Национальный исследовательский
Нижегородский государственный университет им. Н.И. Лобачевского
Институт информационных технологий, математики и механики

Классификация изображений с большим числом категорий с использованием методов глубокого обучения

Кустикова В.Д.,
к.т.н., доцент каф. МОСТ ИИТММ
ННГУ им. Н.И. Лобачевского

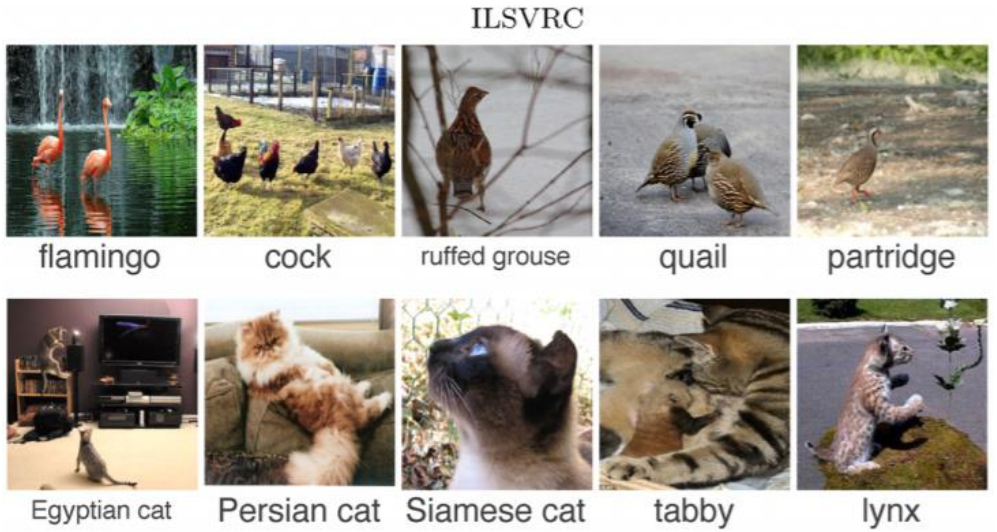
Содержание

- Постановка задачи классификации изображений
- ImageNet Large Scale Visual Recognition Challenge и набор данных ImageNet
- Обзор глубоких моделей для классификации изображений на наборе данных ImageNet
- Сравнение качества классификации и сложности глубоких моделей на наборе данных ImageNet
- Заключение

ПОСТАНОВКА ЗАДАЧИ КЛАССИФИКАЦИИ ИЗОБРАЖЕНИЙ

Постановка задачи (1)

- Задача классификации изображений состоит в том, чтобы поставить в соответствие изображению класс объектов, содержащихся на этом изображении
- Примеры изображений и соответствующих им классов:



* Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A.C., Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge // International Journal of Computer Vision, 2015.

Постановка задачи (2)

- Исходное изображение представлено набором интенсивностей пикселей $I = \left(I_{ij}^{k}\right)_{\substack{0 \le i < w \\ 0 \le j < h \\ 0 \le k < 3}}$, где w и h – ширина и высота изображения, k – количество каналов
- Определено множество допустимых классов объектов на изображении $C = \{0, 1, \ldots N - 1\}$, множество идентификаторов классов однозначно соответствует множеству названий классов
- ***Задача классификации изображений*** состоит в том, чтобы каждому изображению поставить в соответствие класс, которому оно принадлежит

$$\varphi : I \to C$$

IMAGENET LARGE SCALE VISUAL RECOGNITION CHALLENGE И НАБОР ДАННЫХ IMAGENET

ImageNet Large Scale Visual Recognition Challenge

- ImageNet – открытый набор данных, предоставляемый в рамках конкурса по классификации изображений с большим числом категорий и детектированию объектов на изображениях ILSVRC (ImageNet Large Scale Visual Recognition Challenge)
- С 2010 по 2017 годы базировался на [http://www.image-net.org], с 2017 года переехал на платформу Kaggle

* Russakovsky O., et al. ImageNet Large Scale Visual Recognition Challenge. – 2015. – [https://arxiv.org/pdf/1409.0575.pdf].

Набор данных ImageNet

- Состоит из 14 197 122 изображений, принадлежащих 21 841 категориям из иерархии WordNet*
- Иерархия содержит 27 категорий объектов высокого уровня
- 1 034 908 изображений содержат разметку для задачи детектирования объектов (размечены окаймляющие прямоугольники для объектов), эти данные используются и для задачи классификации
- Изображения собраны из различных источников, создатели набора данных не имеют авторских прав на изображения

* Jia D., Dong W., Socher R., Li L.-J., Li K., Li F.-F. ImageNet: A large-scale hierarchical image database // In the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2009. – P. 248-255. – [https://ieeexplore.ieee.org/document/5206848].

Набор данных ImageNet для классификации изображений по данным конкурса ILSVRC 2012

- 1 000 категорий изображений
- Минимальное разрешение – 75x56 пикселей
- Максимальное разрешение – 4288x2848 пикселей
- Размер тренировочной выборки – 1 200 000 изображений
- Размер валидационной выборки – 50 000 изображений
- Размер тестовой выборки – 150 000 изображений

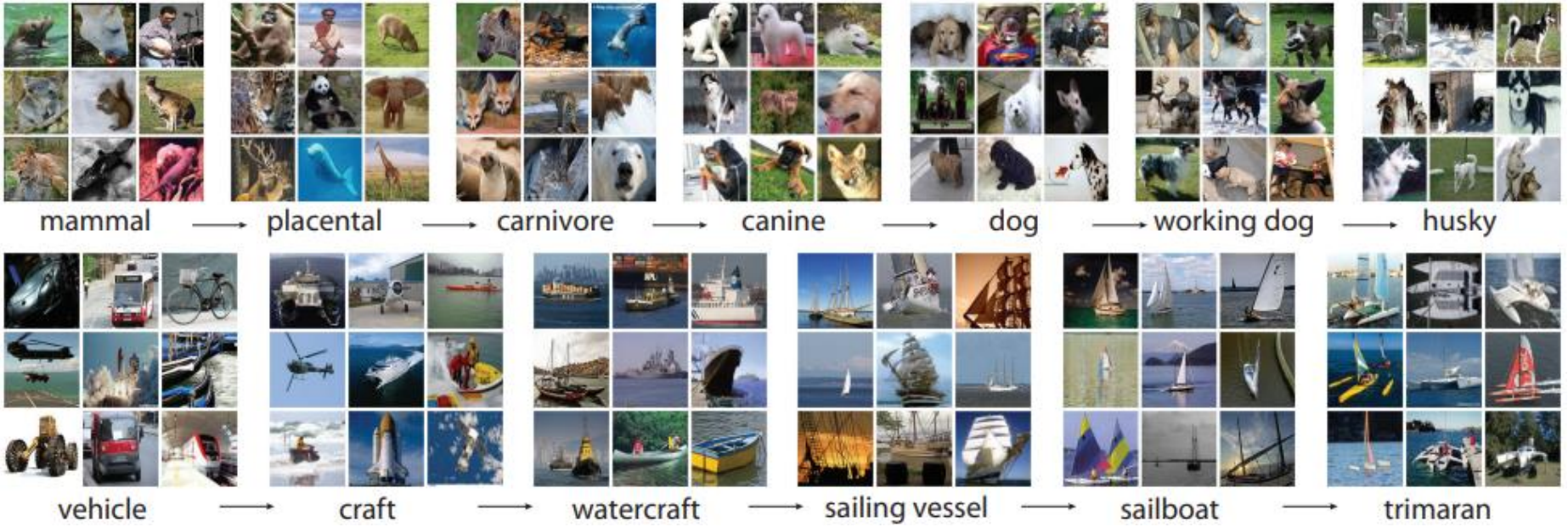
Иерархия классов WordNet (1)

- WordNet* – большая лексическая база слов английского языка
- Основным отношением между словами в WordNet является ***синонимия***
- ***Синонимы*** – слова, близкие по значению, взаимозаменяемые во многих контекстах
- Синонимы сгруппированы в ***неупорядоченные множества*** (synset)
- Группы синонимов связаны следующими отношениями:
  - ***Гиперонимия*** (гипонимия) – связь общего и частного (например, кровать – это мебель)
  - ***Меронимия*** (партонимия) – связь между объектами и их частями (например, «двигатель» – мероним по отношению к термину «автомобиль»)

* WordNet. A Lexical Database for English [https://wordnet.princeton.edu].

