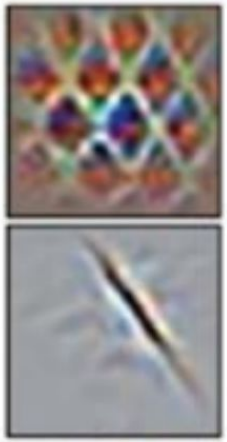
Imagenet-trained CNNs are biased towards texture

Как CNN распознает объекты

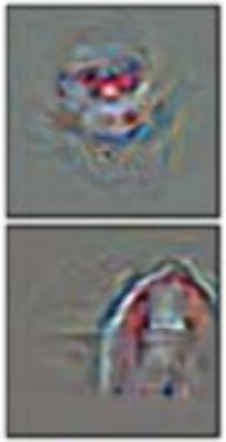
Стандартное объяснение - распознавая их форму

Слой 1



Слой 2

Слой 3

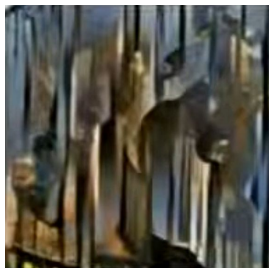


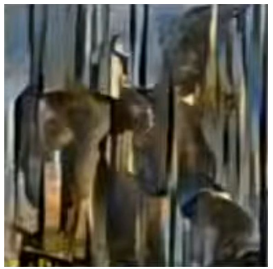
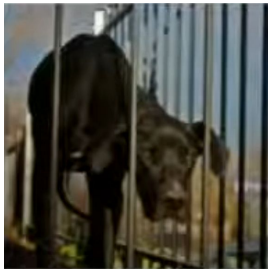
Слой 4

Слой 5

Как CNN распознает объекты

Но при этом распознаются и просто текстурированные объекты, которые лишены формы



Гипотезы формы и текстуры

кот

форма кота

текстура кота

Гипотезы формы и текстуры

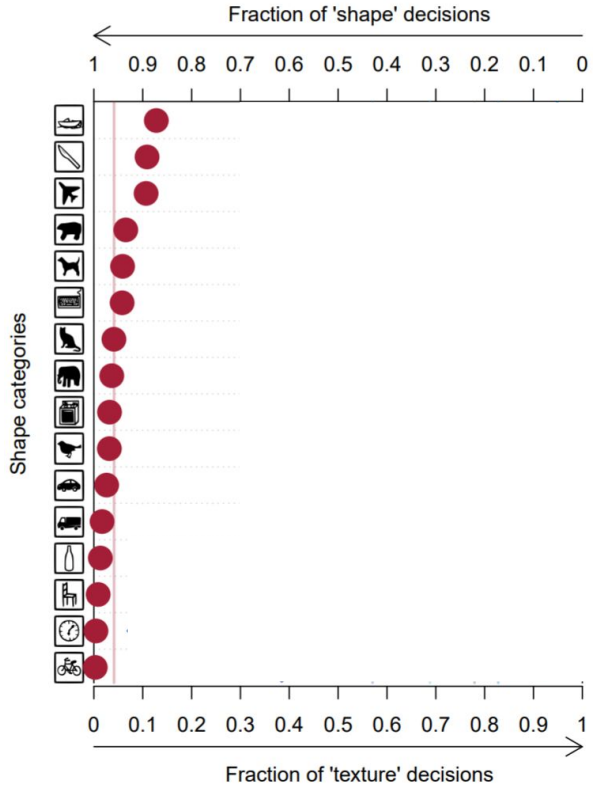
это кот или слон?

