Интерпретируемость нейронных сетей

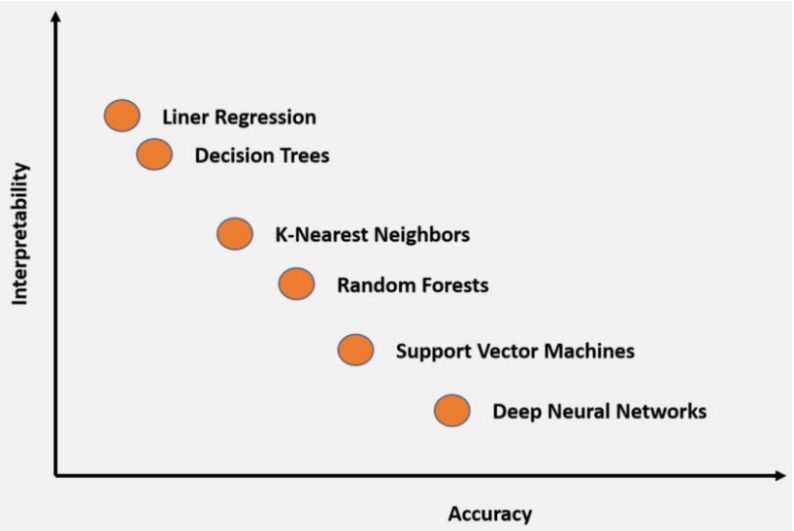
Работу выполнил:
студент НИУ ВШЭ ПМИ 182
Пак Ди Ун

Проблематика

1. Как объяснить заказчику, что мы сделали?
2. А себе?
3. Чем отличаются модели и почему они ведут себя по-разному на определенном предсказании?

Вспомним, что умеем

1. Линейная регрессия
2. Linear SVM
3. Решающие деревья
4. Случайный лес



Local Interpretable Model-Agnostic Explanations

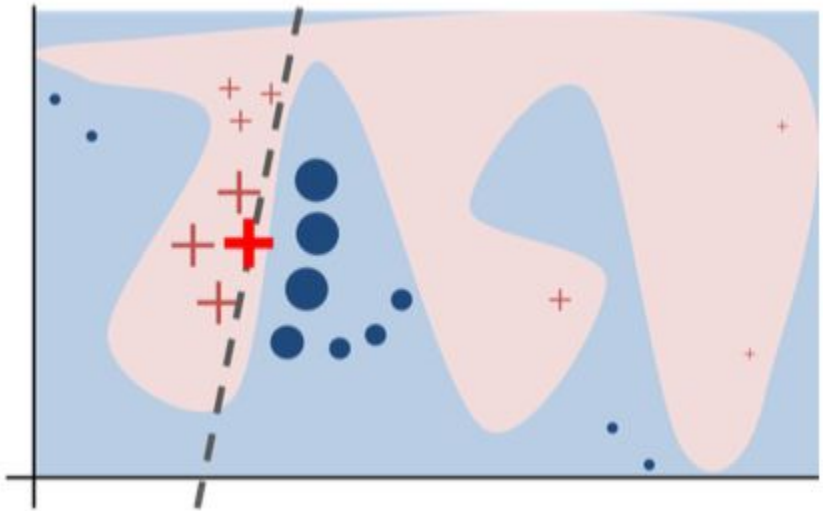
Local: локальная точность. Результаты LIME верны хотя бы в окрестности данной точки