Иерархия классов WordNet (2)

- WordNet* содержит около 80 000 существительных
- Цель разработки набора данных ImageNet для каждого множества синонимов подобрать 500-1000 изображений



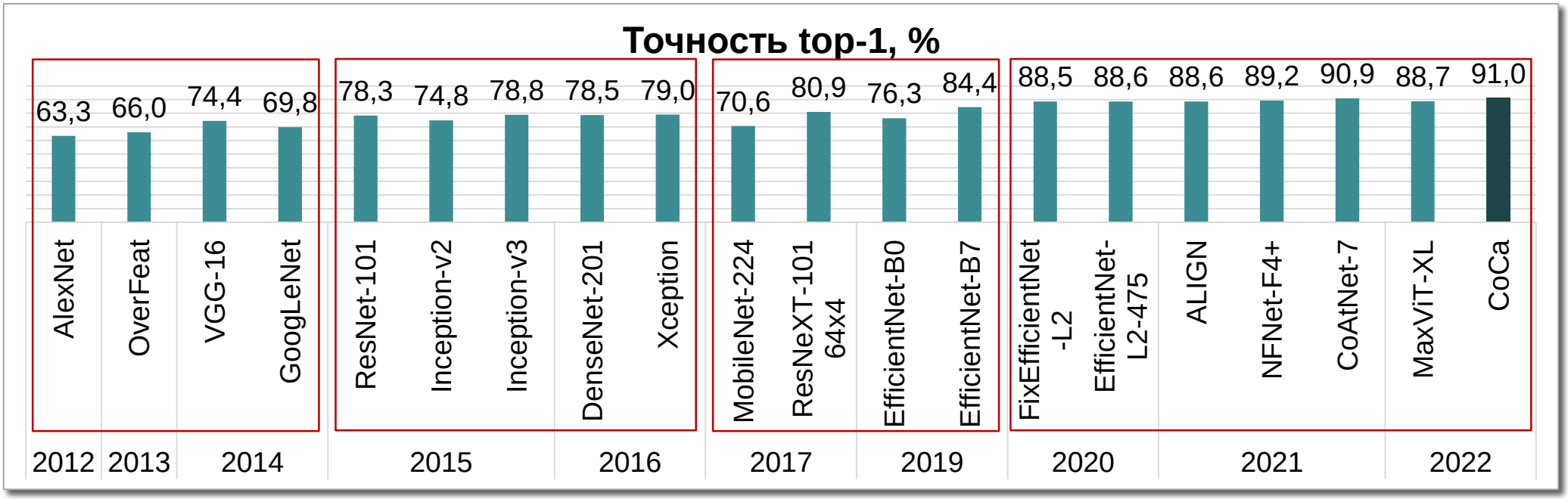
* WordNet. A Lexical Database for English [https://wordnet.princeton.edu].

** Ye T. Visual Object Detection from Lifelogs using Visual Non-lifelog Data. – 2018. – [https://www.researchgate.net/publication/324797660_Visual_Object_Detection_from_Lifelogs_using_Visual_Non-lifelog_Data].

ГЛУБОКИЕ МОДЕЛИ ДЛЯ КЛАССИФИКАЦИИ ИЗОБРАЖЕНИЙ НА НАБОРЕ ДАННЫХ IMAGENET

Обзор моделей. Точность

- Изменение точности top-1 на наборе данных ImageNet для избранных моделей:



Увеличение глубины | **Проблема деградации** | **Снижение сложности (число параметров + число операций)** | **Применение трансформеров**

- ***За 10 лет точность выросла на ~28%***

* Image Classification on ImageNet [https://paperswithcode.com/sota/image-classification-on-imagenet].

Обзор моделей (1)

Наращивание глубины

- ***AlexNet (2012)***
  - Krizhevsky A., Sutskever I., Hinton G.E. ImageNet Classification with Deep Convolutional Neural Networks // Advances in neural information processing systems. – 2012. – [http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf].
- ***OverFeat (2013)***
  - Sermanet P., Eigen D., Zhang X., Mathieu M., Fergus R., LeCun Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. – 2013. – [https://arxiv.org/pdf/1312.6229.pdf].
- ***VGG-16, VGG-19, GoogLeNet (2014)***
  - Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. – 2014. – [https://arxiv.org/pdf/1409.1556.pdf].
  - Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., Rabinovich A. Going Deeper with Convolutions. – 2014. – [https://arxiv.org/pdf/1409.4842.pdf].

Обзор моделей (2)

Решение проблемы деградации модели

- ***ResNet-*(50, 101, 152), Inception-v*(2,3) (2015)***
  - He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. – 2015. – [https://arxiv.org/pdf/1512.03385.pdf].
  - Ioffe S., Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. – 2015. – [https://arxiv.org/pdf/1502.03167.pdf].
  - Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. Rethinking the Inception Architecture for Computer Vision. – 2015. – [https://arxiv.org/pdf/1512.00567.pdf], [https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.pdf] (опубликованная версия).
- ***DenseNet-* (121, 169, 201, 264), Xception (2016)***
  - Huang G., Liu Z., Maaten L., Weinberger K.Q. Densely Connected Convolutional Networks. – 2016. – [https://arxiv.org/pdf/1608.06993.pdf].
  - Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. – 2016. – [https://arxiv.org/pdf/1610.02357.pdf].

Снижение количества параметров модели

Обзор моделей (3)

Снижение сложности модели

- ***MobileNet, ResNeXT-* (2017)***
  - Howard A.G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. – 2017. – [https://arxiv.org/pdf/1704.04861.pdf].
  - Xie S., Girshick R., Dollar P., Tu Z., He K. Aggregated Residual Transformations for Deep Neural Networks. – 2017. – [https://arxiv.org/pdf/1611.05431v2.pdf].
- ***MobileNetV2 (2018)***
  - Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. – 2018. – [https://arxiv.org/pdf/1801.04381.pdf].
- ***EfficientNet-* (B0,…,B7), SENet (2019)***
  - Tan M., Le Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. – 2019. – [https://arxiv.org/pdf/1905.11946.pdf].
  - Hu J., Shen L., Albanie S., Sun G., Wu E. Squeeze-and-Excitation Networks. – 2019. – [https://arxiv.org/pdf/1709.01507.pdf].

Обзор моделей (4)

Применение трансформеров

- ***FixEfficientNet-L2, EfficientNet-L2-475 (2020, 2021)***
  - Touvron H., Vedaldi A., Douze M., Jégou H. Fixing the train-test resolution discrepancy: FixEfficientNet. – 2020. – [https://arxiv.org/pdf/2003.08237v5.pdf].
  - Foret P., Kleiner A., Mobahi H., Neyshabur B. Sharpness-Aware Minimization for Efficiently Improving Generalization. – 2021. – [https://openreview.net/pdf?id=6Tm1mposlrM].
- ***ALIGN, NFNet-F4+, CoAtNet-7 (2021)***
  - Jia Ch., et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. – 2021. – [https://arxiv.org/pdf/2102.05918v2.pdf].
  - Brock A., De S., Smith S.L., Simonyan K. High-Performance Large-Scale Image Recognition Without Normalization. – 2021. – [https://arxiv.org/pdf/2102.06171v1.pdf].
  - Dai Z., Liu H., Le Q.V., Tan M. CoAtNet: Marrying Convolution and Attention for All Data Sizes. – 2021. – [https://proceedings.neurips.cc//paper/2021/file/20568692db622456cc42a2e853ca21f8-Paper.pdf].