форма кота

текстура слона

Гипотезы формы и текстуры

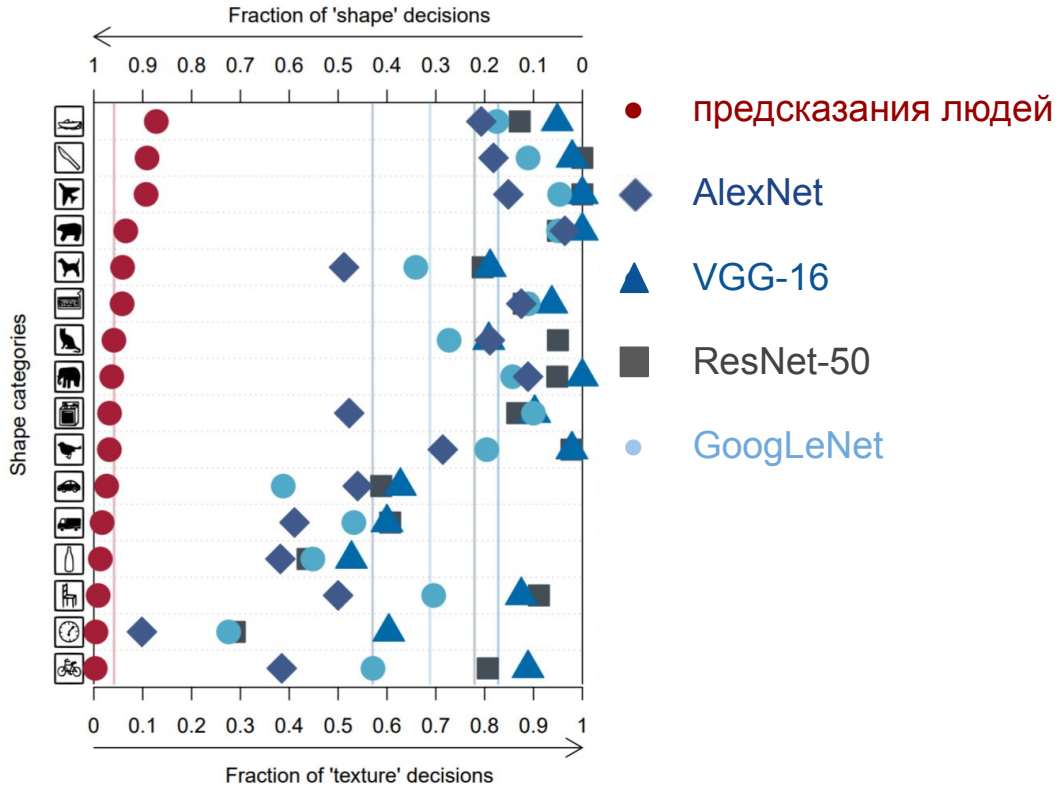
это кот или слон?



- предсказания людей

Гипотезы формы и текстуры

это кот или слон?



Выводы

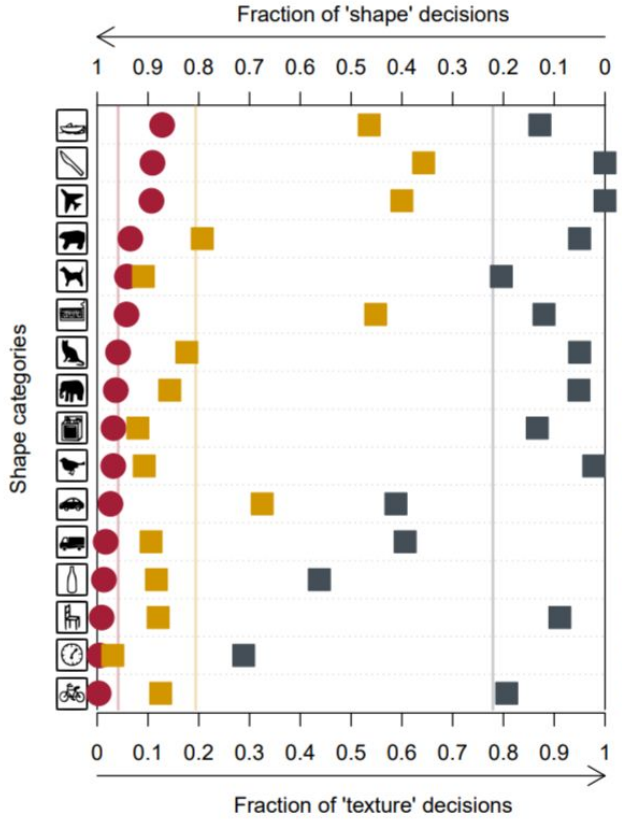
Вывод - CNN в основном опираются на текстуру, а не на форму

Но можем ли мы сместить фокус именно на форму?

