УДК 004.8

Современные тенденции в развитии нейронных сетей

Насибулов Илья Андреевич (Институт систем информатики СО РАН, Новосибирский государственный университет),

Насибулов Егор Андреевич (Институт систем информатики СО РАН, Новосибирский государственный университет)

В последние 30 лет нейронные сети являются одним из наиболее бурно развивающихся направлений искусственного интеллекта. Они широко применяются в обработке звука и изображения, медицине, задачах анализа и генерации контента и других. Это стало возможным благодаря значительному росту вычислительных мощностей, возможности обработки больших объёмов данных и развитию теории нейросетей.

В данной работе приведён анализ развития алгоритмов обучения и архитектур нейросетей от их зарождения до современного состояния. Были выделены наиболее активно развивающиеся направления, такие как большие языковые модели, сети-гиганты и мультимодальные модели. Также упомянуто перспективное направление развитие, связанное с сетями Колмогорова-Арнольда.

Ключевые слова: *искусственный интеллект, нейронная сеть, машинное обучение, глубокое обучение, свёрточные нейронные сети, языковые модели, сети-трансформеры, сети Колмогорова-Арнольда.*

1. Введение

Подход к искусственному интеллекту (*ИИ*) в XXI веке значительно изменился. Термин ввёл Джон Маккарти [1], определяя ИИ как машинные вычисления, способные решать задачи, предназначенные для человека и даже имитировать поведение человека. Термин подразумевал под собой следующее: искусственный — эта часть от машины, а интеллект — человеческая часть, способная решать задачи, в том числе и когнитивные, и выдавать себя за человека. Подход к реализации такого ИИ описывался принципом, что ИИ должен самовоспроизводиться с помощью несложных инструкций кода [2]. Несмотря на то, что в определении ИИ отсутствуют технические детали реализации, подразумевая, что способов

реализаций может быть несколько, в последнее время ИИ всё чаще ассоциируется с нейросетями [3], особенно глубокого обучения.

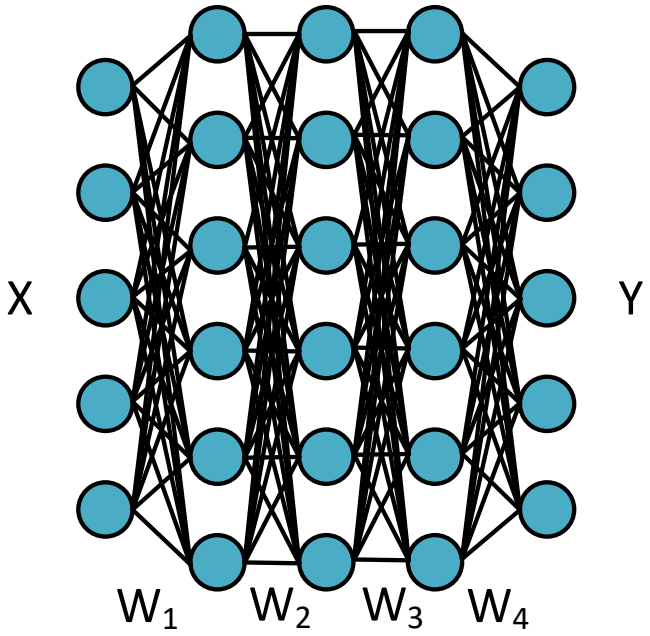

Рисунок 1. Представление нейросети в виде вычислительного графа. X — входные данные, Y — выходные данные, W — матрицы весов.

Первые нейросети задумывались как вычислительные модели, основанные на имитации искусственными нейронами биологических [4], способные обучаться на примерах и применять полученную информацию для решения аналогичных задач. Развивая эту идею, можно воспринимать нейросети как вычислительные графы [5], на рёбрах которых происходят несложные арифметические действия (рис. 1). Таким образом, решение задач с помощью нейросетей технически относится к моделированию. Перспективы создания ИИ с помощью нейросетей послужили выделению отдельного направления в науке, получившего название общий искусственный интеллект [6]. Несмотря на то, что масштабирование нейросетей позволило получить ряд значимых результатов, chatGPT и аналогичные архитектуры всё ещё не могут полноценно заменить живых специалистов, что показал опыт ведущих наукоёмких компаний. Тем не менее, спектр задач, которые решаются с применением нейросетей, довольно обширен.

Одной из типичных задач является анализ изображений. С ним связаны задачи от распознавания рукописного текста до распознавания лиц. Одной из задач обеспечения безопасности является распознавание номеров машин с помощью видеонаблюдения с последующим выявлением нарушений правил дорожного движения [7]. Анализ изображений

находит себя и в медицине [8, 9] — по симптомам и соответствующим снимкам пациентов нейросети помогают соответствующим специалистам поставить диагноз. Есть применение и в аграрной области — если анализировать снимки полей с урожаем, например, поля пшеницы, то можно выявить потенциально больные колосья для последующей обработки посевов. Или же можно получить достаточно много информации по сорту будущего посева и потенциальному процессу роста по снимку семян или листов растений [10].

Нейросети активно применяются в задачах компьютерной лингвистики, психологии и изучения мышления [11, 12]. Есть подходы, позволяющие использовать нейросети для переводов различных текстов и анализа аргументации [13]. В числе успешно решаемых задач — переводы узкоспециализированных технических текстов с определёнными лексическими оборотами [14, 15]. Также ведутся работы в направлении переводов с малораспространённых языков различных народов России и не только [16].

Своё применение нейросети находят и в биоинформатике. Например, анализ последовательностей белков можно частично упростить [17, 18], что в дальнейшем облегчает расшифровку последовательностей аминокислот в организме.