Interpretable: результаты LIME могут быть проинтерпретированы человеком

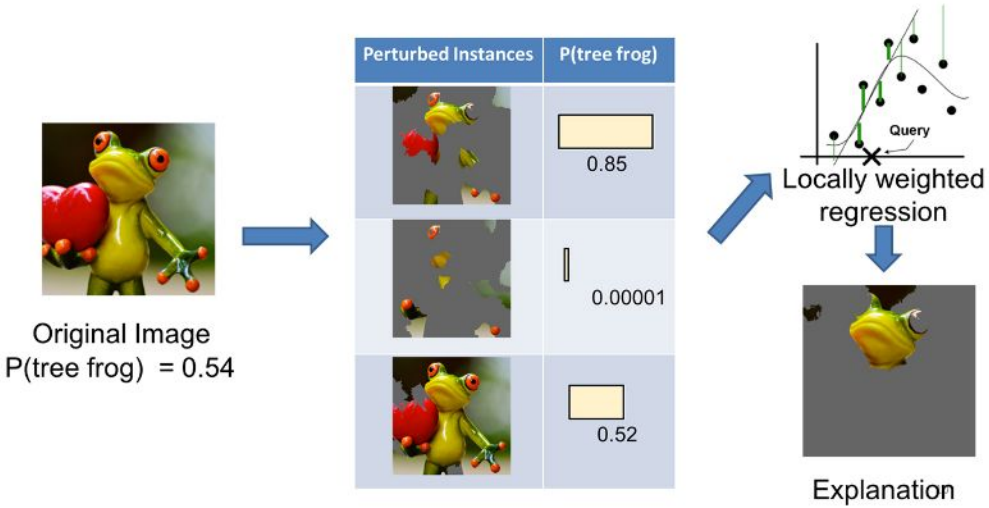
Model-Agnostic: LIME обращается с моделью как с черной коробкой

LIME: how it works

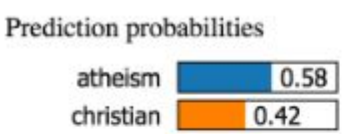
1. Чуть-чуть поменяем вывод и прогоним через модель
2. Присвоим веса в соответствии с расстоянием до исходной точки
3. Построим линейную модель на получившихся точках
4. Интерпретируем линейную модель

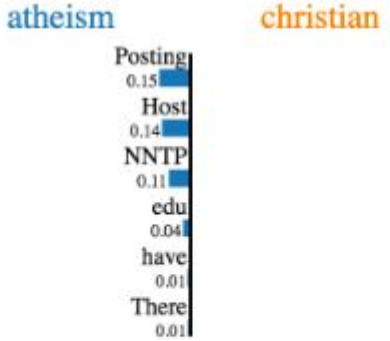


Примеры результатов работы LIME



Примеры результатов работы LIME





Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

Плюсы и минусы LIME

Плюсы:

- Независимость от основной модели
- Не требует сложных и долгих вычислений

Минусы:

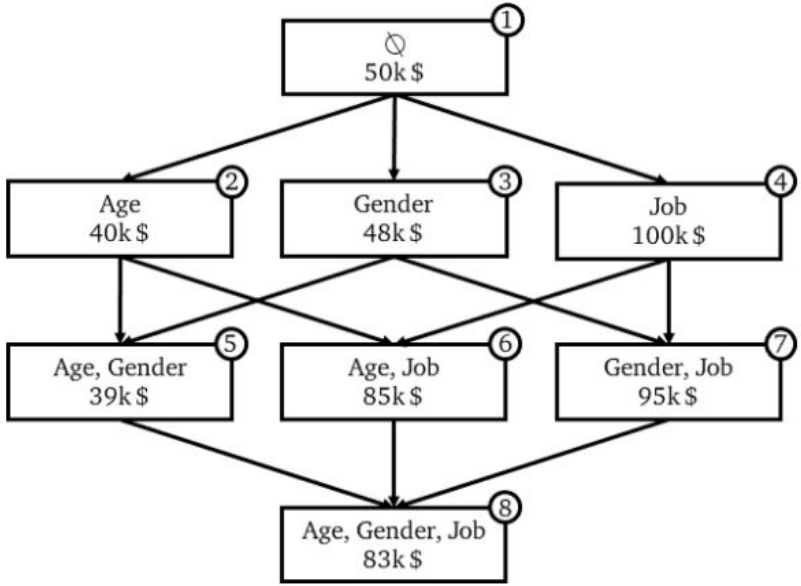
- Не во всех случаях можно локально точно предсказать линейной моделью
- Нет глобальной интерпретации

SHapley Additive exPlanations (SHAP)

1. Результаты SHAP показывают, как особенность влияет на предсказания модели в целом
2. Они также объясняют локальный вклад в одно конкретное предсказание
3. SHAP так же как и LIME обращается с моделью как с черной коробкой

SHAP: how it works

1. Обучим 2^N моделей на всевозможных подмножествах особенностей
2. Построим граф и проставим ребра в соответствии с правилами
3. Зафиксируем наблюдение и для каждой особенности посчитаем взвешенную сумму вкладов в оценку
4. Для общей оценки важности особенности в модели возьмем средний вклад по всем наблюдениям

