание**

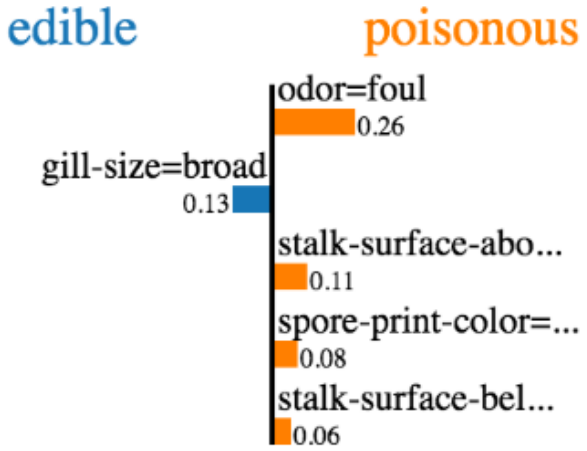
LIME-Local Interpretable Model-Agnostic

- Метод LIME может быть применен к любой модели
- Рассматривает только один пример
- Аппроксимирует модель



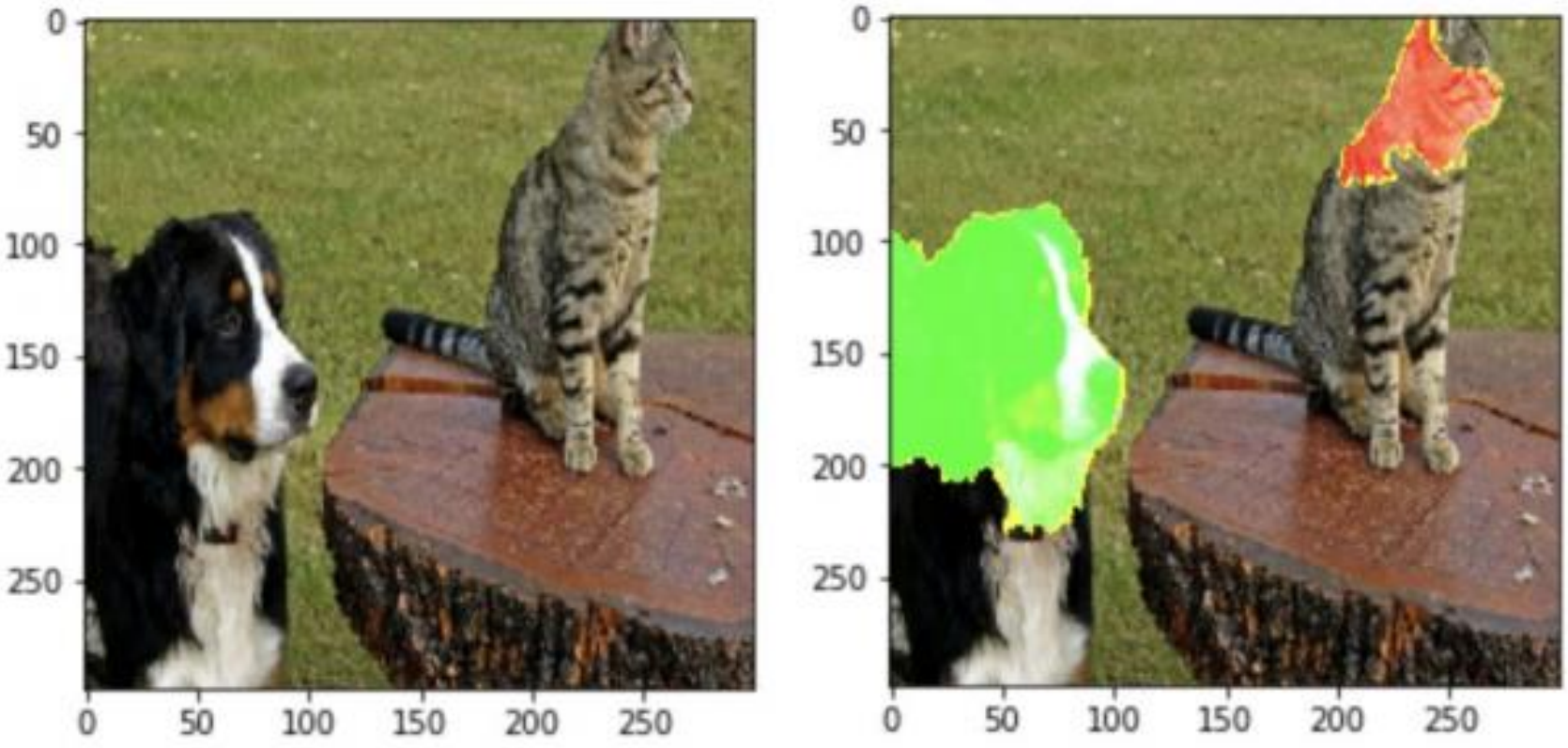
Пример LIME





Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

Пример LIME


0
50
100
150
200
250
0
50
100
150
200
250
0
50
100
150
200
250
0
50
100
150
200
250