Ещё одним применением нейросетей является генерация контента, такого как текст, изображения, звук и т.д. Свою нишу нейросети постепенно занимают и в генерации программного кода [19]. По данному вопросу существуют две точки зрения исследователей. Одни считают, что через декаду лет программисты как таковые будут нужны не в качестве специалистов по написанию кода, а скорее операторов нейросетей и ИИ для формирования конечного программного продукта. Контраргументом является то, что код генерируется на основе уже существующих материалов, заложенных в нейросети на этапе обучения. Большая часть сгенерированного кода не является кодом высокого уровня в смысле оптимального по используемым ресурсам ЭВМ и их производных. Большинство программного кода написано программистами низкой квалификации, так называемыми младшими программистами, и в условиях ограниченности по времени разработки, что несёт в себе изначальные изъяны, которым обучается нейросеть и, в дальнейшем, при генерации нового кода считает эти изъяны допустимой нормой. Первые считают, что для исправления таких проблем и нужно будет «оперирование» нейросетей, т. е. дальнейшее улучшение с помощью них же сгенерированного изначального кода. Обе позиции имеют место быть, однако сложно спорить с тем, что нейросети уже как минимум сдают ЕГЭ по разным предметам на проходной балл для поступления в ВУЗы. Это активно обсуждалось в прессе и образовательных учреждениях [20].

Так называемые из-за своего размера сети-гиганты объединяют в той или иной степени все вышеперечисленные возможности. Актуальные на 2025 г. архитектуры сетей-гигантов содержат миллиарды параметров. ChatGPT и его аналоги, в числе которых Гигачат от Сбербанка, YandexGPT от Яндекса и китайская Deepseek уже используются широкими слоями населения. Возможности этих нейросетей обширны, и есть публикации в прессе о случаях, когда пользователи предпочитают общаться с нейросетями вместо своих коллег и друзей. С тех пор как тест Тьюринга был пройден нейросетями, уровень доверия к нейросетям растёт, в том числе в таких важных сферах, как здравоохранение [21]. Тем не менее, не стоит забывать, что алгоритмы нейросетей не являются абсолютно надёжными и в большинстве моделей недоступны для верификации. Таким образом, ошибки и галлюцинации нейросетей становятся непредсказуемыми и опасными. В последние годы интерес к исследованию данной проблемы держится на устойчиво высоком уровне как с технической [22, 23, 24], так и с этической стороны [25, 26, 27].

Целью данной работы является дать краткий обзор развития нейронных сетей от первых персептронов до сетей Колмогорова-Арнольда (*KAN*). К сожалению, все подобные обзоры [28, 29, 30] быстро устаревают в связи с бурным развитием области и появлением новых направлений, таких как сети Колмогорова-Арнольда [31]. Несмотря на то, что KAN появились сравнительно недавно и большого количества результатов с их использованием ещё не было получено на момент написания данной статьи, архитектура уже зарекомендовала себя как перспективный инструмент для решения нелинейных задач.

2. Зарождение нейросетей

Первой нейросетью считается искусственный нейрон, предложенный У. Маккалаком и У. Питтсом [32]. Они же и ввели понятие искусственной нейронной сети (*ИНС*). Нейрон имеет строение аналогичное сумматору, однако имеются только логические сигналы 0 и 1. В качестве функции активации была выбрана пороговая функция Хевисайда. Это служит неким аналогом активации человеческих нейронов в нервной системе. Таким образом, мы получаем

$$s = w \cdot x + b, z = g(s),$$

Где x и w — вектора входных данных и весов соответственно, $x \in Z_{\{0,1\}}^{N}$, b — смещение, z — выход, g — активационная функция, в данном случае

$$g = \begin{cases} 1, \text{если } s > a, \\ 0, \text{если } s \leq a, \end{cases}$$

где $a > 0$ — порог активации. Если учесть факт, что при формировании нейросетей нейроны объединяются в нейросетевые слои, то вышеописанные уравнения можно записать в матричном виде

$$s = Wx + b, z = g(s),$$

где W — матрица весов.

Первой ИНС, которая способна была решить задачу классификации и широко применялась на практике, была нейросеть Ф. Розенблатта [33], так называемый персептрон. Нейросеть интересна наличием одного скрытого слоя нейронов в ней. Скрытыми слоями называются слои между входными данными и выходным слоем. Веса и смещения задаются случайным образом из {-1, 0, 1}, а в качестве функции активации используется sign. Таким образом, получаем

$$s_1 = W_1 x + b_1, y = g_0(s_1),$$

$$s_2 = W_2 y + b_2, z = g_1(s_2),$$

где $z = g_1(s) = sign(s), x \in Z_{\{0,1\}}^N$. y является выходом первого слоя и одновременно входными данными для второго слоя.

М. Минский и С. Паперт в 1969 году в своей работе [5] рассматривают персептроны Розенблатта в качестве вычислительных графов, что открывает возможности для применения нейросетей в математическом моделировании. Тем не менее, никакого выхода на новые практические задачи продемонстрировано в работе не было, и на какое-то время интерес исследователей к этой области ИИ угас.

