

# Глубокое обучение

Дмитрий Никулин

26 мая 2021 г.

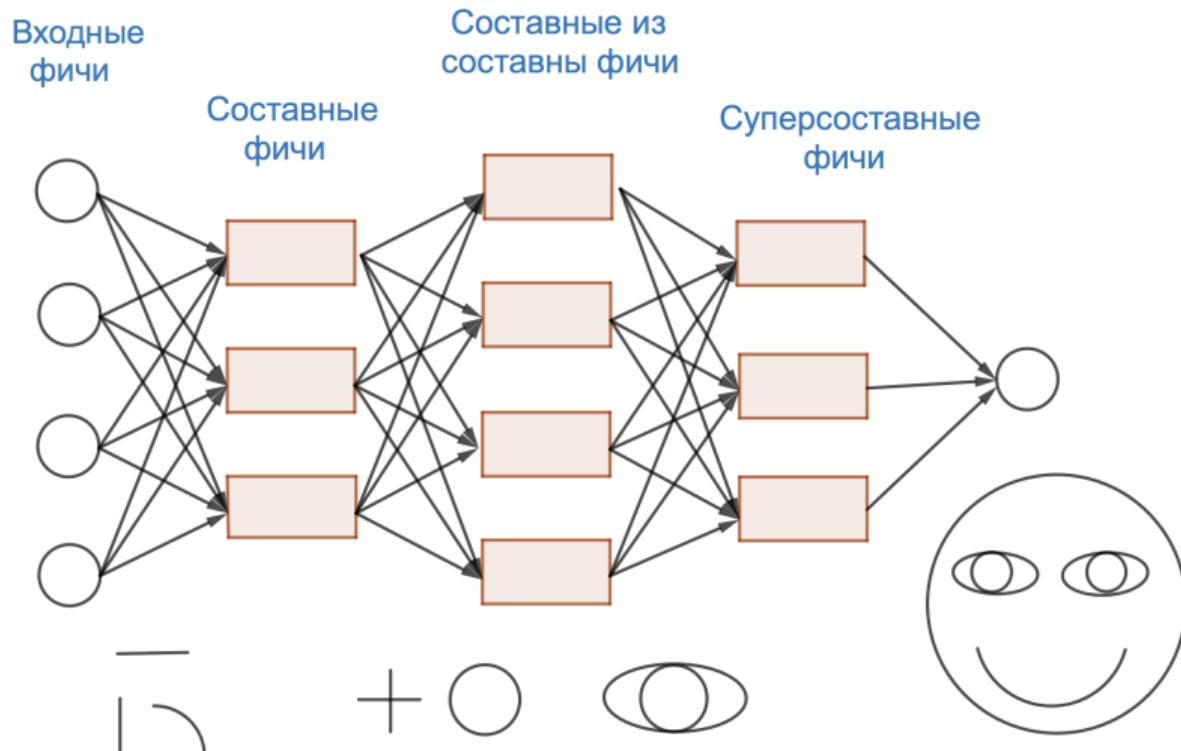
**Неделя 9:** Интерпретируемость. DeepDream

# Agenda

- Что выучивают нейросети
- Обзор некоторых статей
- DeepDream

# Что выучивают нейросети

# Что выучивают нейросети



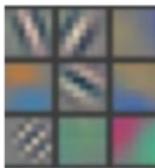
# Visualizing and Understanding Convolutional Networks

Статья далёкого 2013 года:

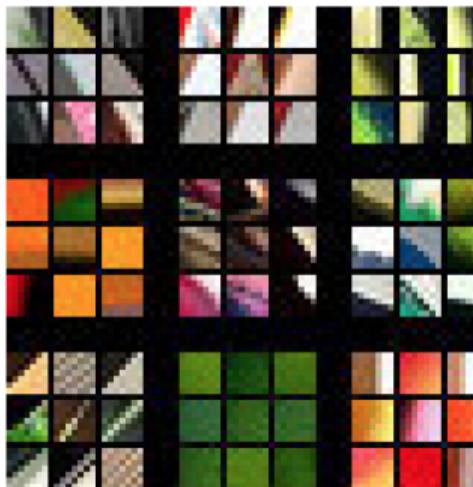
- Берём обученную классификационную свёрточную нейронку
- Прогоняем через неё датасет картинок, следим, когда и какие нейроны при этом активируются наиболее сильно
- Для каждой такой сильной активации рисуем:
  - соответствующий патч в картинке из датасета
  - градиент активации по этому патчу (то есть бэкпропаем из активации в пиксели картинки)