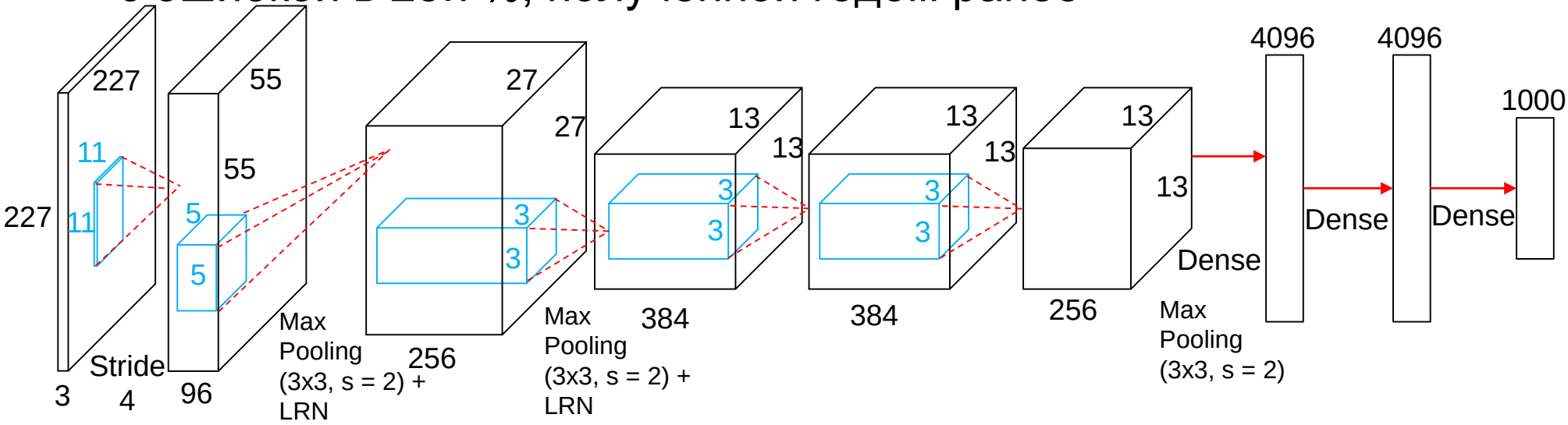
Обзор моделей (5)

- ***MaxViT-XL, CoCa (2022)***
  - Tu Zh., Talebi H., Zhang H., Yang F., Milanfar P., Bovik A., Li Y. MaxViT: Multi-Axis Vision Transformer. – 2022. – [https://arxiv.org/pdf/2204.01697v1.pdf].
  - Yu J., Wang Z., Vasudevan V., Yeung L., Seyedhosseini M., Wu Y. CoCa: Contrastive Captioners are Image-Text Foundation Models. – 2022. – [https://arxiv.org/pdf/2205.01917v2.pdf].

ALEXNET

Архитектура сети AlexNet (1)

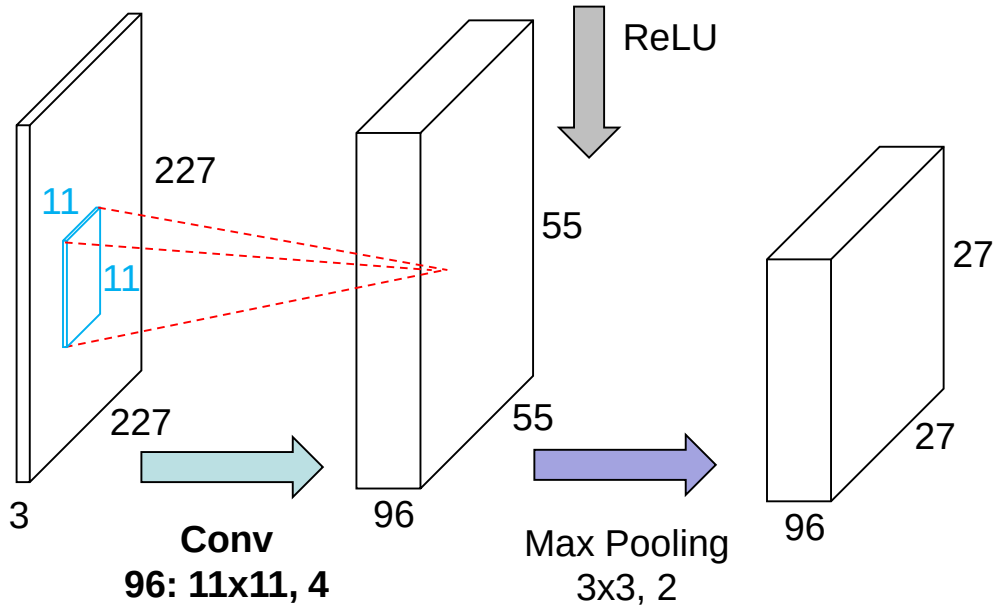
- AlexNet – первая глубокая сверточная нейронная сеть
- Разработчики сети выиграли конкурс по классификации изображений LSVRC-2012 на наборе данных ImageNet
- Ошибка классификации составила 15.3% по сравнению с ошибкой в 25.7%, полученной годом ранее



* Krizhevsky A., Sutskever I., Hinton G.E. ImageNet Classification with Deep Convolutional Neural Networks // Advances in neural information processing systems. – 2012. – [http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf].

Архитектура сети AlexNet (2)

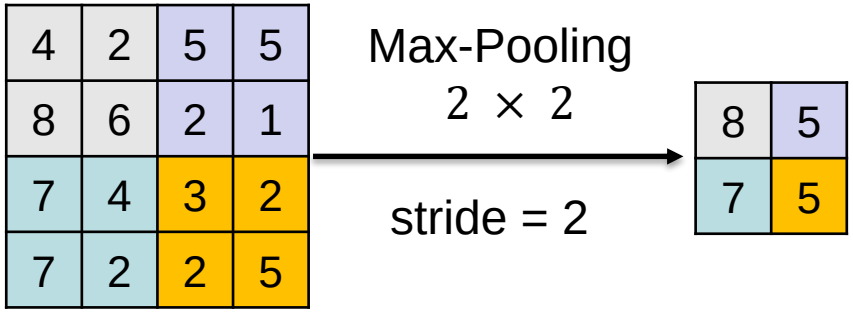
- Структура первого сверточного блока:

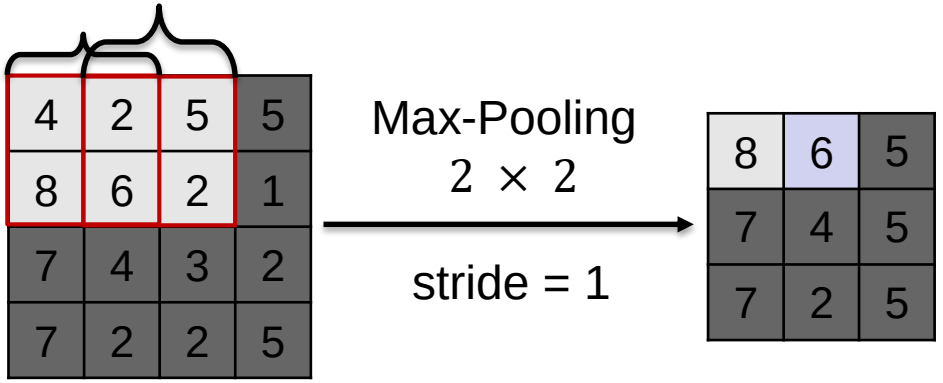


1. Входное изображение: $3 \cdot 227 \cdot 227 = 2\,043$
2. Карта признаков после свертки: $96 \cdot \left(\frac{227-11}{4} + 1\right) \cdot \left(\frac{227-11}{4} + 1\right) = 96 \cdot 55 \cdot 55 = 290\,400$
3. Карта признаков после ReLU: $96 \cdot \left(\frac{227-11}{4} + 1\right) \cdot \left(\frac{227-11}{4} + 1\right) = 96 \cdot 55 \cdot 55 = 290\,400$
4. Карта признаков после max pooling: $96 \cdot \left(\frac{(55-3)}{2} + 1\right) \cdot \left(\frac{(55-3)}{2} + 1\right) = 96 \cdot 27 \cdot 27 = 69\,984$