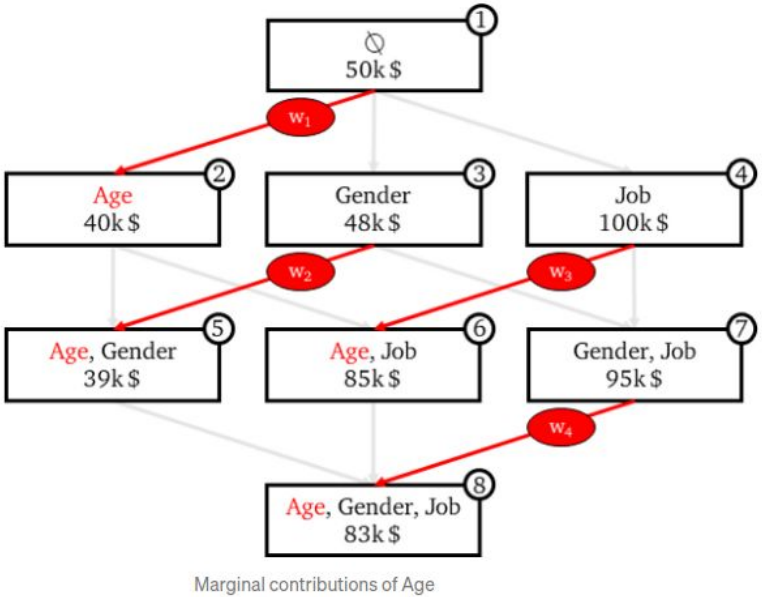


Marginal contributions of Age

Выставление весов

1. Суммы весов в одном ряду должны сохранятся ($w_1 = w_2 + w_3 = w_4$)
2. Веса в одной одном ряду равны ($w_2 = w_3$)
3. Сумма весов равна 1