SHAP-Shapley Additive exPlanations

- Для оценки важности признаков рассчитывается значение Шэпли:

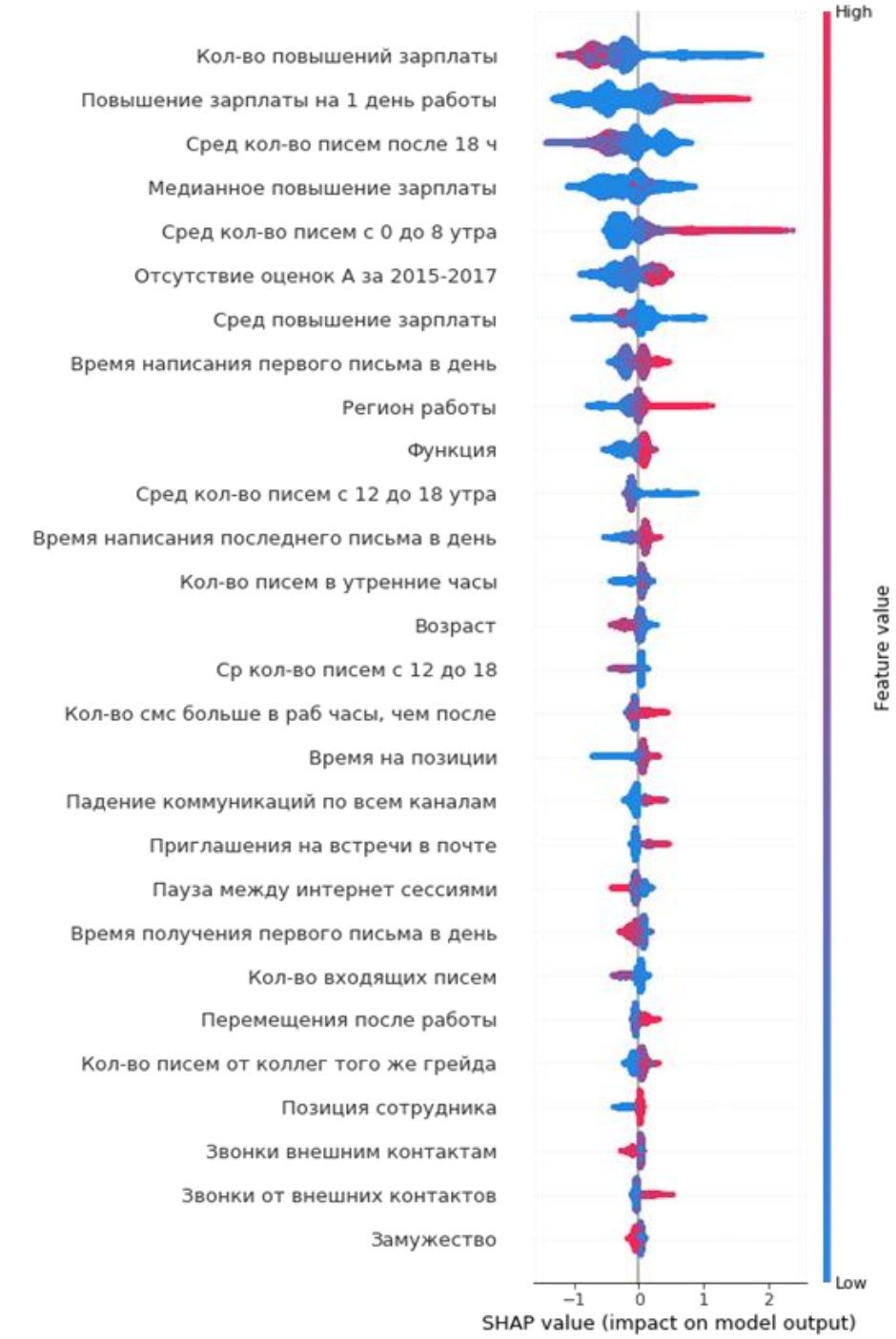
$$f_i(x) = \sum_{S \subseteq N/\{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (p(S \cup \{i\}) - p(S))$$