Stylized-ImageNet

Stylized-ImageNet

это кот или слон?



предсказания людей

ResNet-50

ResNet-50 на Stylized-ImageNet

Выводы

Вывод - CNN в основном опираются на текстуру, а не на форму

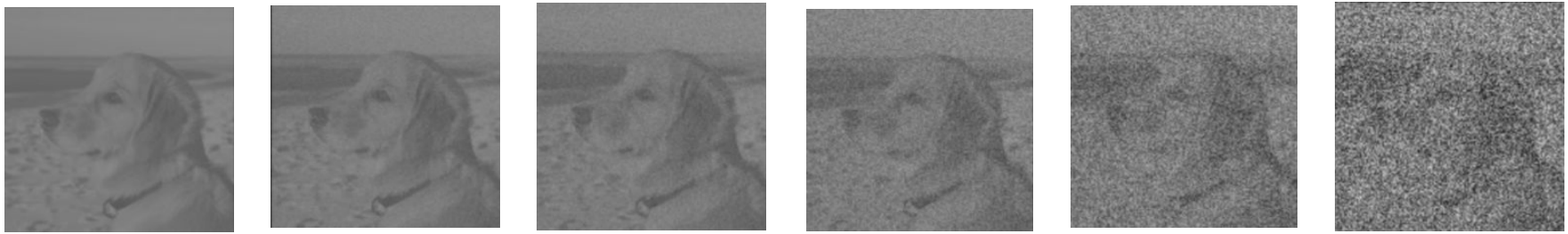
Но на обучение на подходящем наборе данных позволяет сместить фокус именно на форму

Однако дает ли нам это какие-то преимущества?

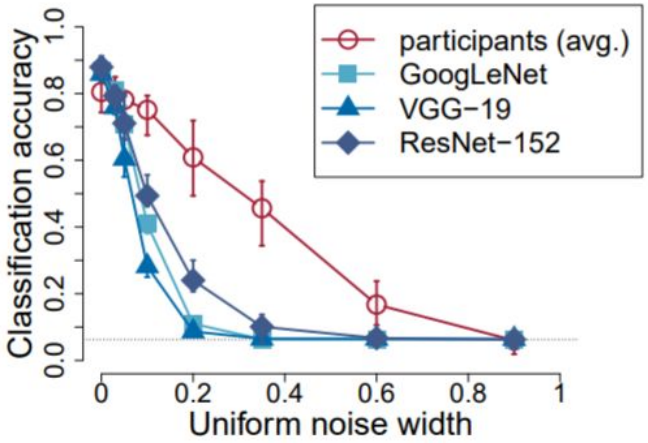
Преимущества формы над текстурой

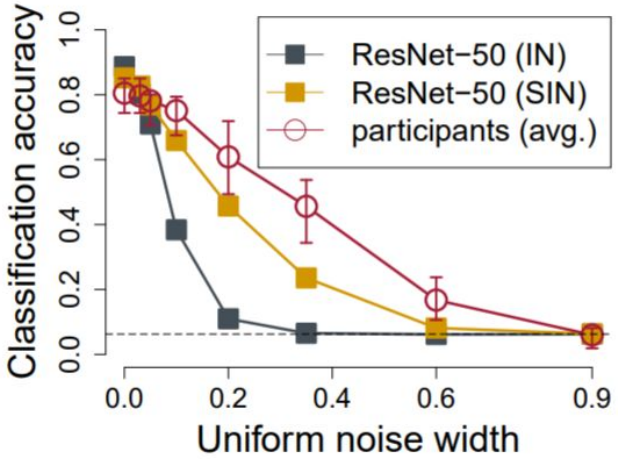
Модель	top-1 IN accuracy (%)	top-5 IN accuracy (%)	Pascal VOC mAP50 (%)	MS COCO mAP50 (%)
ResNet-50 с опорой на текстуру	76.13	92.86	70.7	52.3
ResNet-50 с опорой на форму	**76.72**	**93.28**	**75.1**	**55.2**