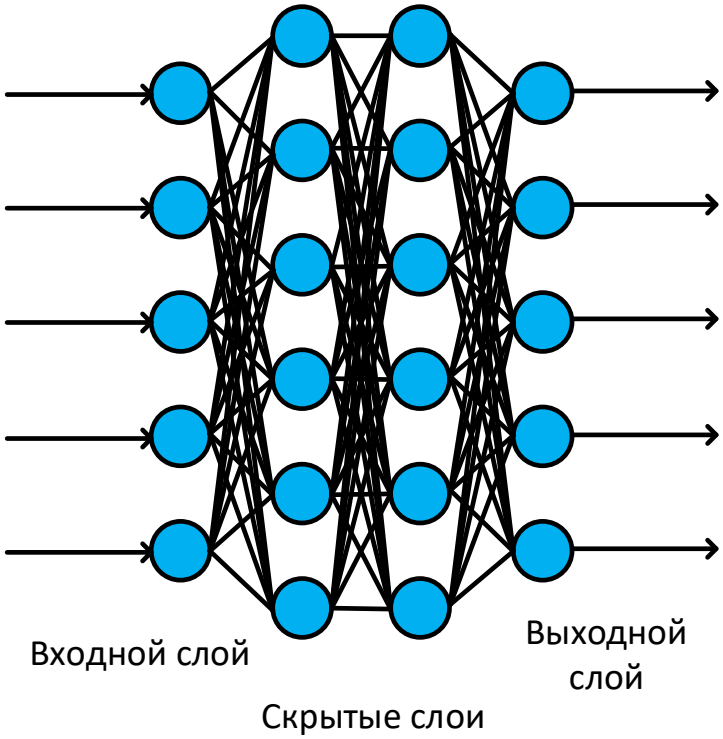

Рисунок 2. Многослойный персептрон. Все нейроны предыдущего слоя соединены со всеми нейронами следующего слоя, и только с ними.

3. Нейронные сети со скрытыми слоями

В 1986 вышла работа Д. Румельхарта [34], в которой был продемонстрирован потенциал многослойных персептронов Розенблатта (рис. 2) с некоторыми улучшениями:

$$s_1 = W_1 x + b_1, y = \tilde{g}_0(s_1),$$

$$s_2 = W_2 y + b_2, z = \tilde{g}_1(s_2),$$

где $\tilde{g}_i(s) = \tilde{g}_{sg}(s) = \frac{1}{1+e^{-x}}$ является сигмоидальной функцией или $\tilde{g}_i(s) = \tilde{g}_{th}(s) = th(s)$ гиперболический тангенс, $i = 0, 1$. Входной слой уже не ограничен исключительно логическими сигналами, а принимает вещественные значения $x \in \mathbb{R}^N$. В обучении нейросетей в данной работе активно применяется метод обратного распространения ошибки, который предложили и развили А. Галушкин и П. Вербос в [35, 36, 37, 38].

Для развития нейронных сетей критически важными являются теоремы о суперпозиции Колмогорова-Арнольда [39] и универсальная теорема аппроксимации [40]. Согласно первой, $\forall f \in \mathbb{C}[0,1]^d$ существует $d(2d+1)$ функций одного аргумента $\phi_{ij} \in \mathbb{C}[0,1]$ таких, что f может быть представлена в виде

$$f(x_1,\dots,x_d)=\sum_{i=1}^{2d+1}\chi_i\left(\sum_{j=1}^{d}\phi_{ij}(x_j)\right)$$

для некоторых $\chi_i \in \mathbb{C}[0,1]$, зависящих от f.

Универсальная теорема аппроксимации представляет собой адаптацию теоремы суперпозиции Колмогорова-Арнольда к области нейронных сетей и говорит, что искусственная нейронная сеть прямого распространения с одним скрытым слоем может аппроксимировать любую непрерывную функцию многих переменных с любой точностью, при условии, что сеть имеет в скрытом слое достаточное число нейронов N, имеющих сигмоидальную функцию активации s_{sg}.

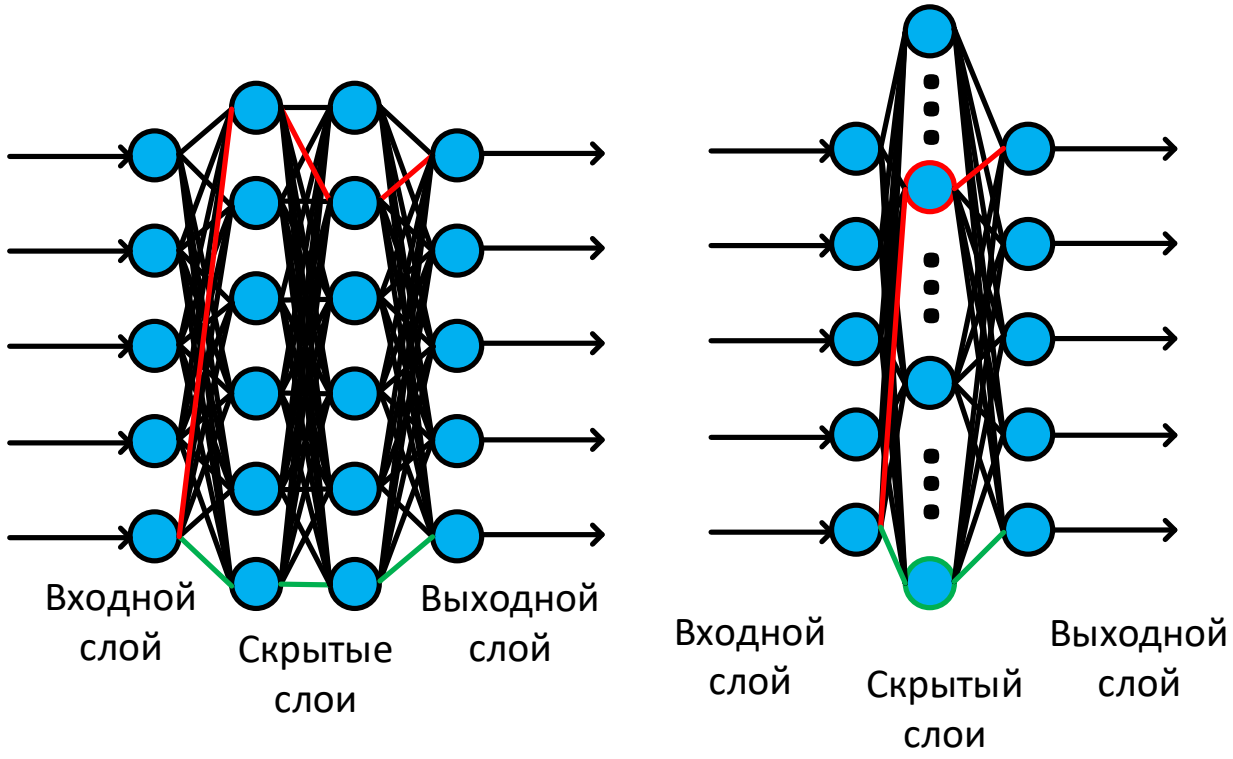

Рисунок 3. Многослойная нейронная сеть (слева) и эквивалентная ей сеть с одним скрытым слоем (справа). Цветом выделены эквивалентные пути, дающие одинаковый вклад в выходной слой, в разных архитектурах.