Тогда вес w на i-ом слое равен $\left(i \frac{i}{N}\right)$



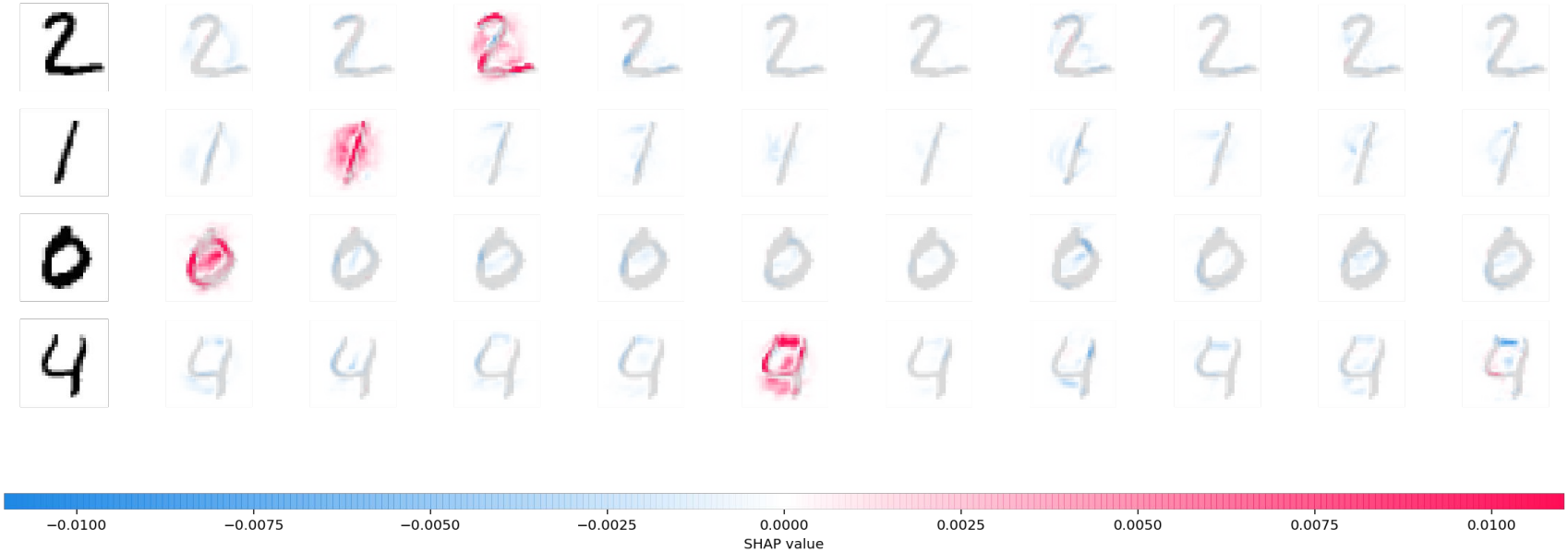
Marginal contributions of Age

Итоговый подсчет

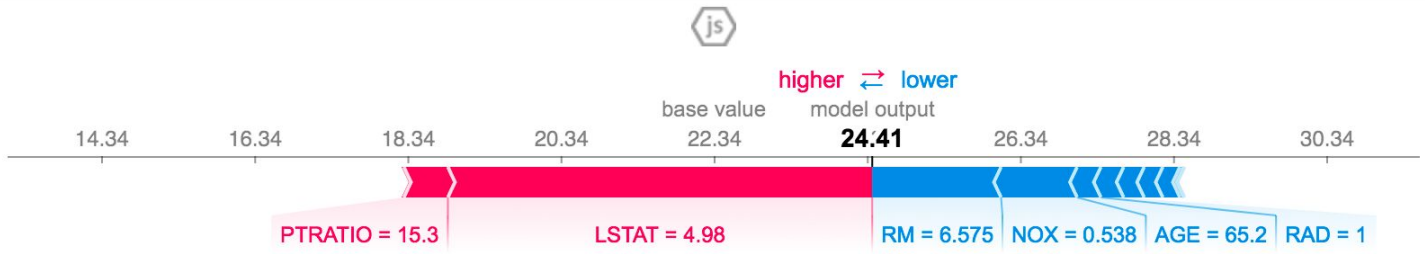
$$SHAP_{feature}(x) = \sum_{set:feature \in set} [|set| \times \binom{F}{|set|}]^{-1} [Predict_{set}(x) - Predict_{set \setminus feature}(x)]$$

$$\begin{aligned}
SHAP_{Age}(x_0) = {} & [(1 \times \binom{3}{1}]^{-1} \times MC_{Age,\{Age\}}(x_0) \; + \\
& [(2 \times \binom{3}{2}]^{-1} \times MC_{Age,\{Age,Gender\}}(x_0) \; + \\
& [(2 \times \binom{3}{2}]^{-1} \times MC_{Age,\{Age,Job\}}(x_0) \; + \\
& [(3 \times \binom{3}{3}]^{-1} \times MC_{Age,\{Age,Gender,Job\}}(x_0) \; + \\
= {} & \frac{1}{3} \times (-10k\$) + \frac{1}{6} \times (-9k\$) + \frac{1}{6} \times (-15k\$) + \frac{1}{3} \times (-12k\$) \\
= {} & -11.33k\$
\end{aligned}$$

Примеры результатов работы SHAP



Примеры результатов работы SHAP



Плюсы и минусы SHAP

Плюсы:

- Независимость от основной модели
- Дает как локальные, так и глобальные интерпретации

Минусы:

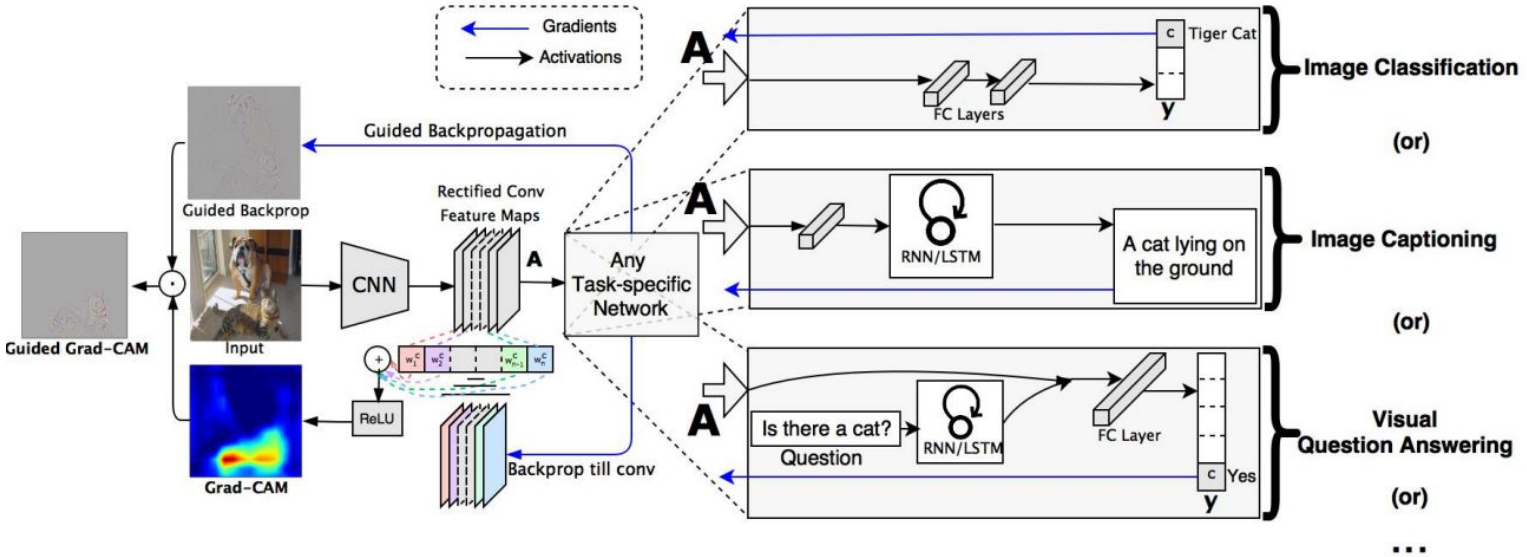
- Очень долго работает

Gradient-weighted Class Activation Mapping

А как объяснить сверточные нейронные сети?

Где находятся важные признаки?

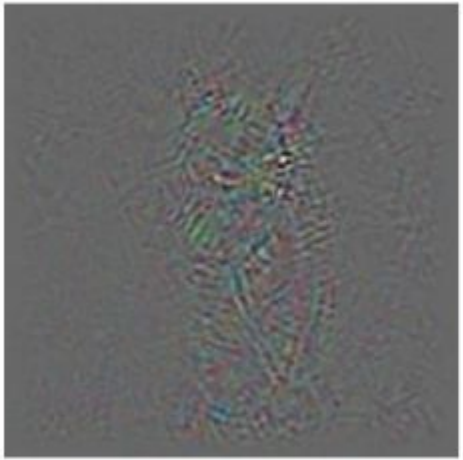
Grad-CAM architecture



Backprop vs guided backprop

Для поиска важных частей картинки сделаем backpropagation

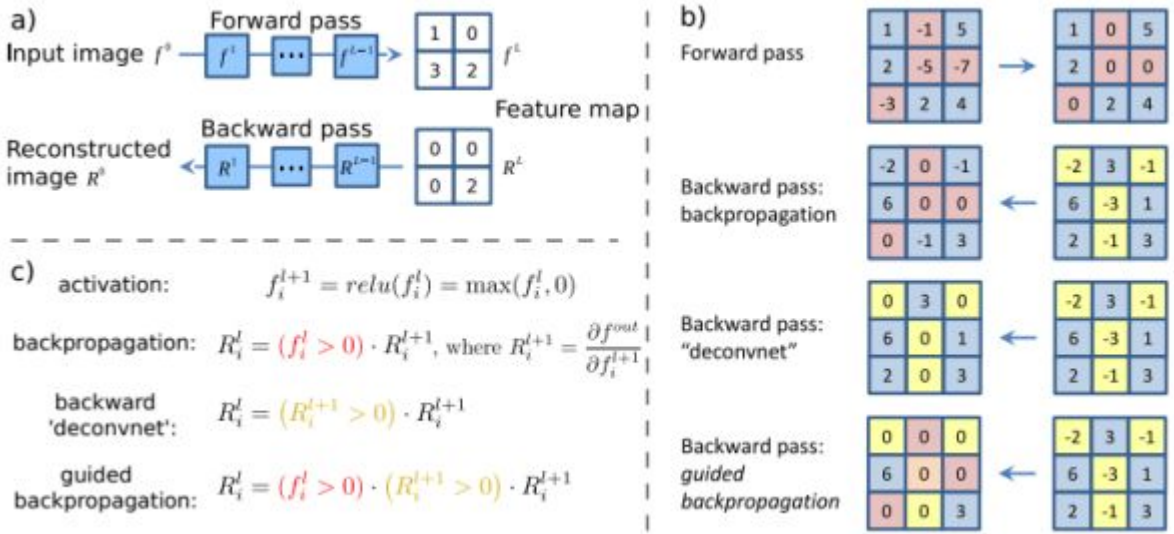
Guided backpropagation помогает убрать шум, обращая внимание только на положительные части градиента