Особенности архитектуры сети AlexNet (1)

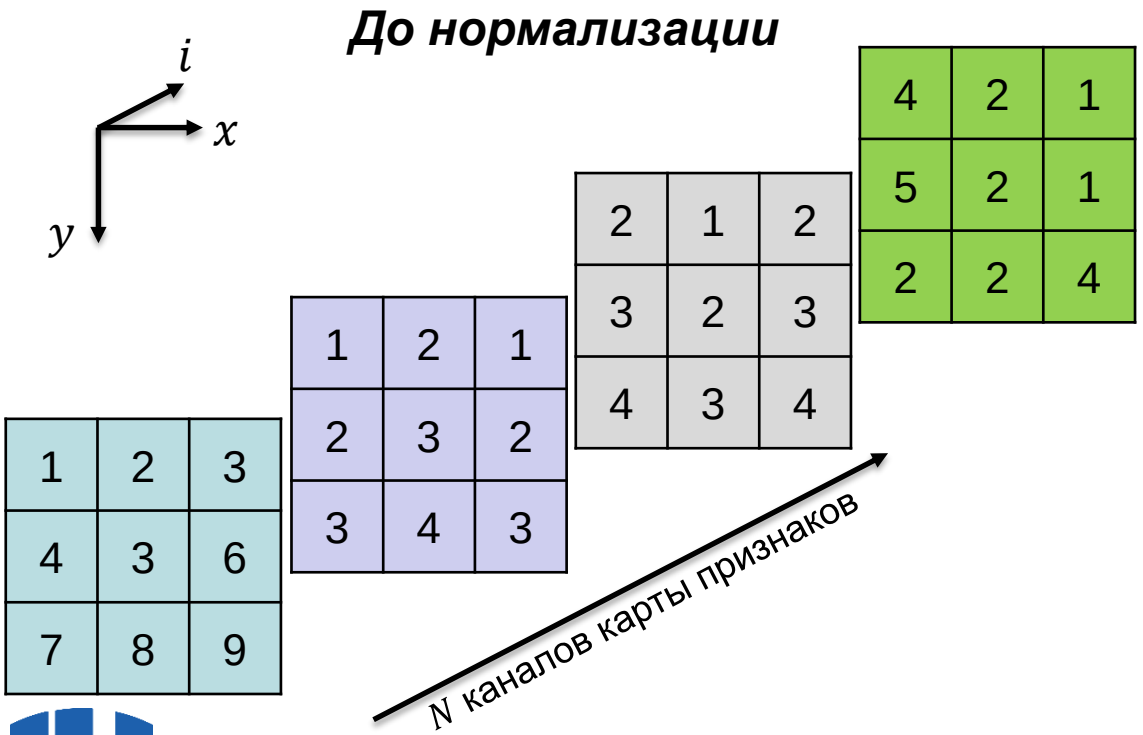
- Вход сети – трехканальное изображение 227x227 пикселей
- В качестве функции активации используется «положительная срезка» (Rectified Linear Unit, ReLU) после каждого сверточного и полносвязного слоев
- Использование слоев пространственного объединения с перекрытиями (overlapping pooling)





Особенности архитектуры сети AlexNet (2.1)

- Локальная нормализация выходов (Local Response Normalization, LRN) – нормализация выходных значений по размерности, соответствующей глубине выходной карты признаков



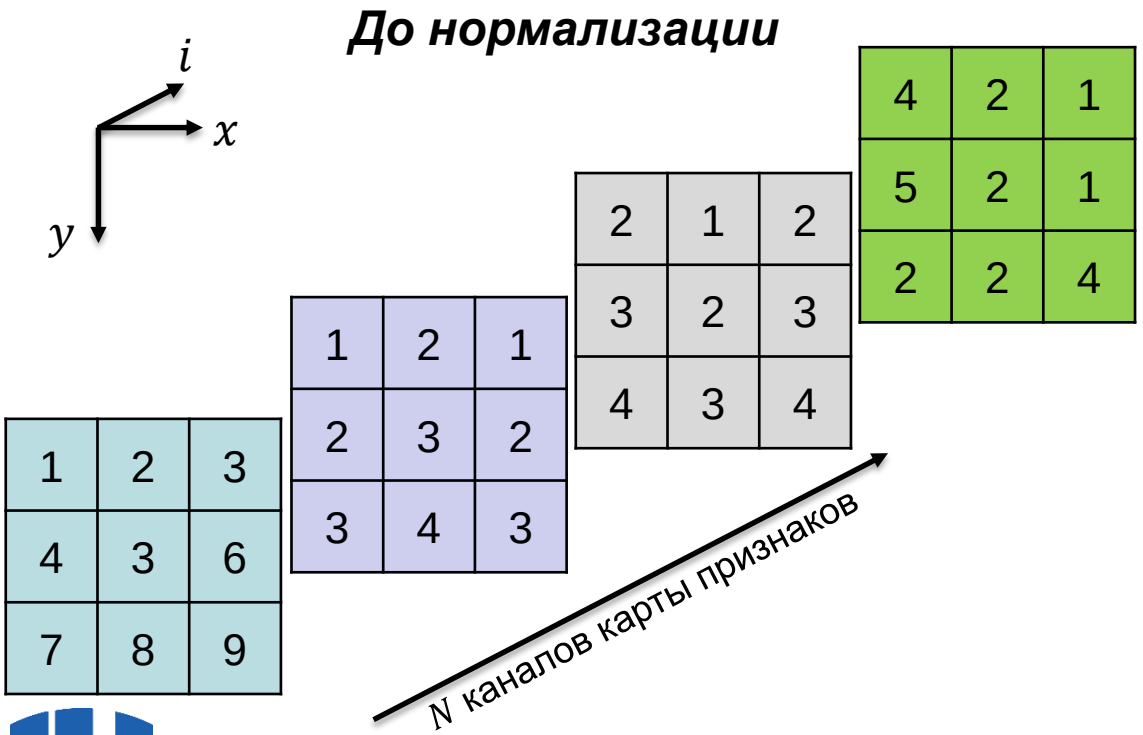
$$b^i_{x,y} = \frac{a^i_{x,y}}{\left(k + \alpha \cdot \sum_{j=\max\left(0,i-\frac{n}{2}\right)}^{\min\left(N-1,i+\frac{n}{2}\right)} \left(a^j_{x,y}\right)^2\right)^{\beta}},$$

$a^i_{x,y}$, $b^i_{x,y}$ – старое и новое значение элемента в позиции (x, y) для канала i

(k, α, β, n) – гиперпараметры:
k помогает избежать деления на ноль,
α – нормировочная константа,
β – ограничивающая константа,
n – размер окрестности

Особенности архитектуры сети AlexNet (2.2)

- Локальная нормализация выходов (Local Response Normalization, LRN) – нормализация выходных значений по размерности, соответствующей глубине выходной карты признаков