Действительно, если каждый уникальный путь через нейроны от входного к выходному слою (рис. 3) представить в виде нейрона этого единственного слоя, получим полную эквивалентность с точки зрения результата работы нейросетей как функционалов. Сам факт

существования такого гомеоморфизма оказался крайне полезным для доказательства ряда теорем.

Важной для развития рекуррентных нейронных сетей является теорема о полной тьюринговости, доказанная Х. Зигельманом и Э. Сонтагом [41]: Любые машины Тьюринга могут моделироваться полностью связанными рекуррентными сетями, созданными из нейронов с сигмоидальными функциями активации, при условии, что сеть имеет достаточное число нейронов в скрытом слое M и достаточное число шагов временной памяти K.

В работе К. Хорника [42] 1991 года значительно расширено понимание возможностей нейронных сетей. Была обобщена универсальная теорема аппроксимации для случая произвольных нелинейных функций активации. Это позволило сделать следующие фундаментальные выводы: нейросети способны к аппроксимации, а свойства последней определяются архитектурой; выбор конкретной функции активации менее критичен, чем считалось ранее; для построения эффективных нейронных сетей может быть использован широкий класс функций активации.

Ещё одним доказанным в 1992 году результатом является универсальная аппроксимационная теорема рекуррентных нейронных сетей, доказанная К. Фунахаши и Ю. Накамурой [43]: Любая нелинейная динамическая система может быть аппроксимирована рекуррентной нейронной сетью с любой точностью, без ограничений на компактность пространства состояний системы, при условии, что сеть имеет достаточное число нейронов в скрытом слое.

Т. Чоу и Х. Ли в 2000 году расширили универсальную аппроксимационную теорему рекуррентных нейронных сетей на случай неавтономных нелинейных обыкновенных дифференциальных уравнений [44].

Несмотря на то, что появились теоремы, раскрывающие область применимости нейронных сетей для решения различных задач машинного обучения, на практике возможности ИНС с одним скрытым слоем достаточно ограничены и не подходят для решения большого класса задач. Однослойная нейросеть, которую можно построить для любой многослойной сети, крайне плохо обучается и годится только в качестве абстрактной математической модели. Попытки же добавления скрытых слоёв нейронов не приносят значительного продвижения из-за проблемы затухания градиента [45]. В процессе обучения поправки из выходного слоя просто не доходят до входа сети.

4. Глубокие (многослойные) нейросети

Вторая половина 2000-х годов приносит плоды, благодаря которым интерес к нейросетям снова разжигается в научном сообществе. Работы Д. Хинтона и Р. Салахутдинова [46] предлагают некоторые способы обучения нейросетей со многими скрытыми слоями, однако на практике эти способы получились затратными в плане вычислений и неустойчивыми для сетей с более чем 3–5 слоями. Для решения проблемы обучения нейросетей со многими скрытыми слоями оказались принципиальными два фактора. Один из них — подобрать правильную функцию активации, что сделал в 2010 году В. Наир, предложив использовать функцию ReLU (*Rectified Linear Unit*) [47]. Функция ReLU представляет собой $g_{ReLU}(s) = \max(0, s)$. Вторым фактором стал способ начальной инициализации весов нейросетей. К. Глорот и Й. Бенжио в 2010 году предложили [48] дисперсию инициализирующего шума находить по формуле

$$Var(w) = \frac{2}{N_{in} + N_{out}},$$

где N_{in} и N_{out} — число искусственных нейронов в предыдущем и следующем нейрослое соответственно. Эти факторы дали старт для очередного быстрого развития нейросетей, а как сопутствующий результат, сформировали понятие глубокой нейронной сети — нейросети, содержащей 2 и более скрытых нейрослоёв (рис. 3).

5. Свёрточные нейросети

Отдельного упоминания стоит история развития свёрточных нейросетей, так как они показывали и показывают себя как один из самых эффективных инструментов для анализа изображений. История создания свёрточных нейросетей уходит в 50-е — 60-е годы, когда работы таких исследователей, как Д. Хебба [4] и других нейрофизиологов показали, что зрительная кора головного мозга для распознавания объектов имеет отдельный вид нейронов, которые реагируют на определённые шаблоны (*pattern*) в получаемых сигналах — линии, углы, движение.