Работа с шумом



Работа с шумом





Работа с шумом

Вырабатывается устойчивость к зашумлению, при этом с самим зашумлением модель не сталкивалась во время обучения

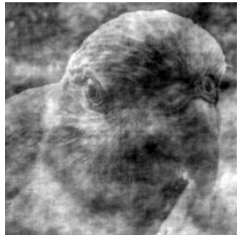
			Noise			Blur			
training	ft	**mCE**	Gaussian	Shot	Impulse	Defocus	Glas	Motion	Zoom
IN (vanilla ResNet-50)	-	76.7	79.8	81.6	82.6	74.7	88.6	78.0	79.9
SIN	-	77.3	71.2	73.3	72.1	88.8	85.0	79.7	90.9
SIN+IN	-	**69.3**	**66.2**	**66.8**	**68.1**	**69.6**	**81.9**	**69.4**	80.5
SIN+IN	IN	73.8	75.9	77.0	77.5	71.7	86.0	74.0	**79.7**

		Weather				Digital			
training	ft	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	JPEG
IN (vanilla ResNet-50)	-	77.8	74.8	66.1	56.6	71.4	84.8	76.9	76.8
SIN	-	71.8	74.4	66.0	79.0	**63.6**	81.1	72.9	89.3
SIN+IN	-	**68.0**	**70.6**	**64.7**	57.8	66.4	**78.2**	**61.9**	**69.7**
SIN+IN	IN	74.5	72.3	66.2	**55.7**	67.6	80.8	75.0	73.2

Выводы

Вывод - CNN в основном опираются на текстуру, а не на форму

Но на обучение на подходящем наборе данных позволяет сместить фокус именно на форму



Благодаря чему вырабатывается устойчивость к зашумлению

Review

Положительные стороны

- Статья хорошо написана: понятный текст и переходы, мало ошибок, красивые понятные графики
- Воспроизводимость: все эксперименты очень подробно описаны в аппендиксе с точностью до мелких деталей; есть открытые репозитории с кодом и с понятным README

- Многочисленные разнообразные эксперименты.


original


greyscale


silhouette


edges


texture


cue conflict

Отрицательные стороны

- Отсутствие теоретической обоснованности.
- Отсутствие пояснения выбора методов/датасетах в экспериментах и объяснения полученных результатов.
- Нет консистентности в экспериментах. Разные эксперименты проводятся на разных архитектурах.

OpenReview

- Высокое качество написания статьи: “surprising”, “inspiring”, “well-written”
- Формализм: придирки к словам “novelty”, “conclusion”; просьбы ввести определения терминов (например, object shape)
- Отсутствие теоретической обоснованности.

Оценка: 8

Уверенность: 4

Практик-исследователь

Сибагатова Софья

Общие сведения

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

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

28 Sept 2018 (modified: 18 Feb 2019) ICLR 2019 Conference Blind Submission Readers:

Tue May 07 02:00 PM -- 02:15 PM (PDT) @ Great Hall AD Oral

***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

Версии статьи

вопрос

1) Were the networks trained to recognize only the 16 classes or the typical 1,000 of imagenet? If the latter, how was a random prediction restricted to the 16 classes?
2) How were these categories selected? They have both distinct appearance and texture in most cases.