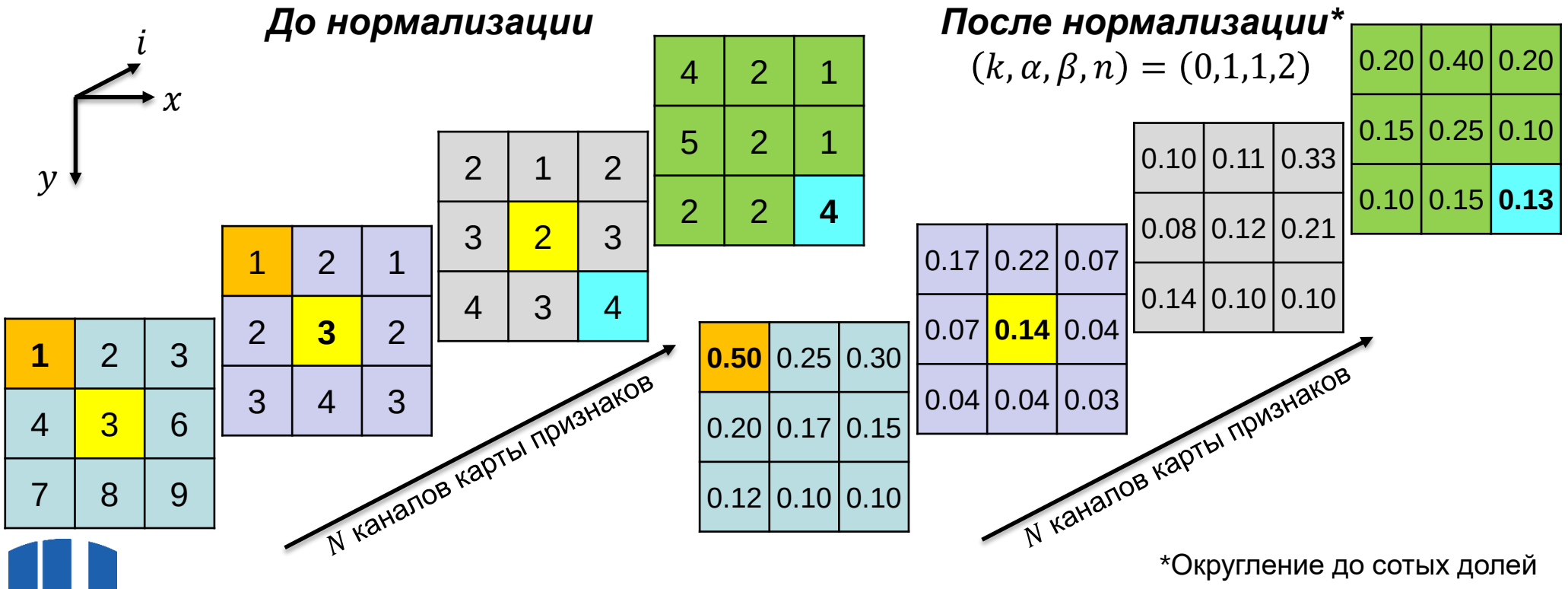
$$b^i_{x,y} = \frac{a^i_{x,y}}{\left(k + \alpha \cdot \sum_{j=\max\left(0,i-\frac{n}{2}\right)}^{\min\left(N-1,i+\frac{n}{2}\right)} \left(a^j_{x,y}\right)^2\right)^{\beta}},$$

Если $(k, \alpha, \beta, n) = (0,1,1,N)$, где N – количество каналов, то

$$b^i_{x,y} = \frac{a^i_{x,y}}{\sum_{j=\max\left(0,i-\frac{N}{2}\right)}^{\min\left(N-1,i+\frac{N}{2}\right)} \left(a^j_{x,y}\right)^2}$$

Особенности архитектуры сети AlexNet (2.3)

- Локальная нормализация выходов (Local Response Normalization, LRN) – нормализация выходных значений по размерности, соответствующей глубине выходной карты признаков



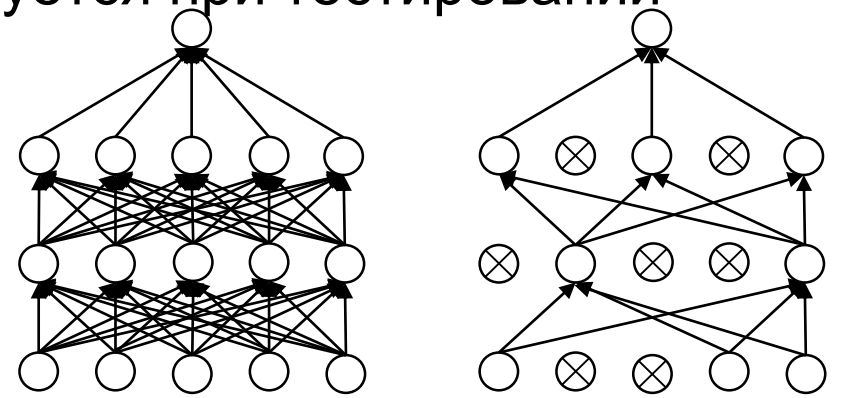
*Округление до сотых долей

Особенности обучения сети AlexNet (1)

- Увеличение количества данных (data augmentation)
  - Применение операций сдвига и зеркального отражения изображений из тренировочного набора данных
- Обучение на двух видеокартах
  - Распределение вычислений посредством разделения размерности, соответствующей глубине карт признаков

Особенности обучения сети AlexNet (2)

- Использование dropout-слоев перед первыми двумя полносвязными слоями размера
  - Обнуление выходов нейронов на каждой итерации (эпохе) обучения с вероятностью 0.5
  - Основная идея – вместо обучения одной глубокой модели обучить ансамбль, а затем усреднить результаты
  - Слева – нейронная сеть до применения Dropout, справа – та же сеть после Dropout
  - Сеть, показанная слева, используется при тестировании



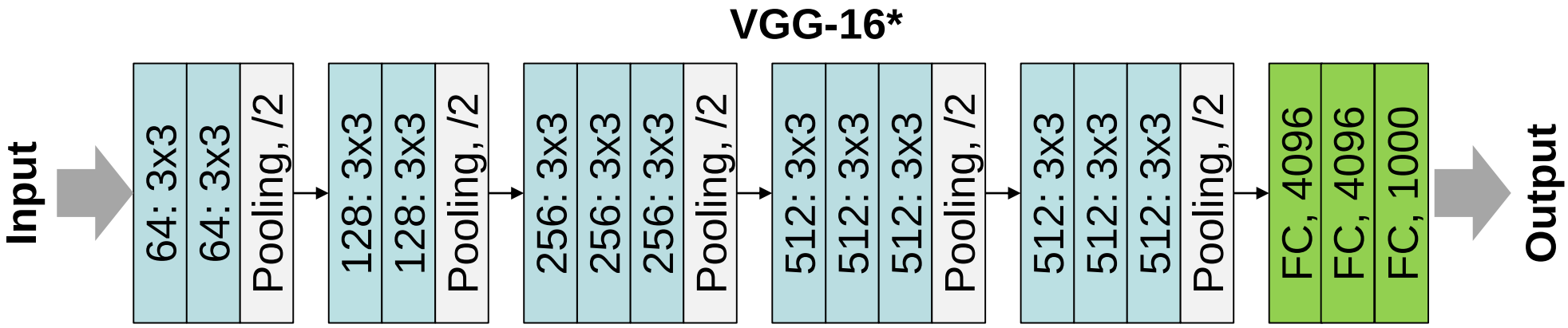
Сложность модели AlexNet

- Сложность модели:
  - Сеть содержит ~60 млн. параметров
  - Прямой проход требует выполнения ~1 миллиарда операций
  - Сверточные слои, на которые приходится 6% всех параметров, производят 95% вычислений

VGG-16, 19

VGG-16, 19 (1)

- VGG-* является улучшением модели AlexNet, принципиальное отличие состоит в том, что большие ядра сверточных фильтров (11 и 5) заменены последовательностью сверток размера 3x3