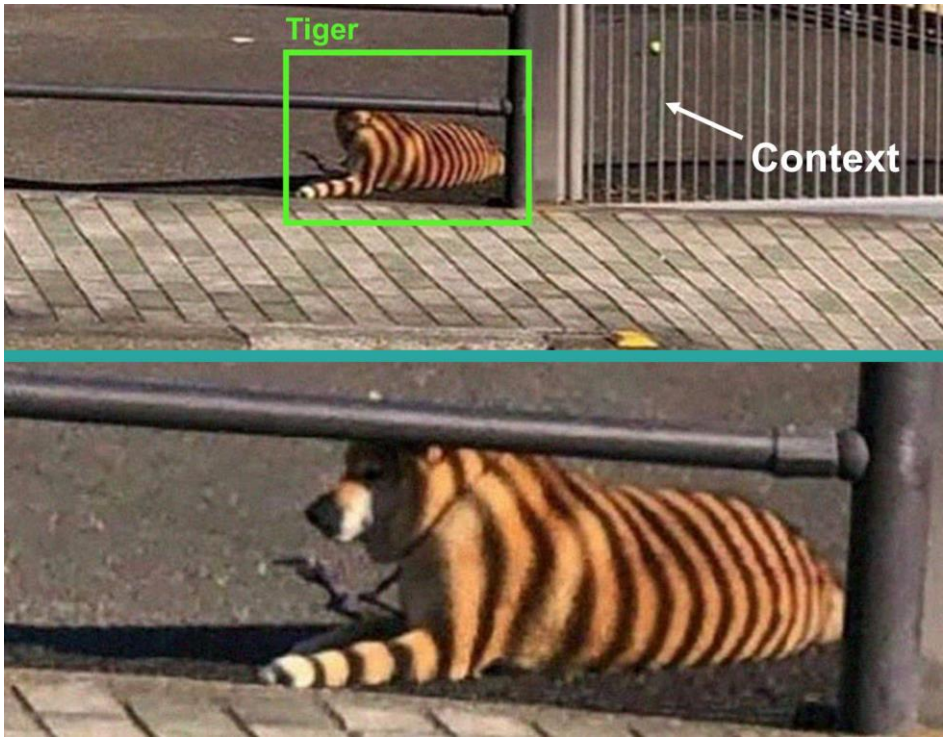

Рисунок 4. Завирусившееся в интернете изображение, на котором IntelexVision вместо собаки видит тигра, не учитывая контекст в виде забора, дающего полосатую тень.

В связи с этим принцип работы нейросетей основан на поиске тех или иных признаков объекта, который подаётся в качестве входных данных. Этот процесс называется поиском шаблонов в данных. Однако входные данные могут содержать персональный контекст, который может искажать исходные признаки объекта. Ярким примером такого искажения является следующее изображение (рис. 4), на котором человек определит собаку, однако нейросеть, ранее обученная решать задачу выявления животных на изображении и классифицировать выявленное животное, относит данную собаку к тигру.

Подобные случаи неверного истолкования признаков входных объектов доказывают важность персонального контекста и в целом обработки данных при работе с нейросетями, а не только выявления самой эффективной архитектуры и получения лучших метрик в выходных данных.

Первые свёрточные нейросети работали на следующих принципах: к изображению применяется операция свёртки; для обнаружения шаблонов используются соответствующие фильтры; изображение обрабатывается слой за слоем с целью извлечения более сложных шаблонов; для обучения сетей используется метод обратного распространения ошибки.

Первой практически работающей архитектурой свёрточной нейросети стала LeNet, которую в 1989 году разработал Я. Лекун с соавторами [49] для распознавания написанных от руки цифр почтового индекса, предоставляемых почтовой службой США. Работа нейросети основана на принципе разделения весов, когда несколько нейронов или групп нейронов используют одни и те же веса для снижения количества уникальных параметров в модели и оптимизации процесса обучения. К 1998 году идеи коллектива выливаются в архитектуру LeNet-5 [50], состоящей из чередования свёрточных слоёв и слоёв субдискретизации, завершая выход нейросети двумя полносвязными свёрточными слоями.

В 2010 году Д. К. Кирешан и Ю. Шмидхубер публикуют препринт [51], в котором реализованная архитектура нейросети добивается рекордных на то время показателей точности определения рукописных символов эталонных тестов MNIST. Для достижения цели была спроектирована нейросеть, содержащая 9 скрытых слоёв, а для её обучения использовался графический процессор, что позволило снизить время обучения сети.

М. Цейлер с соавторами публикуют работу [52], в которой предложили использование слоя деконволюции. Если рассмотреть его на примере входного изображения y_i с K_0 входными каналами $y_1, y_2, .., y_{K_0}$, то каждый из этих каналов представляется в виде линейной суммы K_1 скрытых карт признаков z_k^i, свёрнутых фильтрами $f_{k,c}$:

$$\sum_{k=1}^{K_1} z_k^i \oplus f_{k,c} = y_c^i.$$

Если изображение y_c^i имеет размер $N_r \times N_c$, а фильтры имеют размер $H \times H$, то карты скрытых объектов имеют размер $(N_r + H - 1) \times (N_c + H - 1)$.

В 2012 году **AlexNet** — **с**вёрточная нейронная сеть, разработанная командой исследователей под руководством А. Крижевского [53] — одерживает революционную победу в конкурсе ImageNet LSVRC-2012, где показывает впечатляющий результат с ошибками топ-1 и топ-5 в 37,5% и 17,0% против 45,7% и 25,7% у конкурентов, усреднявших показания двух классификаторов, обученных на Фишеровских векторах [54]. Ключевыми особенностями AlexNet стали применение функции активации ReLU, эффективное использование графического процессора для обучения и применение техники Dropout для борьбы с переобучением, предложенная ранее тем же коллективом [55]. Техника основана на выключении в течение одной итерации обучения части

нейронов скрытых слоёв с определённым шансом с последующим их возвращением в конце итерации с последующей нормировкой весов нейронов.

М. Лин с соавторами в 2013 году [56] предложили ввести слой глобальной усредняющей субдискретизации, что в дальнейшем позволило создавать полносвязные свёрточные нейросети без полносвязных слоёв в конце сети. Также было показано, что вставка многослойных персептронов между свёрточными слоями усиливает свёрточные свойства.