<https://arxiv.org/pdf/1311.2901.pdf>

# Visualizing and Understanding Convolutional Networks

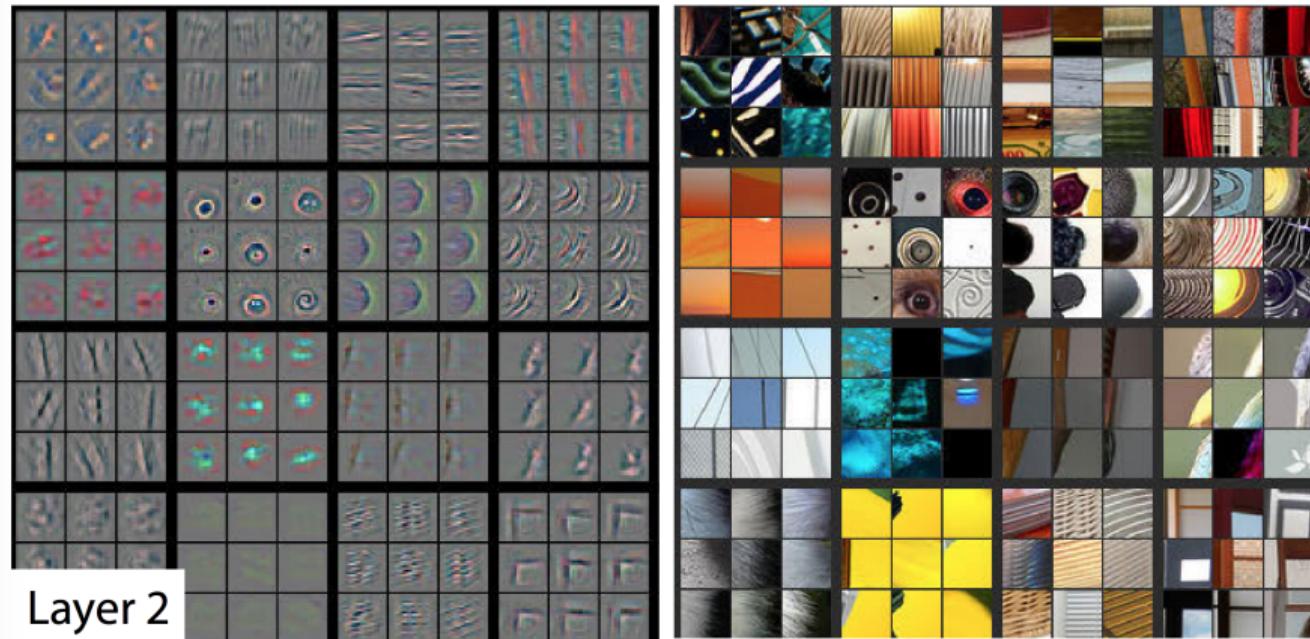


Layer 1



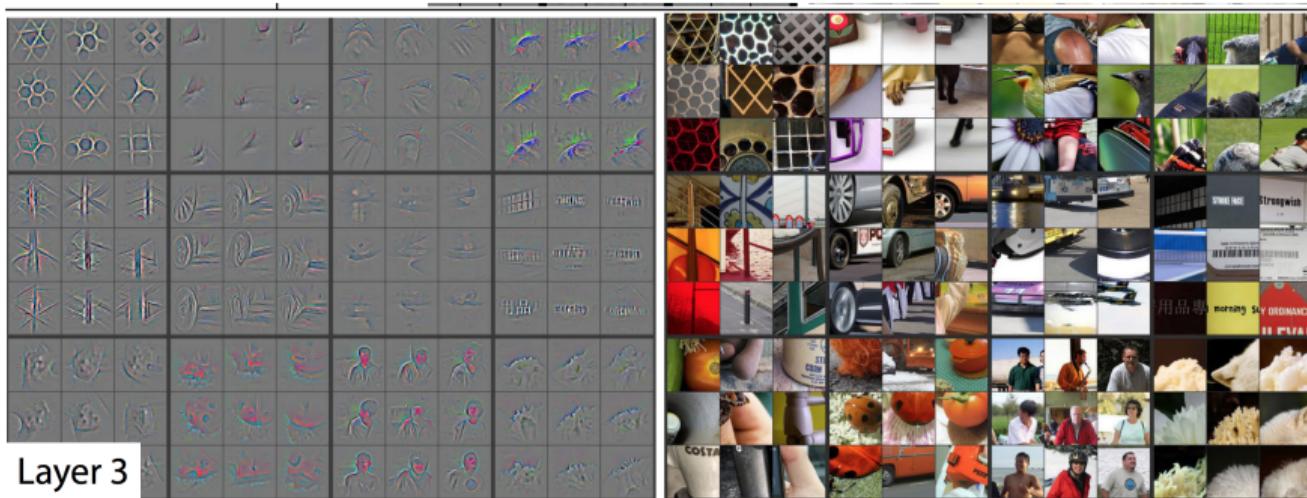
<https://arxiv.org/pdf/1311.2901.pdf>

# Visualizing and Understanding Convolutional Networks



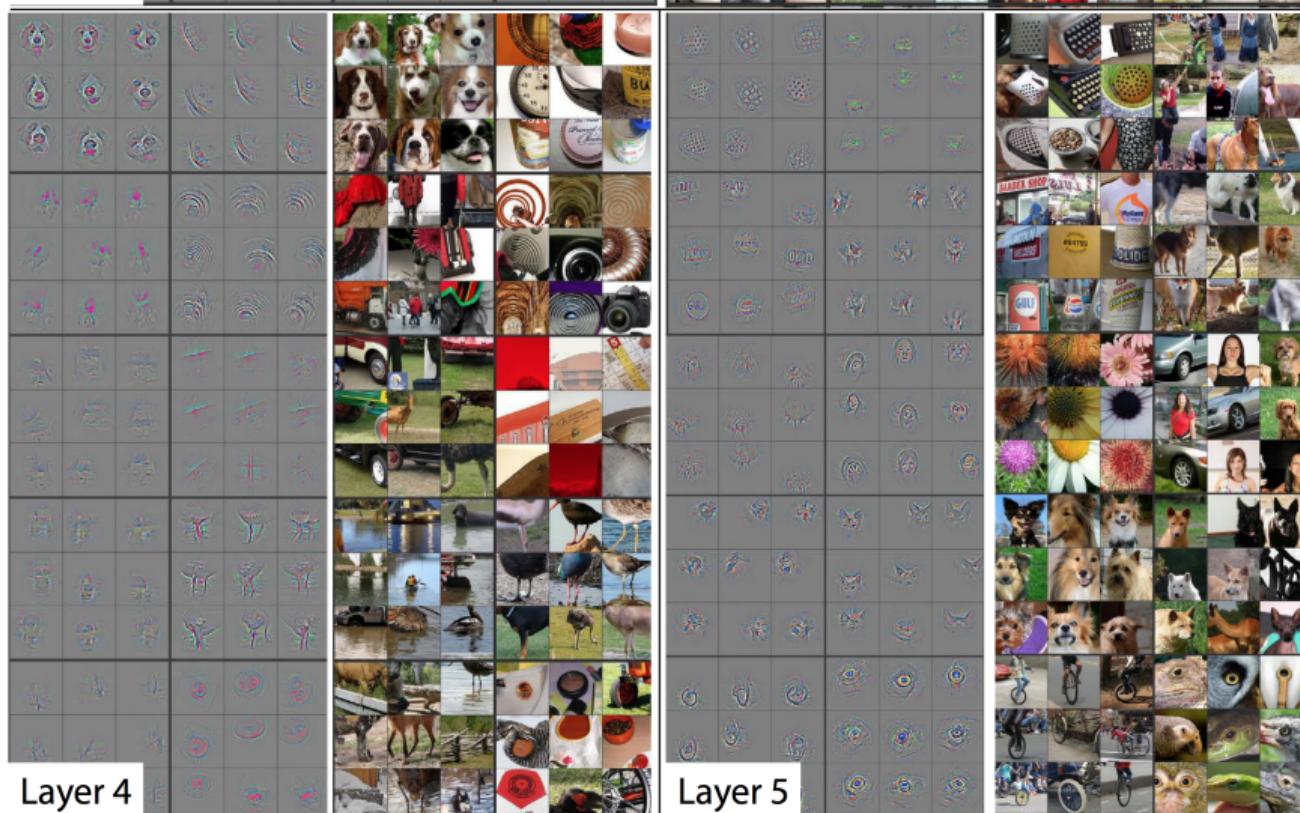
<https://arxiv.org/pdf/1311.2901.pdf>

# Visualizing and Understanding Convolutional Networks



<https://arxiv.org/pdf/1311.2901.pdf>

