# Разработка адверсативных моделей для оценки качества изображений

***Кириллов Алексей Константинович***

*Студент*

*Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия*

*E-mail:* alexey.kirillov@graphics.cs.msu.ru

***Научный руководитель*** *— Ватолин Дмитрий Сергеевич*

Задача оценки качества изображений (Image quality assessment) — это задача определения технического качества изображения, то есть степени присутствующих на нем дефектов. К ним можно отнести шум, размытие или артефакты сжатия jpeg.

В настоящее время существуют модели оценки качества изображений, которые достигают уровня точности, сравнимого с человеческим [1]. Интересным является вопрос о том, как именно работают такие модели, то есть на какие области изображения они обращают больше внимания, а на какие — меньше.

Для интерпретации моделей существует множество методов. Одним из наиболее распространённых является GradCAM [2]. Этот метод усредняет каналы признаков $A^k$ с последнего слоя с весами $\alpha_c^k$, вычисляемыми как значение градиента по предсказанию модели. Большинство других существущих методов являются вариациями GradCAM. Например, в HiResCAM [3] карты активаций поточечно перемножаются с градиентами, а в GradCAM++ [4] используется вторые производные.

В данной работе представлена дифференцируемая версия GradCAM. С её помощью можно обучать модели, которые смотрят в определённые области изображения. Например, на салиентные регионы, в углы или на определённый паттерн.

Для создания дифференцируемой версии GradCAM предлагается зафиксировать коэффициенты $\alpha_c^k$, оторвав их от вычислительного графа, и считать градиент только по картам активациям $A^k$. Таким образом, при дифференцировании GradCAM градиенты вычисляются только по весам бекбона.

Для проверки работоспособности предложенного метода были обучены четыре модели: одна без использования лосса по GradCAM и три модеоли, обученные смотреть на салиентные регионы, в углы и на паттерн VG. На рисунке 1 для них приведены визуализации GradCAM.

Результаты экспериментов показали, что оптимизация функции

потерь по GradCAM заставляет модель смотреть на нужные области. При этом точность модели на изначально решаемой задачи не падает.

Предложенный метод позволяет обучать адверсативные модели. Внешне они не отличаются от обычных, но внутри они маркированы и фокусируются только на заданных областях. Это вызывает сомнения в надёжности существующих методов интерпретации моделей.

**Иллюстрации**

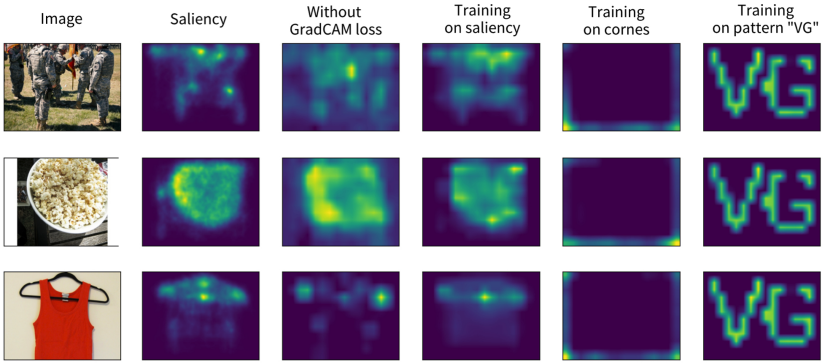


Рис 1. Визуализация GradCAM для обученных моделей

**Литература**

1. C Chen. Topiq: A top-down approach from semantics to distortions for image quality assessment // In Proceedings of the Transactions on Image Processing, 2023.
2. Ramprasaath R. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization // In IEEE International Conference on Computer Vision (ICCV), 2017, P. 234–241.
3. Rachel L. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks // In arxiv prerpint 2020, arXiv:2011.08891
4. Ronneberger O. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks // In arxiv prerpint, 2020, arXiv:1710.11063