В 2014 году К. Симонян и Э. Зиссерман публикуют работу [57], в которой описывают так называемую VGG-сеть. Архитектура предлагает использовать вместо тяжёлых свёрточных слоёв размером 5x5 и более свёрточные слои размером 3x3 с увеличением их количества — сама нейросеть включала 19 нейрослоёв. Как оказалось, такой подход оказывается крайне эффективным за счёт уменьшения числа параметров и уменьшения тяжеловесных арифметических операций.

К. Сегеди с коллегами в 2014 году публикуют работу [58], где предлагают к рассмотрению архитектурный принцип под названием Inception и конкретную нейросеть с 22-мя нейрослоями GoogLeNet, основанную на данной архитектуре и успешно применяемую в задачах классификации и обнаружения. В архитектуре Inception ключевым принципом является использование ядер размера 1x1, что является переосмыслением идей Лин [56].

В том же 2014 году Л. Сифре защищает кандидатскую диссертацию [59], посвящённую получению шаблонов изображений, стабильных относительно трансляции и поворота посредством каскада вейвлет-преобразований по пространственным и угловым координатам. Для этого он вводит Depthwise Separable свёрточный слой нейронов, в котором также являются важными ядра размера 1x1.

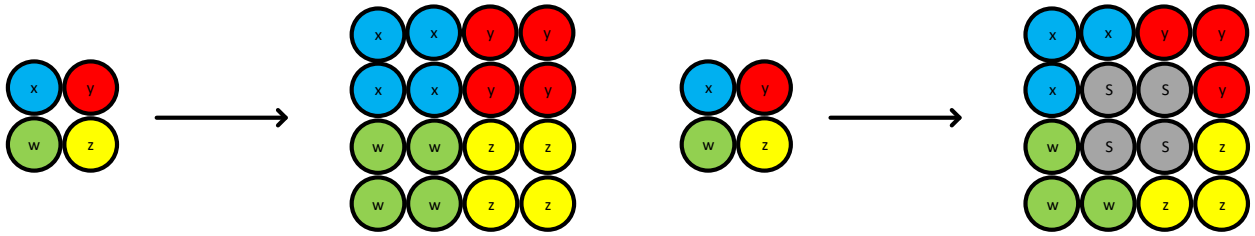

Рисунок 5. Пример операции пространственного расширения карты признаков. Слева — простейшая схема с дублированием существующих, справа — с применением интерполяции и созданием синтетических признаков. В данном примере $S = \frac{x+y+z+w}{4}$

Развивая идеи [56], в 2014 году Дж. Лонг с соавторами в своей работе [60] предложили концепцию полносвёрточной сети для задач, результатом в которых является изображение такого же размера, как исходное. Предлагается использовать операцию пространственного расширения карты признаков (*Upsampling*) для решения проблемы несбалансированных наборов данных (рис. 5). В простейшем случае операция представляет собой метод копирования существующих предметов, а может применяться и интерполяция для создания новых синтетических примеров в классе на основе существующих. В дальнейшем подход с применением интерполяции доработали и нейрослой стал называться слоем транспонированной свёртки [61].

Рисунок 6. Появление дефектов в случае использования только деконволюции (сверху), для сравнения — использование деконволюции в сочетании с расширением карты признаков (снизу). Источник: [62].

В работе [62] А. Одена изучает пиксельные дефекты, возникающие при анализе изображений с использованием нейросетей с пространственным расширением карты признаков. Коллектив приходит к выводу, что стандартный подход к созданию изображений с помощью деконволюции имеет значительные успехи, но также имеет некоторые концептуально простые проблемы, которые приводят к появлению дефектов в создаваемых изображениях (рис. 6). Тщательное же продумывание архитектуры нейросетей с использованием операций расширения карты признаков и слоя транспонированной свёртки может оказаться лучшим решением, чем уже существовавшие решения в виде свёрточных нейросетей.

В 2015 году С. Иоффе с командой разрабатывают пакетную нормализацию данных (*Batch Normalization*) при передаче между слоями [63]. Метод основан на том, что нормализация становится частью архитектуры нейросети и выполняется для каждого обучающего мини-пакета, что позволяет использовать более быстрые темпы обучения сетей и снизить требования к инициализации исходных данных, а в некоторых случаях не применять технику борьбы с переобучением Dropout.

О. Роннебергер с коллегами разрабатывают в 2015 году подход к построению архитектуры нейросети и её обучению в задаче сегментации изображений [64], позволяющий эффективно использовать имеющиеся выборки данных. Архитектура основана на двух частях — сжимающей, активно использующей свёрточные слои для извлечения шаблонов из входных данных, и расширяющей, использующей операцию пространственного расширения карты признаков. Связи между этими двумя частями архитектуры позволяют добиться эффективного обучения на относительно меньших наборах входных данных и лучшей сегментации изображений, что и позволило выиграть ISBI cell tracking challenge 2015 с большим отрывом относительно конкурентов (свёрточных нейросетей с подвижным окном).

Ф. Ю и В. Колтун в своей работе [65] предлагают использовать архитектурный модуль нейросети, содержащий операцию разреженной свёртки. Коллектив показывает, что данный модуль можно применять для систематического объединения многомасштабной контекстной информации без потери разрешения, что выливается в повышение точности современных систем семантической сегментации.

В конце 2015 года выходит работа К. Хэ с коллегами [66], в которой предлагается к рассмотрению архитектура Residual Network (*ResNet*) — свёрточная нейросеть с остаточными блоками. Переформулировав слои как обучающие остаточные функции со ссылкой на входные слои, коллектив делает упор на глубину сети, сравнивая до 152 слоёв с 16–19 в архитектуре VGG. Несмотря на количество слоёв, ResNet является сетью с меньшей сложностью. Совокупность остаточных блоков даёт погрешность в 3,57% на эталонном наборе тестов ImageNet. Также данная архитектура побеждает на ILSVRC 2015 в задачах классификации.