# Visualizing and Understanding Convolutional Networks



# Визуализация через оптимизацию

- Генерируем случайное изображение (например, с помощью `torch.rand`) и оптимизируем его таким образом, чтобы некоторый нейрон внутри сетки активировался как можно сильнее
- Для этого надо бэкпропнуть из активации в пиксели изображения и сделать шаг градиентного спуска
- На изображении будет постепенно прорисовываться шаблон, который возбуждает соответствующий нейрон

# Deep Inside Convolutional Networks

Здесь картинки максимизируют предсказываемые вероятности классов...

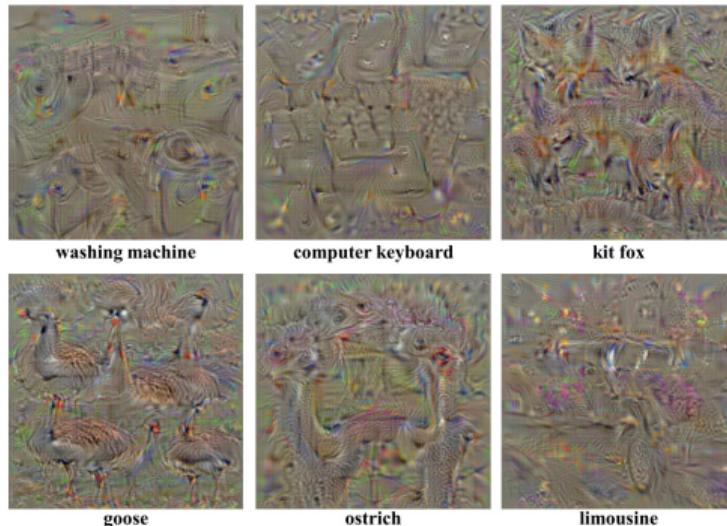


Figure 1: Numerically computed images, illustrating the class appearance models, learnt by a ConvNet, trained on ILSVRC-2013. Note how different aspects of class appearance are captured in a single image. Better viewed in colour.

<https://arxiv.org/pdf/1312.6034.pdf>

# Deep Inside Convolutional Networks

...а тут нарисован просто градиент вероятности класса по пикселям картинок.