ответ

1) All networks were trained on full ImageNet (or full Stylized-ImageNet); they thus recognize 1,000 classes. Concerning the mapping from 1,000 classes to 16 classes, we followed the procedure introduced in [1] (the paper where 16-class-ImageNet was proposed). In order to achieve a fair comparison to the forced-choice paradigm for human observers (who were given a choice of 16 categories on the lab response screen), only those ImageNet categories corresponding to one of the 16 entry-level categories were considered for the network response. The mapping between ImageNet and 16-class-ImageNet categories was achieved via the WordNet hierarchy [2] - e.g. ImageNet category "tabby cat" would be mapped to "cat".

2) We used the 16-class-ImageNet categories introduced in [1]. These are the 16 entry-level categories from MS COCO that have the highest number of ImageNet classes mapped to them via the WordNet hierarchy.

Авторы

Robert Geirhos
University of Tübingen & IMPRS-IS
robert.geirhos@bethgelab.org

Patricia Rubisch
University of Tübingen & U. of Edinburgh
p.rubisch@sms.ed.ac.uk

Claudio Michaelis
University of Tübingen & IMPRS-IS
claudio.michaelis@bethgelab.org

Matthias Bethge*
University of Tübingen
matthias.bethge@bethgelab.org

Felix A. Wichmann*
University of Tübingen
felix.wichmann@uni-tuebingen.de

Wieland Brendel*
University of Tübingen
wieland.brendel@bethgelab.org

*Joint senior authors

Robert Geirhos

I'm a PhD student in deep learning and vision science, working in the labs of Felix Wichmann, Matthias Bethge and Wieland Brendel at the University of Tübingen and the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

I'm in the last year of my PhD (expected graduation date: ~Feb 2022). During summer 2021, I was a Research Intern at FAIR with Ari Morcos.

Why do Deep Neural Networks see the world as they do?

I'm interested in the fascinating area that lies at the intersection of Deep Learning and Visual Perception.

I want to understand why Deep Neural Networks (DNNs) see the world as they do. Visual perception is a process of inferring—typically reasonably accurate—hypotheses about the world. But what are the hypotheses and assumptions that DNNs make? Answering this question involves understanding the limits of their abilities (when do machines fail, and why?), the biases that they incorporate (e.g. texture bias, a reliance on local features) and the underlying pattern behind their success (such as shortcut learning, or "cheating").

When comparing DNNs to human perception, I develop quantitative methods to identify areas where DNNs are still falling short of the remarkably robust, flexible and general representations of the human visual system and in a second step seek to overcome these differences. Ultimately, I am convinced that understanding why DNNs see the world as they do holds the key towards making them more interpretable, robust and reliable: Once we have understood DNNs, we can build DNNs that truly "understand".

Robert Geirhos

I'm a PhD student in deep learning and vision science, working in the labs of Felix Wichmann, Matthias Bethge and Wieland Brendel at the University of Tübingen and the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

I'm in the last year of my PhD (expected graduation date: ~Feb 2022). During summer 2021, I was a Research Intern at FAIR with Ari Morcos.

Wichmann, F. A., Janssen, D. H., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., & Bethge, M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging, Human Vision and Electronic Imaging, 2017*(14), 36–45.

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*.

Geirhos, R., Medina Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems 31* (pp. 7548–7560).

[ORAL] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations*.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems 33*.

Huber, L. S., Geirhos, R., & Wichmann, F. A. (2021). Out-of-distribution robustness: Limited image exposure of a four-year-old is enough to outperform ResNet-50. *NeurIPS workshop on Shared Visual Representations in Human & Machine Intelligence*.

Matthias Bethge

I did my undergraduate studies in physics and started working in computational neuroscience when I joined the MPI for dynamics and self-organization for my diploma project. Since then my research aims at understanding perceptual inference and self-organized collective information processing in distributed systems---two puzzling phenomena that contribute much to our fascination about living systems.