Х. Хан и Б. Енер в 2018 году на конференции представляют работу [67], в которой используют слой вейвлет-деконволюции для спектрального разложения временных рядов вместо предобработки сигналов в свёрточных нейросетях, что позволяет уменьшить количество параметров обучения и повысить интерпретируемость классификатора временных рядов. Данный метод позволил уменьшить ошибку в распознавании телефонных сигналов на 4% до 18,1%.

С. Фудзиеда с коллегами представили в своей работе [68] использование вейвлет-преобразований непосредственно в нейросети в качестве вейвлет-нейрослоёв субдискретизации. Такой подход позволяет использовать спектральную информацию, обычно теряющуюся в обычных свёрточных нейросетях, для эффективного решения задач классификации текстур и аннотации 2D-изображений.

В 2019 году П. Лю с коллегами предлагает использование [69] мультивейвлет-свёрточной нейросети, архитектура которой включает встройку вейвлет-преобразований в нейросеть для уменьшения карт признаков и увеличить поле восприятия соответствующих фильтров. Также данную архитектуру можно применять для восстановления карт объектов с высоким разрешением с использованием обратных вейвлет-преобразований в архитектуре.

6. Глубокие рекуррентные нейросети

В конце 1980-х — 1990-х годах М. Джордан и Дж. Элман внесли фундаментальный вклад в развитие рекуррентных нейронных сетей (*RNN*). В 1990 году Элман в своей работе [70] предлагает модель рекуррентной ИНС с обратной связью и наличием одношагового временного контекста как обобщение высказанных идей различными исследователями в 86-90-х годах, первым из которых был Джордан, говоривший про временной контекст в техническом отчёте [71]:

$$h_k = g_h(W_h x_k + U_h h_{k-1} + b_h),$$

$$y_k = g_y(W_y h_k + b_y),$$

где k — дискретное время, h_k — вектор скрытого состояния нейросети в момент времени k, а $U_h h_{k-1}$ отвечает за обратную связь и временной контекст.

В 1997 году Джордан предложил модификацию сети Элмана [70] в работе [72] по изучению коартикуляционных явлений в речи, которая заключается в том, что контекст решения определяется выходом сети, а не скрытым слоем:

$$h_k = g_h(W_h x_k + U_h y_{k-1} + b_h),$$

$$y_k = g_y(W_y h_k + b_y),$$

Такие сети как в [72] и [70] называются SimpleRNN и имеют некоторые проблемы. Однако рекуррентные сети, состоящие из стандартных рекуррентных ячеек, не способны обрабатывать долгосрочные зависимости: по мере увеличения разрыва между соответствующими входными данными становится трудно получить информацию о соединении. А сигналы об ошибках, поступающие в обратном направлении во времени, имеют тенденцию либо усиливаться, либо исчезать [73].

Для решения проблем SimpleRNN в 1997 году З. Хохрайтер предлагает архитектуру LSTM — Long Short Term Memory [74], или же долгая краткосрочная память. Они улучшили запоминающую способность стандартной рекуррентной ячейки, введя в нее шлюзы (*gates*). После этой новаторской работы LSTM были модифицированы и популяризированы многими исследователями. Варианты включают LSTM без шлюза забывания, LSTM с шлюзом забывания и LSTM с подключением через глазок. Обычно термин "ячейка LSTM" обозначает LSTM со шлюзом забывания [73].

Ф. Морин и Й. Бенгиа в 2005 году в своей работе про языковое моделирование распознавателей речи [75] предлагают использование иерархической декомпозиции условных вероятностей отношений языковых конструкций к определённым классам семантической иерархии WordNet, что по сути является модификацией нейрослоя Softmax. Это позволяет эффективно решать задачи классификации в компьютерной лингвистике с десятками тысяч классов.

А. Гравес и Ю. Шмидбухер в своей работе 2005 года [76] совершенствуют архитектуру LSTM и представляют архитектуру сети Bidirectional LSTM. Тестируя различные архитектуры нейросетей, исследователи делают вывод о том, что двунаправленные сети превосходят однонаправленные, а архитектура LSTM в целом превосходит обычные RNN сети.

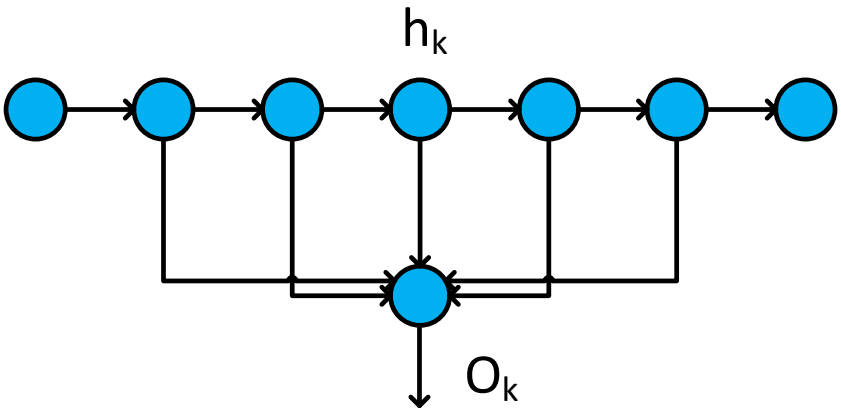

Рисунок 7. Слой внимания (*Attention layer*). Выходное значение формируется как взвешенное среднее выхода.

В 2013 году Гравес [77] предложил реализацию слоя внимания (*Attention layer*). Выход слоя внимания является взвешенным средним выхода рекуррентного слоя (рис. 7).

