ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix Wichmann, Wieland Brendel

Практик-исследователь Сусла Диана

Работа написана 28 сентября 2018 года, отредактирована в последний раз 18 февраля 2019 года. Что интересно - всего было 8 версий статьи. Представлена работа на ICLR 2019 в виде орал выступления.

Все авторы связаны с Тюбингенский университетом Германии.

Авторы статьи:

- **Robert Geirhos.** Работает в Тюбингенском университете и в международной исследовательской школе интеллектуальных систем Макса Планка. Научные интересы: Understanding CNNs, Deep Learning, Human Vision, Psychophysics, Robustness. Индекс Хирша 9. Исследует CNN и сравнение работы нейронных сетей с человеческими возможностями. 21 публикация.

- **Patricia Rubisch**. Аспирантка кафедры вычислительной нейронауки Эдинбургского университета. Научные интересы: Synaptic Plasticity, Supervised and unsupervised learning for Spiking Neural Networks, Reservoir Computing. Индекс Хирша 2.
Не очень понятно, чем Патриция конкретно занимается. У нее 3 публикации - две из них посвящены CNN и сравнение работы нейронных сетей с человеческими возможностями.

- **Claudio Michaelis.**. Аспирант в Тюбингенском университете. Научные интересы: Machine Learning, Computer Vision. Индекс Хирша 8. Также занимается сегментацией и One-shot методами.

- **Matthias Bethge..** Пожалуй, самый крутой автор из всех. Приглашенный профессор в Тюбингенском университете и École polytechnique. Научные интересы: Computational Neuroscience, Machine Learning, Vision. Индекс Хирша 61. 301 публикация!!! Самая цитируемая работа: Image style transfer using convolutional neural networks. У него есть несколько работ, которые также связаны с изучением смещения CNN в сторону текстур и синтеза текстур в целом. В том числе эта статья ссылается на некоторые его работы.

- **Felix Wichmann (Феликс Вихманн).** Тюбингенский университет. Научные интересы: Psychophysics, Vision, Visual Perception, Human Vision. Он тоже довольно крутой исследователь. Индекс Хирша 38. 223 публикации. В основном занимается психофизикой и психометрическими функциями.

- **Wieland Brendel (Виланд Брендель)**. Руководитель группы Эмми Нетер в Тюбингенском университете. Научные интересы: Machine Learning, Computer Vision. Индекс Хирша 22. В основном занимается исследованиями различных методов, влияющих на надежность модели.

Связанные работы:

Не могу сказать, что данная работа является прямым продолжением какой-то другой работы. Но один из авторов Matthias Bethge до этого активно занимался изучением синтеза текстур изображения.
И авторы статьи ссылаются на некоторые его работы:

- **Synthesising Dynamic Textures using Convolutional Neural Networks**. Christina M. Funke, Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, 2017
- **Texture and art with deep neural networks.** Leon A Gatys, Alexander S Ecker, Matthias Bethge, 2017

Также до данной работы уже выдвигались утверждения о том, что CNN больше обращают внимание на текстуру изображения. Например в работе
- **Texture and art with deep neural networks.** Leon A Gatys, Alexander S Ecker, Matthias Bethge, 2017

утверждается, что CNN по-прежнему может классифицировать текстурированные изображения, даже если их форма полностью разрушена.

А в работе
- **On the performance of GoogLeNet and AlexNet applied to sketches.** Pedro L. Ballester, R. M. Araújo 2016

наоборот утверждается, что стандартные CNN плохо распознают объекты, где формы объектов сохранены, но отсутствуют все черты текстуры

Также помимо этого некоторые авторы статьи занимались ранее сравнением способности к распознаванию объектов на изображении у нейронных сетей и у человека (психофизикой).
Например авторы ссылаются на статью (авторами которой являются целых 3 автора рассматриваемой нами статьи)

- **Comparing deep neural networks against humans: object recognition when the signal gets weaker.** Robert Geirhos, David H. J. Janssen, Heiko H. Schütt, Jonas Rauber, Matthias Bethge, Felix A. Wichmann, 2017

В которой изучается обобщающая способность человека и нейронной сети и её изменение при изменении изображений.

Но до рассматриваемой нами статьи как такового вывода о том, что CNN действительно ориентирована на текстуру и систематизированного сравнения с человеческим подходом не было.

Цитирования:

Всего цитирований 1228. Прямых продолжений работы нет. Но если кому-то интересна тема, то я бы советовала почитать работы Matthias Bethge.
Самая популярная статья, в которой ссылаются на данную - YOLO4. Это соответственно самое последнее обновление модели YOLO на основе свёрточных слоёв, которая используется для детекции объектов. Статья вышла в 2020 году.
Цитата из статьи: “Style transfer GAN используется для аугментации данных. Такое использование может эффективно уменьшить ориентацию CNN на текстуру.”
В целом среди цитирований в основном были работы про StyleGAN, про улучшение надежности моделей в рамках задачи классификации и про adversarial обучение.

Примеры работ:

- **YOLOv4: Optimal Speed and Accuracy of Object Detection.** Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao, 2020

- **Analyzing and Improving the Image Quality of StyleGAN.** Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, Timo Aila, 2020

- **Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey.** Naveed Akhtar, Ajmal Mian, 2018

- **Adversarial Examples Are Not Bugs, They Are Features.** Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry, 2019.

Вклад и практическое применение:

- Статья помогает лучше понимать работу нейронных сетей на основе CNN. Сделан вывод о том, что смещение на текстуре

можно сдвигать в сторону формы, если обучать модель на подходящем наборе данных. А это в свою очередь повышает точность и надежность модели.
- Также авторы показывают, что ориентированность на структуру в CNN не является свойством архитектуры по умолчанию, а скорее вызвано особенностями обучающих данных.
- В статье был предложен новый набор данных ImageNet, который называется Stylized-ImageNet (SIN), где текстура заменяется случайно выбранным стилем рисования.
- Ну и также авторы утверждают, что ShapeResNet - это первая сеть, которая приблизилась к надежности на уровне человеческой классификации при искажениях, которые не были частью обучающих данных.

Практическое применение результатов статьи под вопросом. Скорее знания, полученные в ней, помогут для дальнейших исследований

Идеи для исследования:

- Провести эксперименты на других наборах данных и с другими, более широкими нейронными сетями, чтобы понять, сохранятся ли результаты.

(Сами авторы говорят, что боялись, что более широкие нейронные сети могут дать другой результат, так что не стали включать их в статью)

- Использовать знания для создания крутых adversarial моделей

Но в целом, как я уже говорила ранее, множество работ по adversarial обучению уже ссылаются на эту статью