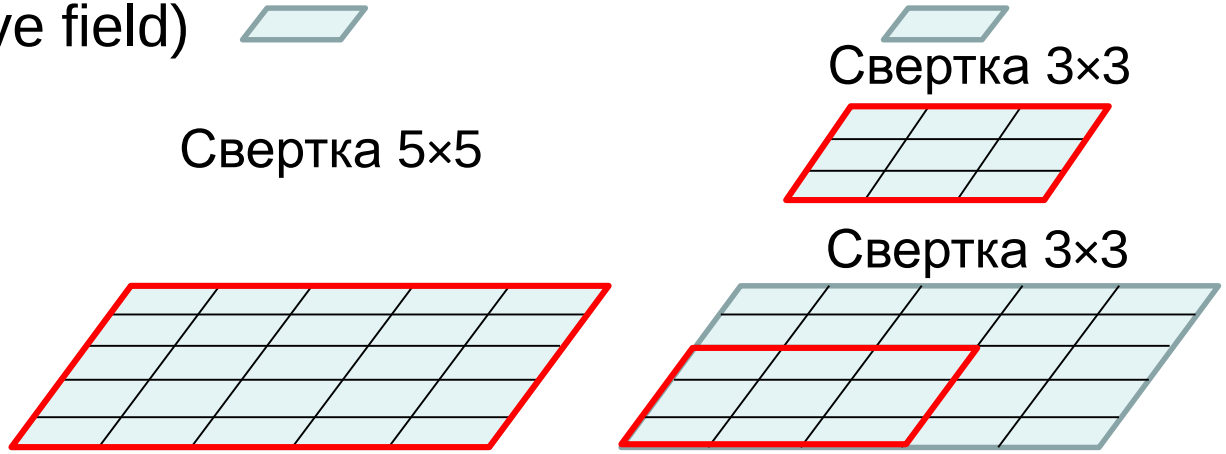


После каждого сверточного и полносвязного слоя функция активации ReLU

* Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. – 2014. – [https://arxiv.org/pdf/1409.1556.pdf].

VGG-16, 19 (2)

- Свертку с фильтром 5x5 можно заменить двумя последовательными свертками с фильтрами размера 3x3
- При этом формируется сеть с меньшим числом параметров (25 vs. 18), но с тем же размером входа и рецептивного поля (receptive field)



- VGG-19 (16 сверточных + 3 полносвязных) содержит большее количество сверточных слоев по сравнению VGG-16. Количество блоков, содержащих последовательность сверток и операцию пространственного объединения одинаковое

RESNET-50, 101, 152

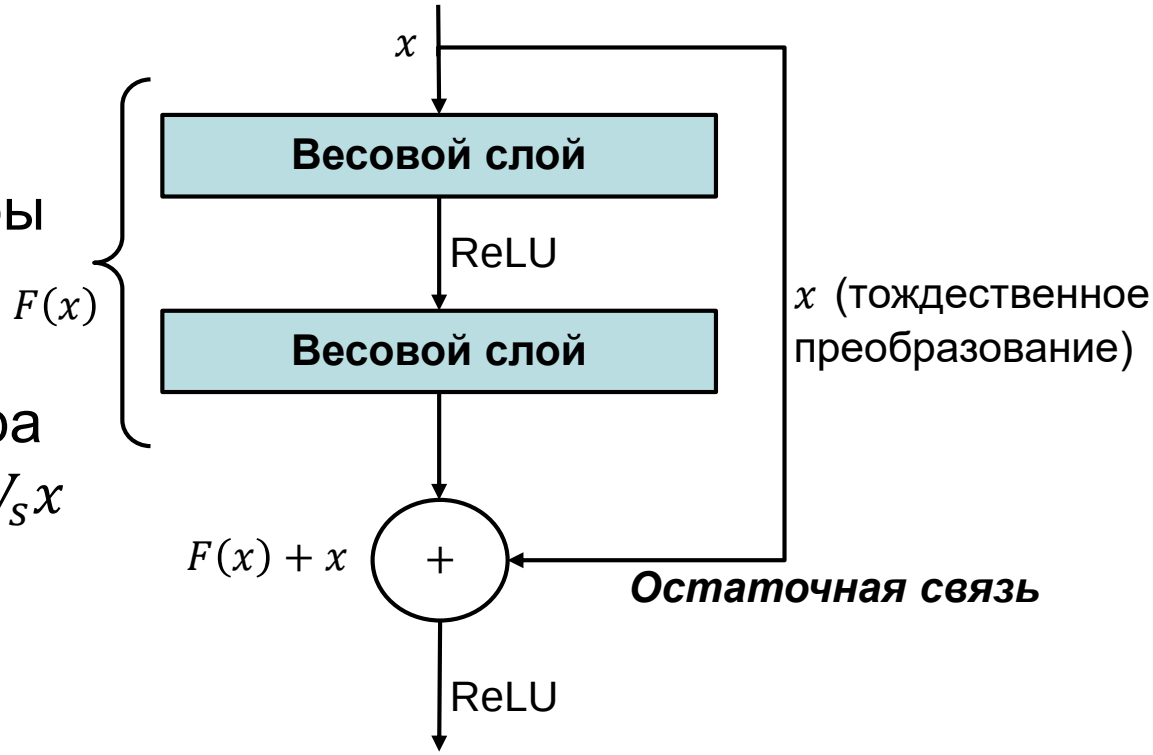
Проблема деградации глубоких моделей

- К началу 2015 года общая тенденция в разработке глубоких моделей состоит в увеличении количества сверточных слоев
- С ростом глубины сети точность насыщается и затем быстро начинает уменьшаться (деградировать)
- ***Проблема деградации глубоких моделей*** не является следствием переобучения модели, добавление дополнительных слоев приводит к еще большему значению тренировочной ошибки из-за затухающих градиентов (vanishing gradients)
- ***Остаточные сети*** (Residual Network, ResNet) решают проблему
- Идея – предположить, что некоторая последовательность слоев сети аппроксимирует не базовое отображение, а остаточное отображение

* He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. – 2015. – [https://arxiv.org/pdf/1512.03385.pdf].

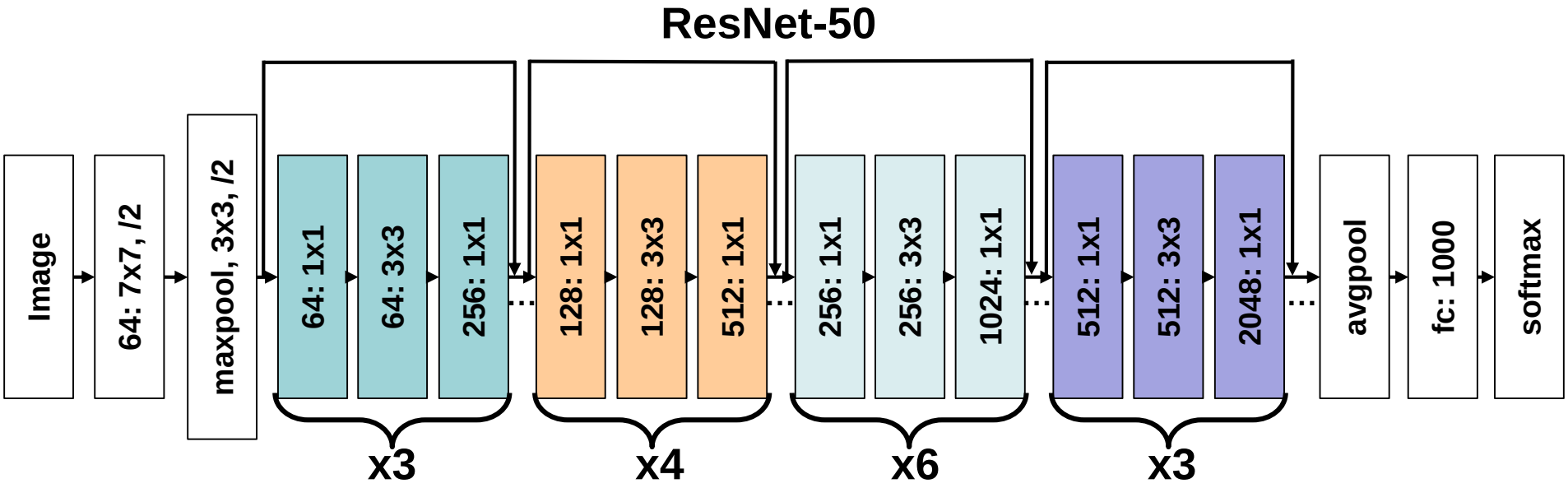
Остаточный блок

- $H(x)$ – базовое отображение
- $F(x) = H(x) - x$ – остаточное отображение
- Базовое отображение можно представить как поэлементное сложение карт признаков $F(x) + x$
- $F(x)$ и x могут иметь разную размерность, чтобы исправить эту ситуацию достаточно выполнить проекцию входного вектора признаков $y = F(x, W_i) + W_s x$