General principles are important but at the same time these principles need to be grounded in reality. Therefore, a large part of my research focuses on the mammalian visual system working closely together with experimentalists (Andreas Tolias , Thomas Euler , Felix Wichmann). I also work on neural coding in other sensory systems (collaborations with Cornelius Schwarz)

http://bethgelab.org/people/matthias/

Matthias Bethge

Comparing DNNs with human/mouse

F. A. Wichmann, D. H. Janssen, R. Geirhos, G. Aguilar, H. H. Schütt, M. Maertens, and M. Bethge
Methods and measurements to compare men against machines
Electronic Imaging, ***2017****(14), 36-45, 2017*

R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann
Generalisation in humans and deep neural networks
Advances in Neural Information Processing Systems 31, 2018

S. A. Cadena, F. H. Sinz, T. Muhammad, E. Froudarakis, E. Cobos, E. Y. Walker, J. Reimer, M. Bethge, *et al.*
How well do deep neural networks trained on object recognition characterize the mouse visual system?
NeurIPS Neuro AI Workshop, 2019

http://bethgelab.org/people/matthias/

Matthias Bethge

Textures/Style transfer

L. A. Gatys, A. S. Ecker, and M. Bethge **Texture Synthesis Using Convolutional Neural Networks** *Advances in Neural Information Processing Systems 28, 2015*
L. A. Gatys, A. S. Ecker, and M. Bethge **Image Style Transfer Using Convolutional Neural Networks** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016*
I. Ustyuzhaninov*, W. Brendel*, L. Gatys, and M. Bethge **What does it take to generate natural textures?** *International Conference on Learning Representations, 2017*
L. A. Gatys, A. S. Ecker, and M. Bethge **Texture and art with deep neural networks** *Current Opinion in Neurobiology, **46**, 178-186, 2017*

http://bethgelab.org/people/matthias/

Felix A. Wichmann

Felix Wichmann

Group Leader

+49 7071 29 70421
Room no. 10-5/A24
felix.wichmann[at]uni-tuebingen.de
Office Hours

Research:

My laboratory investigates human perception combining psychophysical experiments with computational modelling. Currently we have four research foci: First, to improve our image-based model of early spatial vision. Second, to connect early spatial vision with mid-level vision: perceived lightness, brightness and contrast in relation to surface reflectance and illumination in images of real scenes. Third, we investigate differences and similarities between deep convolutional neural networks and human object recognition. Fourth, we explore connections between causality from a perceptual as well as a machine learning perspective.

Felix Wichmann received his B.A. (1994) and DPhil (1999) in Experimental Psychology from the University of Oxford. After post-doctoral research at the University of Leuven (2000-2001), he worked as a research scientist in the Empirical Inference Department at the Max Planck Institute for Biological Cybernetics in Tübingen (2001-2007). From 2007 to 2011 he was Associate Professor (W2) at the Technical University of Berlin and since 2011 he is Full Professor (W3) at the Eberhard Karls Universität Tübingen. He serves on the editorial board of the *Journal of Vision*.

Felix A. Wichmann

Comparing DNNs with human

ei Wichmann, F. A., Janssen, D. H. J., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., Bethge, M. **Methods and measurements to compare men against machines** *Electronic Imaging*, pages: 36-45(10), 2017 (article)

ei Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H., Bethge, M., Wichmann, F. A. **Generalisation in humans and deep neural networks** *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages: 7549-7561, (Editors: S. Bengio and H. Wallach and H. Larochelle and K. Grauman and N. Cesa-Bianchi and R. Garnett), Curran Associates, Inc., 32nd Annual Conference on Neural Information Processing Systems, December 2018 (conference)

Textures

ei Wallis, T. S. A., Funke, C. M., Ecker, A. S., Gatys, L. A., Wichmann, F. A., Bethge, M. **A parametric texture model based on deep convolutional features closely matches texture appearance for humans** *Journal of Vision*, 17(12):5, 2017 (article)

Источники вдохновения

Форма

- Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. PLoS Computational Biology, 12(4):e1004896, 2016.

  “CNNs implicitly learn representations of shape that reflect human shape perception”

- Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. arXiv preprint arXiv:1706.08606, 2017.

  “CNNs develop a so-called “shape bias” just like children, i.e. that object shape is more important than colour”

Текстура

- Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Texture and art with deep neural networks. Current Opinion in Neurobiology, 46:178–186, 2017.