Здесь:
$p(S \cup \{i\})$ - Предсказание модели с i-ым признаком
$p(S)$ - Предсказание модели без i-го признака
n – Кол-во признаков
S – произвольный набор признаков без i-го

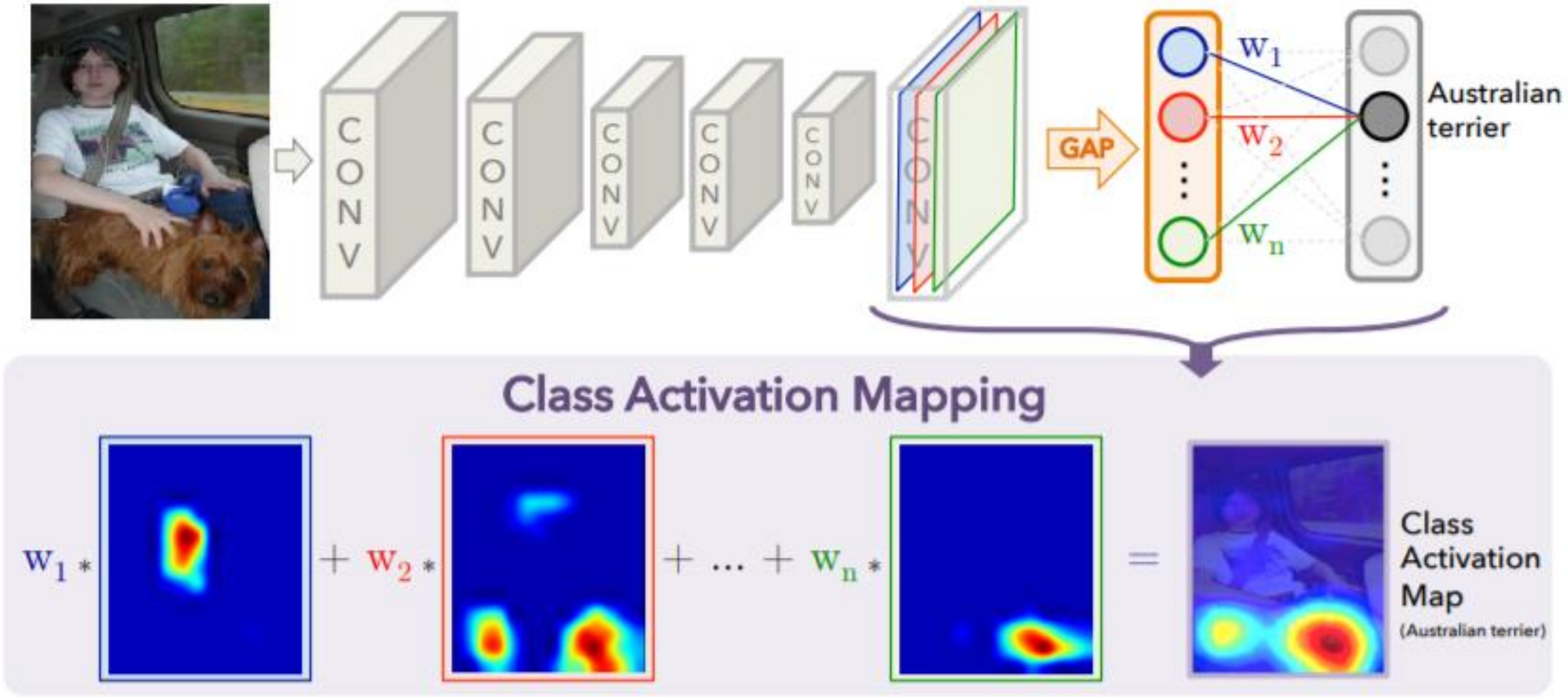
Примеры SHAP

Как читать график важности признаков:

1. значения слева от центральной вертикальной линии — это negative класс, справа — positive
2. чем толще линия на графике, тем больше таких точек наблюдения
3. чем краснее точки на графике, тем выше значения фичи в ней