Структура ResNet-50, 101, 152

- Модели ResNet-50, 101, 152 построены по принципу наращивания сверточных слоев, проблема деградации моделей решается посредством введения остаточных связей для каждой последовательной тройки сверточных слоев

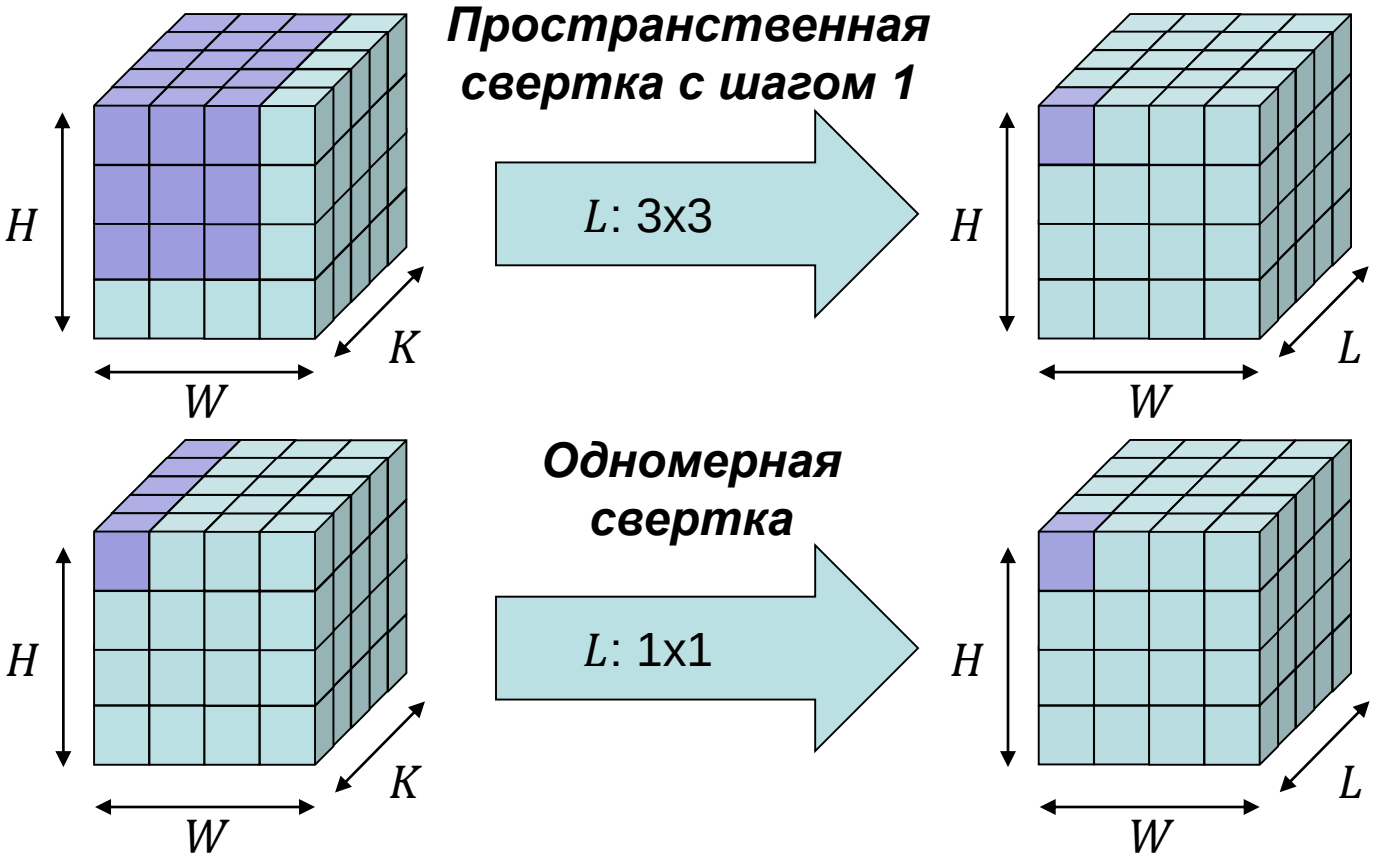


После 1- и 2-ой свертки и после остаточной связи функция активации ReLU

* He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. – 2015. – [https://arxiv.org/pdf/1512.03385.pdf].

Одномерные свертки (1)

- Свертки вида “L: 1x1” – одномерные свертки, учитывающие только особенности вдоль размерности, которая соответствует глубине карты признаков



Одномерные свертки (2)

- Одномерная свертка позволяет изменить количество каналов входной карты признаков, сохранив ее пространственное разрешение
- Применение одномерных сверток в остаточных сетях позволяет привести карты признаков, входящие в остаточный блок, к одинаковым размерностям для их последующей конкатенации

СРАВНЕНИЕ КАЧЕСТВА КЛАССИФИКАЦИИ И СЛОЖНОСТИ ГЛУБОКИХ МОДЕЛЕЙ

Тестовый набор данных

- Сравнение результатов качества классификации показано на тестовой выборке набора данных ImageNet
- Приведенные показатели собраны исследователями по результатам конкурса ILSVRC и опубликованы в Интернет [https://paperswithcode.com/sota/image-classification-on-imagenet]

Показатели качества

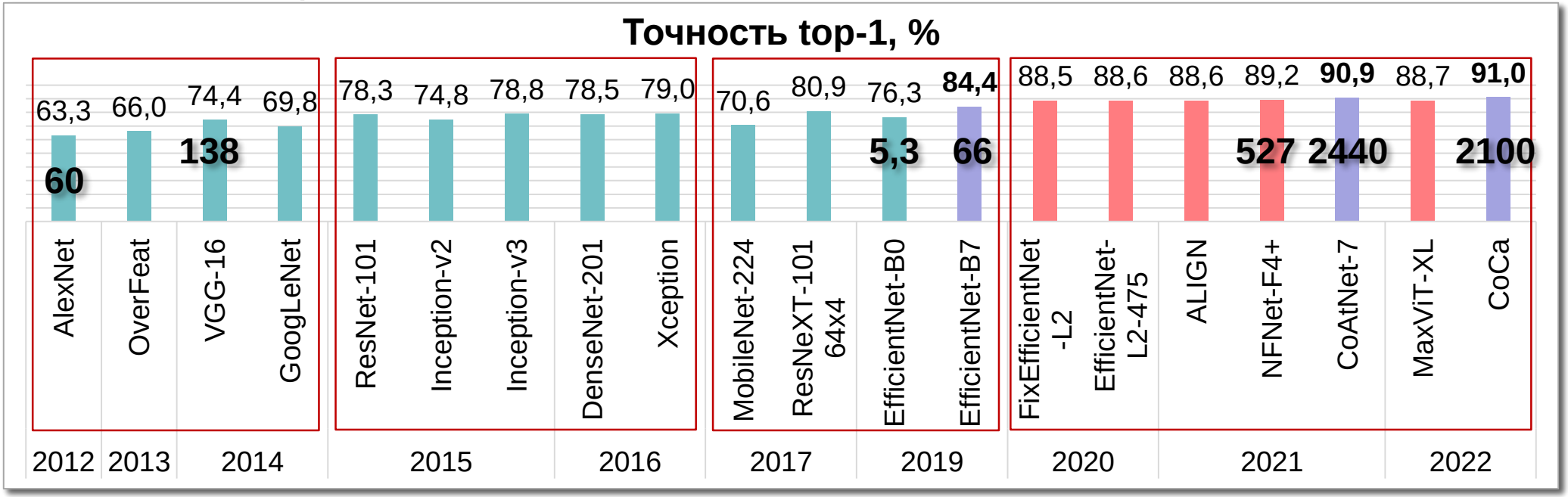
- Предположим, что N – количество категорий изображений
- Для каждого изображения $I_j, j = \overline{1, S}$ в выборке метод строит вектор достоверностей $p^j = \left(p_1^j, p_2^j, \dots, p_N^j\right)$, где p_i^j – достоверность того, что изображение I_j принадлежит классу i
- ***Точность top-K*** (top-K accuracy) определяется следующим образом:

$$topK = \frac{\sum_{j=1}^{S} 1_{\left\{i_1^j, i_2^j, \dots, i_K^j\right\}}\left(l_j\right)}{S},$$

где $\left\{i_1^j, i_2^j, \dots, i_K^j\right\} \subseteq \{1,2, \dots, N\}$, а $p_{i_1^j}^j, p_{i_2^j}^j, \dots, p_{i_K^j}^j$ – K наибольших достоверностей, l_j – класс, которому принадлежит изображение I_j согласно разметке, $1_{\left\{i_1^j, i_2^j, \dots, i_K^j\right\}}\left(l_j\right)$ – индикаторная функция

Сравнение качества классификации и сложности глубоких моделей (1)

- Изменение точности top-1 на наборе данных ImageNet для избранных моделей:

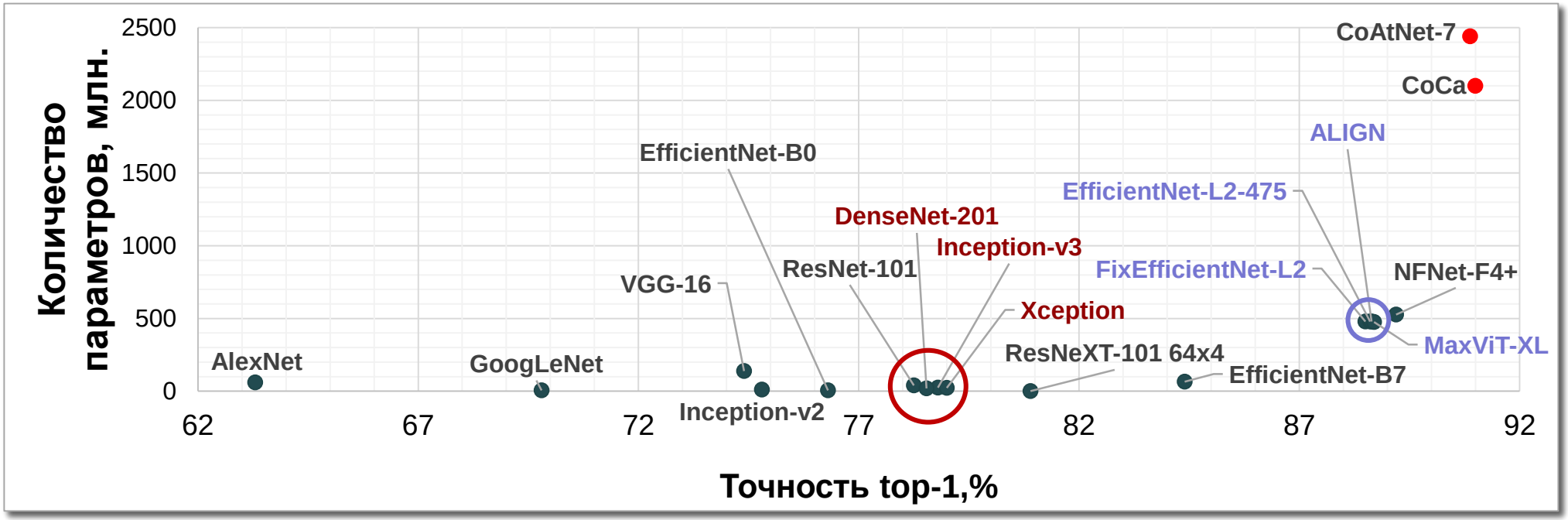


- ***За 10 лет точность выросла на ~28%, а количество параметров варьируется (зависит от цели разработки)***

* Image Classification on ImageNet [https://paperswithcode.com/sota/image-classification-on-imagenet].

Сравнение качества классификации и сложности глубоких моделей (2)

- До 2014 г. цель разработки моделей – повышение качества решения задачи, ***с 2015 г. по 2019 г. – повышение эффективности модели без потери качества, с 2020 г. – повышение качества***



Сравнение качества классификации и сложности глубоких моделей (3)

Год	Модель	top-1,%	Количество параметров, млн.
2012	AlexNet	63,30	60,0
2014	VGG-16	74,40	138,0
	GoogLeNet	69,80	5,0
2015	ResNet-101	78,25	40,0
	Inception-v2	74,80	11,2
	Inception-v3	78,80	23,8
2016	DenseNet-201	78,54	20,0
	Xception	79,00	22,8
2017	ResNeXT-101 64x4	80,90	83.6
2019	EfficientNet-B0	76,30	5,3
	EfficientNet-B7	84,40	66,0
2020	FixEfficientNet-L2	88,50	480,0
	EfficientNet-L2-475	88,61	480,0
2021	ALIGN	88,64	480,0
	NFNet-F4+	89,20	527,0
	CoAtNet-7	90,88	2440,0
2022	MaxViT-XL	88,70	475,0
	CoCa	91,00	2100,0

Сравнение качества классификации и сложности глубоких моделей (4)

- Примечания:
  - Повышение эффективности модели – снижение вычислительной сложности модели (количества выполняемых операций) и уменьшение размеров (количества параметров) модели
  - Вычислительная сложность модели напрямую не связана с числом параметров
  - На практике (при многократном выводе), как правило, важна вычислительная сложность – число операций за прямой проход по сети
- Сравнение вычислительной сложности глубоких моделей
  - Bianco S., Cadene R., Celona L., Napoletano P. Benchmark Analysis of Representative Deep Neural Network Architectures. – 2018. – [https://arxiv.org/pdf/1810.00736.pdf].

Заключение

- Множество глубоких моделей для классификации изображений не ограничивается приведенными в настоящей лекции, существует множество модификаций базовых архитектур
- В настоящее время большое количество моделей для решения задач из других областей используют описанные архитектуры за счет применения переноса обучения (transfer learning), либо используют базовые строительные блоки рассмотренных моделей
- ***Оптимальная модель – компромисс между точностью и сложностью***
  - Точность определяется требованиями, предъявляемыми к решению практической задачи
  - Сложность определяется доступными вычислительными ресурсами и требованиями ко времени выполнения

Основная литература

- Учебно-образовательный курс «Современные методы и технологии глубокого обучения в компьютерном зрении» [http://hpc-education.unn.ru/ru/обучение/курсы/магистратура/deep_learning_in_computer_vision].

Контакты

- **Кустикова Валентина Дмитриевна**
  к.т.н., доцент кафедры МОСТ ИИТММ ННГУ
  valentina.kustikova@itmm.unn.ru