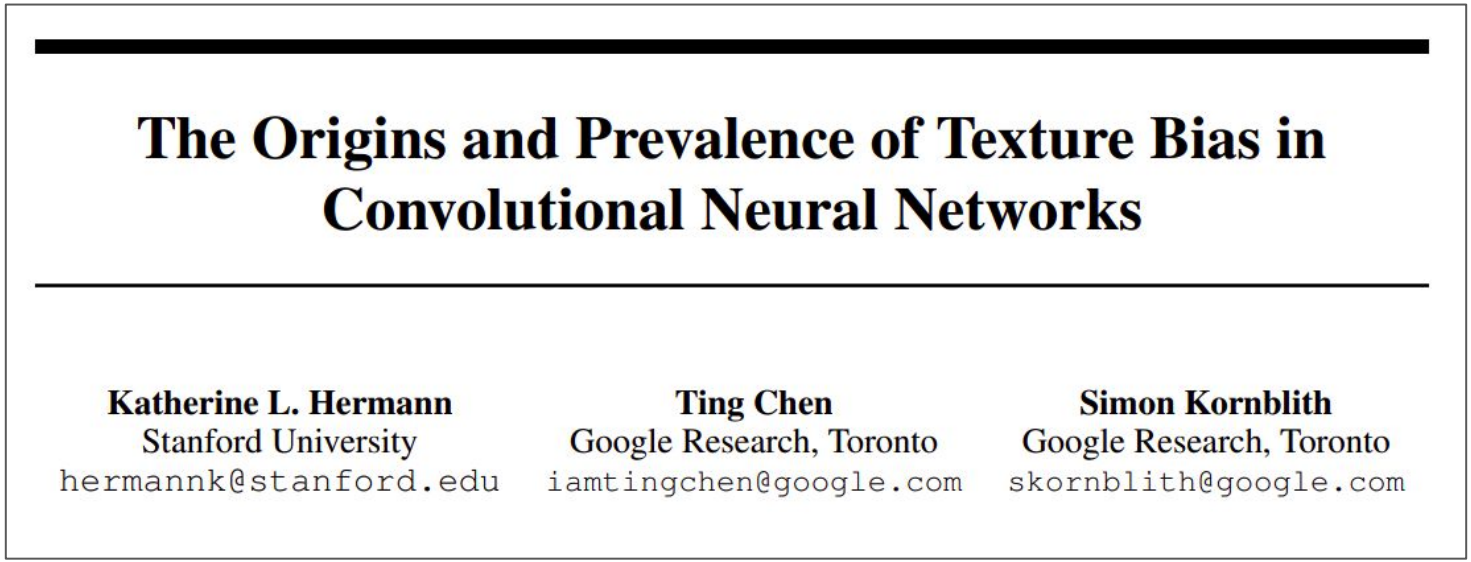
  “CNNs can still classify texturised images perfectly well, even if the global shape structure is completely destroyed”

- Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In International Conference on Learning Representations, 2019.

  “CNNs with explicitly constrained receptive field sizes throughout all layers are able to reach surprisingly high accuracies on ImageNet, even though this effectively limits a model to recognising small local patches rather than integrating object parts for shape recognition”

Цитирования

Всего: 1127



The Origins and Prevalence of Texture Bias in Convolutional Neural Networks

Katherine L. Hermann
Stanford University
hermannk@stanford.edu

Ting Chen
Google Research, Toronto
iamtingchen@google.com

Simon Kornblith
Google Research, Toronto
skornblith@google.com

Прямые конкуренты не были найдены

Дополнительные исследования и применения

Дополнительные исследования: авторы статьи используют рандомные текстуры для формирования набора данных. Можно ли улучшить качество, подбирая для каждого изображения текстуры, наиболее ухудшающие качество (использовать своего сорта adversarial атаку)?

Применение в индустриальных приложениях: распознавание человека с реалистичным гримом/в реалистичном костюме (например, животного)