CAM: Class Activation Maps

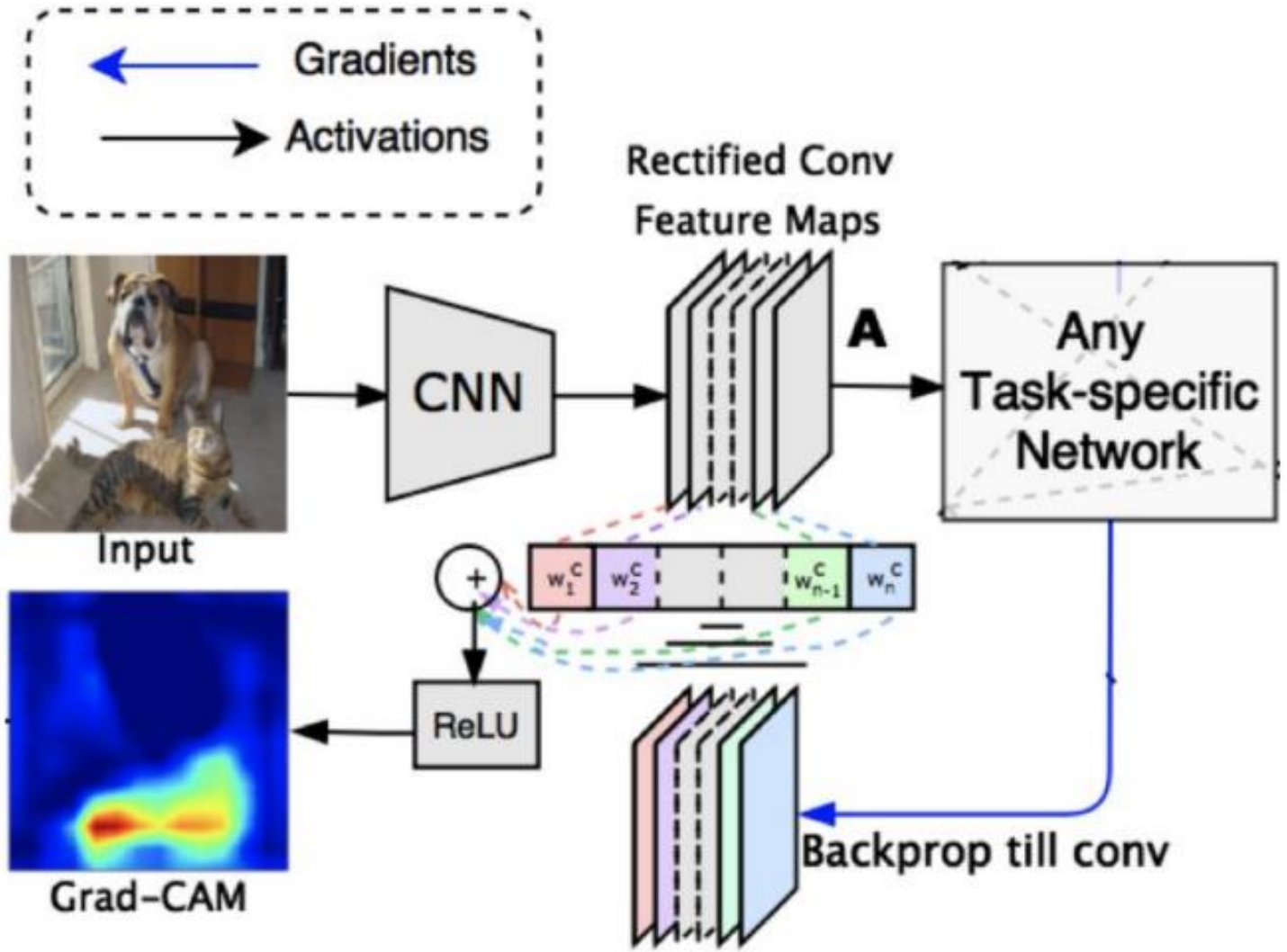


GRAD-CAM

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

α – Веса, показывающие важность фильтра k для таргет-класса c

$$L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$



GRAD-CAM Пример

A bedroom with a bed and a desk

Плюсы и минусы методов:

LIME

Плюсы:

- Подходит для любой модели
- Высокая точность

Минусы:

- Медленно работает – надо сделать много сэмплингов

SHAP

Плюсы:

- Подходит для любой модели
- Хорошая визуализация

Минусы:

- Вычисления значений Шэпли очень затратны

GRAD-CAM

Плюсы:

- Быстро работает
- Достоверный
- Подходит для большинства моделей

Вопросы:

- Как происходит оценка важности отдельного признака с помощью метода SHAP?
- Какие плюсы и минусы метода LIME?
- С помощью какого метода интерпретируется классификация объектов на картинке

Источники

- https://arxiv.org/abs/1610.02391
- https://arxiv.org/abs/1602.04938
- https://habr.com/ru/post/428213/
- https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b