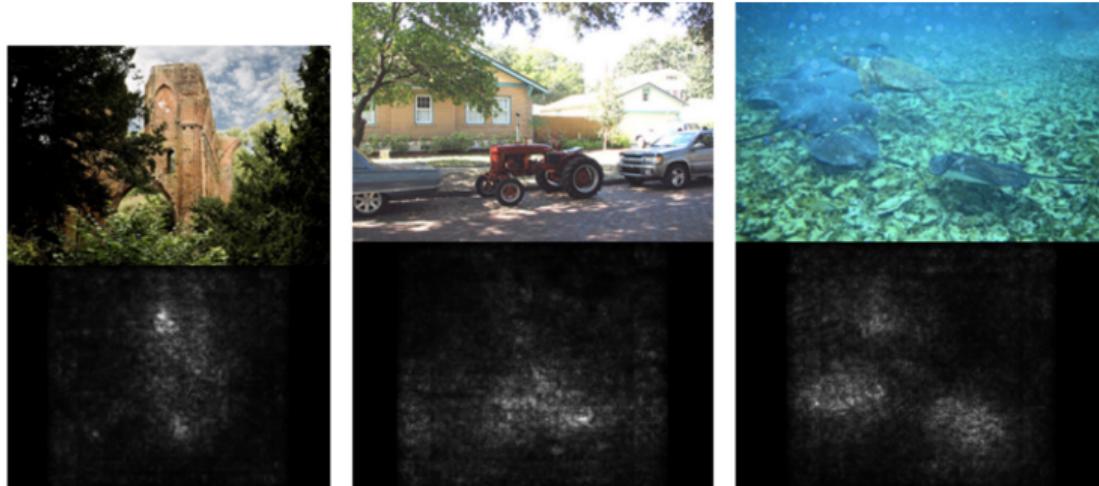
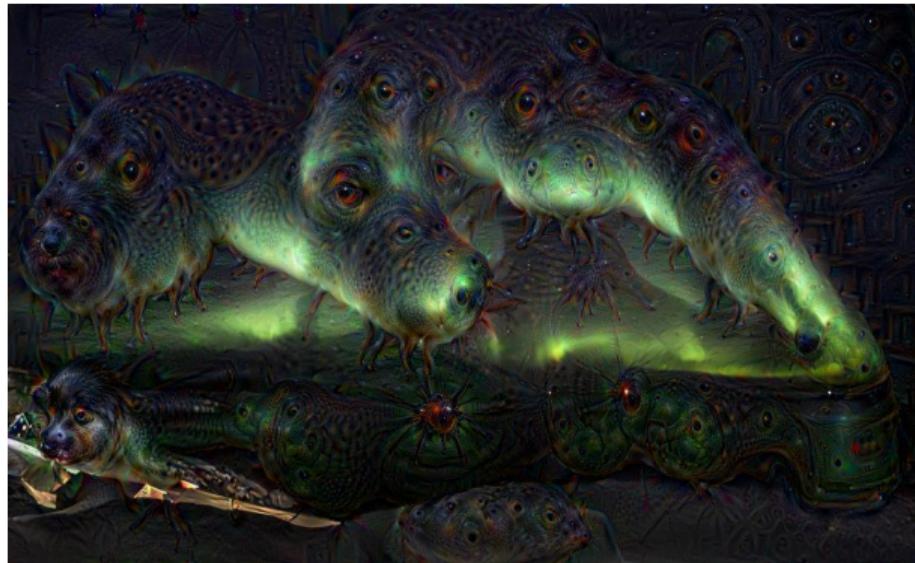


Figure 2: **Image-specific class saliency maps for the top-1 predicted class in ILSVRC-2013 test images.** The maps were extracted using a single back-propagation pass through a classification ConvNet. No additional annotation (except for the image labels) was used in training.

<https://arxiv.org/pdf/1312.6034.pdf>

# DeepDream

Если процедуру начать не со случайной картинки, а с чего-то осмысленного, то шаблон отрисуется поверх такой картинки. Несколько костылей спустя получается вот такая наркомания:



Картинка: <https://nplus1.ru/material/2015/07/13/use>

Анонс: <https://ai.googleblog.com/2015/07/deepdream-code-example-for-visualizing.html>

# Feature Visualization

- Инициализируем картинку случайным шумом
- Градиентным спуском оптимизируем её так, чтобы она максимально активировала нейрон / канал / слой
- На каждом шаге слегка аугментируем картинку (двигаем, крутим, масштабируем, добавляем паддинг, обрезаем)
- (Опционально) делаем градиентный спуск не на пикселях картинки, а на её преобразовании Фурье

<https://distill.pub/2017/feature-visualization/>

# Feature Visualization

Different **optimization objectives** show what different parts of a network are looking for.

**n** layer index

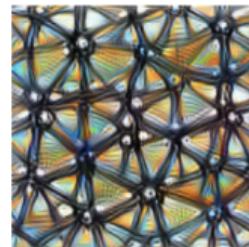
**x,y** spatial position

**z** channel index

**k** class index



Neuron

$$\text{layer}_n[x, y, z]$$


Channel

$$\text{layer}_n[:, :, z]$$


Layer/DeepDream

$$\text{layer}_n[:, :, :]^2$$


Class Logits

$$\text{pre\_softmax}[k]$$


Class Probability

$$\text{softmax}[k]$$

<https://distill.pub/2017/feature-visualization/>

# The Building Blocks of Interpretability

- Можно склеивать друг с другом разные методы для интерпретирования моделей и делать из них интерактивные интерфейсы
- Например, можно нарисовать визуализации нейронов из разных слоёв в соответствии с их пространственным расположением, и при наведении курсора на нейрон показывать, насколько сильно этот нейрон влияет на последующие нейроны