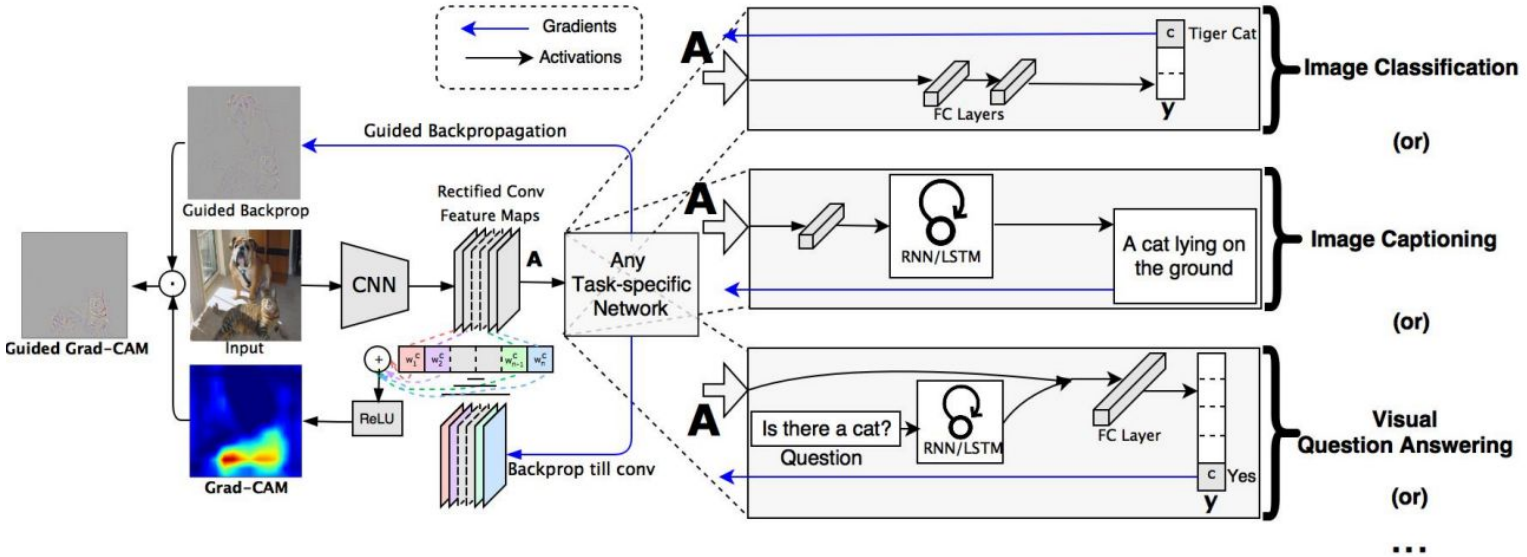
Backprop

Guided Backprop

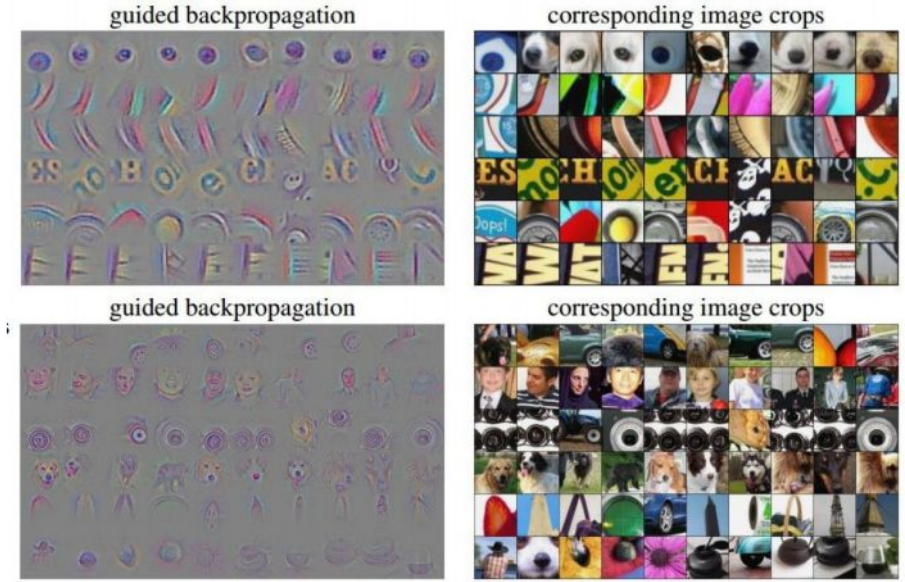
Guided backprop



Grad-CAM architecture

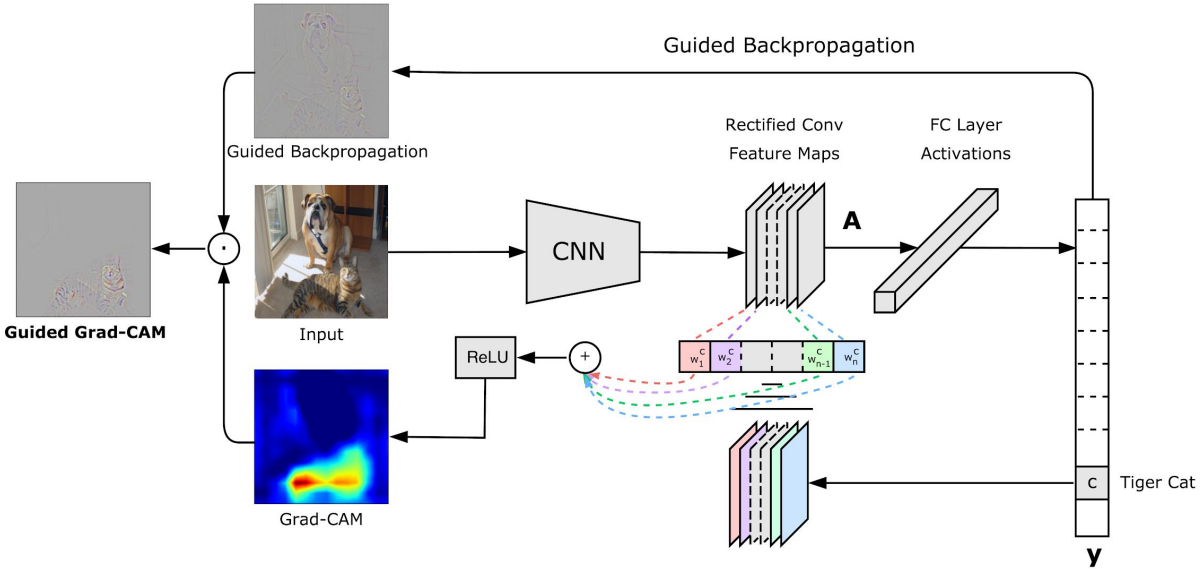


Guided backprop results

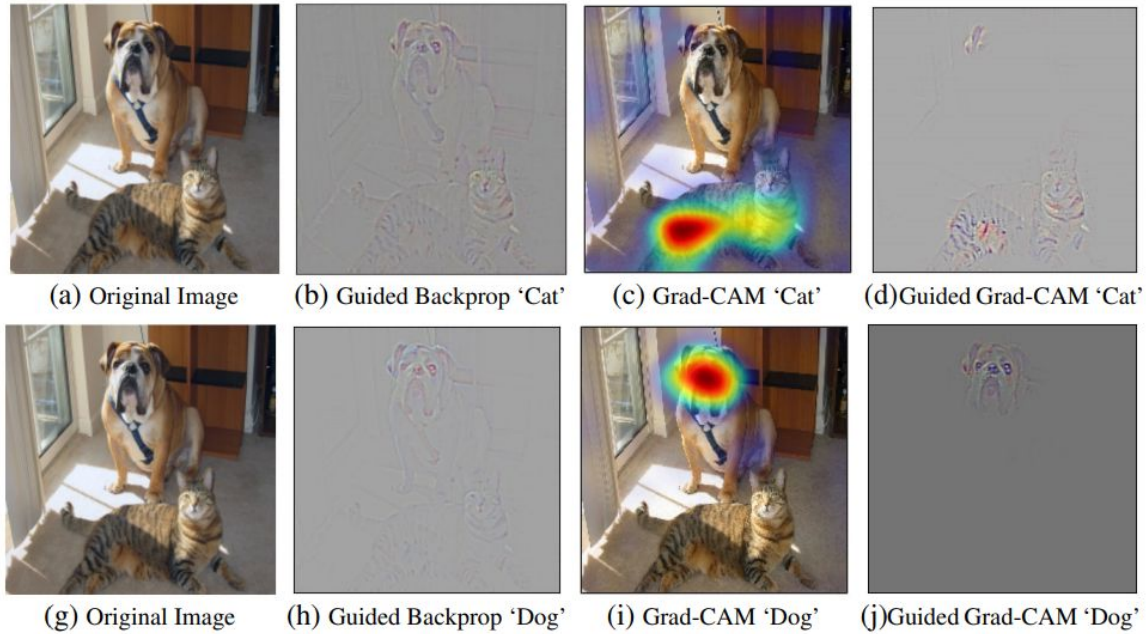


Springerberg et al, Striving for Simplicity: The All Convolutional Net (ICLR 2015 workshops)

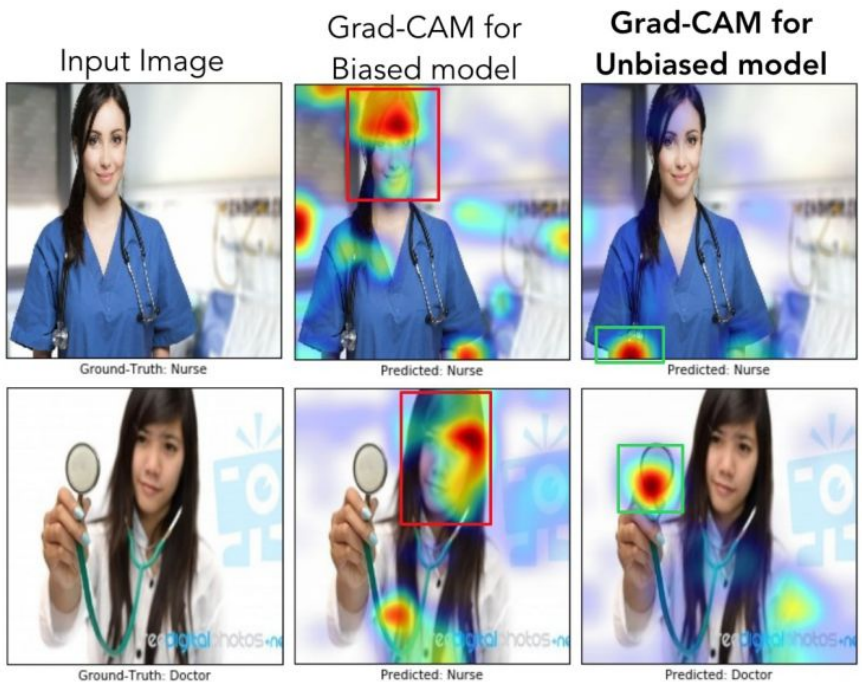
Grad-CAM


Guided Backpropagation
Guided Backpropagation
Rectified Conv
Feature Maps
FC Layer
Activations
A
CNN
Guided Grad-CAM
Input
ReLU
Grad-CAM
C
Tiger Cat
y

Grad-CAM intuition



Примеры результатов работы Grad-CAM



Примеры результатов работы Grad-CAM

Плюсы и минусы Grad-CAM

- Не требует архитектурные изменения и переобучение модели
- Работает быстро
- Работает только со сверточными нейронными сетями

Заключение

1. Обсудили известные способы интерпретации моделей
2. Узнали методы LIME, SHAP, обсудили их особенности для произвольных моделей
3. Посмотрели на способ интерпретировать сверточные нейронные сети Grad-CAM

Источники

SHAP values explained exactly how you wished someone explained to you:
https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30

Explain Your Model with the SHAP Values:
https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

Explain Your Model with LIME:
https://medium.com/analytics-vidhya/explain-your-model-with-lime-5a1a5867b423

Grad-CAM:
https://arxiv.org/abs/1610.02391

Grad-CAM:
https://medium.com/@mohamedchetoui/grad-cam-gradient-weighted-class-activation-mapping-ffd72742243a