<https://distill.pub/2018/building-blocks/>

# The Building Blocks of Interpretability



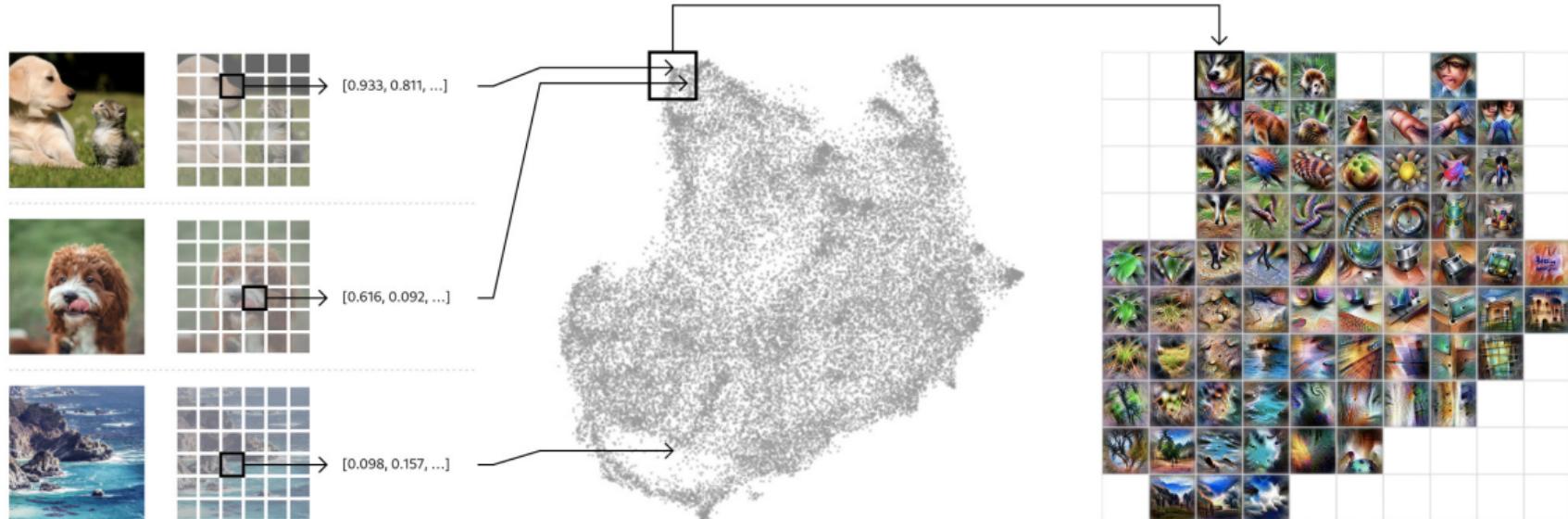
<https://distill.pub/2018/building-blocks/>

# Exploring Neural Networks with Activation Atlases

- Метод для визуализации того, как один слой внутри нейронки воспринимает целый датасет
- Прогоняем через модель весь датасет ( $N$  картинок), для каждой картинки из активаций соответствующего слоя (размером  $C \times H \times W$ ) выбираем рандомно один вектор длиной  $C$
- Полученные  $N C$ -мерных векторов как-нибудь рисуем на плоскости (PCA / t-SNE / UMAP / etc)
- Плоскость режем на квадраты, в каждом квадрате усредняем все попавшие туда векторы, и получившийся средний вектор активаций визуализируем

<https://distill.pub/2019/activation-atlas/>

# Exploring Neural Networks with Activation Atlases



A randomized set of one million images is fed through the network, collecting one random spatial activation per image.

The activations are fed through UMAP to reduce them to two dimensions. They are then plotted, with similar activations placed near each other.

We then draw a grid and average the activations that fall within a cell and run feature inversion on the averaged activation. We also optionally size the grid cells according to the density of the number of activations that are averaged within.

# Пишем код