С. Ши с соавторами в 2015 году представляют в работе [78] свёрточную LSTM нейросеть — инновационную архитектуру, объединяющую преимущества свёрточных и LSTM сетей для краткосрочного прогнозирования осадков. Данная работа демонстрирует

успешное применение глубокого обучения в метеорологии и показывает эффективные методы обработки пространственно-временных данных.

В 2015 году Н. Кальхбреннер в работе [79] представляет архитектуру Grid LSTM, расширяющую возможности LSTM для работы с входными данными, имеющими сетчатую структуру. Отличия от традиционной долгой кратковременной памяти заключается в связи ячеек между нейросетевыми уровнями, а также в пространственно-временных данных. Превосходство архитектуры показано на наборе эталонных тестов предсказания символов Википедии, в задаче перевода текста с китайского на английский. На бенчмарке MNIST для определения рукописных символов архитектура показала конкурентоспособный процент ошибки.

К. Лаурент с коллегами в своей работе 2015 года показывает как пакетная нормализация позволяет значительно сократить время обучения сети [80]. Коллектив отмечает, что хотя эта техника может привести к более быстрой сходимости критериев обучения, как такового преимущества при решении задач языкового моделирования и распознавания речи не даёт.

Д. Амодей с соавторами [81] добиваются ускорения работы нейросети для распознавания английской и китайской речи в своей работе. Ключом к повышению эффективности является использование архитектур нейросетей, которые позволяют применение схемы формирования пакетов для графических процессоров. Применение данной схемы, называющейся диспетчеризацией пакетов (*Batch Dispatch*), позволило добиться эффективной работы программного пакета в реальном времени на серверах разработки.

В 2016 году Дж. Л. Ба публикует работу [82], в которой метод пакетной нормализации модифицируется в метод нормализации слоя. Отличием служит применение адаптивных смещений нейронов до применения элементов нелинейности. Также схожие операции происходят и во время обучения и тестирования нейросети. Такой подход эффективен для стабилизации динамики скрытых состояний в рекуррентных нейросетях и может значительно сократить время, требуемое для обучения сети.

А. Васвани с коллегами представляют в своей работе [83] архитектуру Transformer, которая с помощью механизма многоголового внимания (*Multi-head attention*) может эффективно моделировать зависимости между данными без использования рекуррентных или свёрточных нейрослоёв, заменяя их массивами ячеек внимания. Данная архитектура показала значительные улучшения показателей в задачах обработки естественного языка и эффективность обучения в условиях ограниченности обучающих данных. Однако у

данной сети есть недостаток в виде ограниченности операционного контекста — всего несколько десятков токенов.

М. Дехгани и С. Гувс в 2018 году представляют архитектуру Universal Transformer, которая является обобщением архитектуры сети-трансформера с использованием рекуррентных последовательностей, что позволяет преодолеть ограничения сетей-трансформеров во многих простых задачах [84].

Работа Дж. Девлина с соавторами 2018 года [85] представляет модель BERT (*Bidirectional Encoder Representations from Transformers*), которая предлагает новый подход к предварительному обучению языковых моделей, учитывающий левый и правый контекст во всех слоях. Архитектура показала свою эффективность в задачах обработки естественного языка [86], а операционный контекст в несколько сотен токенов позволяет преодолеть недостатки LSTM.

В 2019 году Ц. Даи с коллегами представляет в [87] архитектуру нейросети Transformer-XL, которая позволяет изучать зависимости за пределами фиксированной длины без нарушения временной согласованности. Улучшение составляет на 80% в случае RNN, и 450% для обычных нейросетей архитектуры Transformer.

Современное состояние архитектур рекуррентных нейронных сетей позволяет решать широкий класс задач [88, 89, 90].

7. Современные тенденции глубокого обучения

7.1 Обработка естественного языка

Одним из современным направлений развития является улучшение архитектуры больших языковых моделей. Развитие рекуррентных нейронных сетей, ячеек долгой краткосрочной памяти и архитектуры трансформеров позволило развиваться таким семействам нейросетей как вышеописанная BERT [85], модель от Google Brain Team, названная T5 [91, 92], или ChatGPT [93]. Принципиальным вкладом T5 в развитие нейросетей является унифицирование задачи обработки естественного языка в единую схему преобразования текста. В отличие от существовавших до неё моделей, в том числе BERT, T5 обрабатывает входной текст (*энкодер*) и на его основе генерирует выходной текст (*декодер*).

Развитие ChatGPT демонстрирует впечатляющий прогресс в области искусственного интеллекта и обработки естественного языка. Первой архитектурой для ChatGTP была GPT-1, появившаяся в 2018 году и представлявшая из себя архитектуру трансформера, умеющего

генерировать простые тексты, отвечать на несложные вопросы, завершать предложения, генерировать небольшие описания на основе описания характеристик. Но GPT-1 имела недостатки в виде ограниченной длины генерации текстов, слабое понимание контекста, сложности с решением сложных задач, несвязное ведение диалога с пользователем. По-настоящему революционной стала архитектура GPT-3, которая колоссально увеличила количество параметров модели — 175 миллиардов, значительно улучшила качество генерации и уменьшила выборки данных для обучения. В 2022 году появился сам ChatGPT, расширивший возможности пользовательского интерфейса, а на текущий момент активно используется архитектура GPT-4o в комбинации с GPT-5, вышедшей в релиз 7 августа 2025 года.

В 2022 году Google AI представляет архитектуру Pathways Language Model (*PaLM*) [94]. Нейросеть представляет собой языковую модель-трансформер с 540 миллиардами параметров, а в обучении использовалась Pathways, новая система ML, которая обеспечивает высокоэффективное обучение в нескольких модулях тензорных процессоров Google. PaLM обладает широкими возможностями в решении многоязычных задач и генерации исходного кода. В работе [94] продемонстрировано превосходство модели над GPT-3.

Нейросети-трансформеры нашли своё применение и в задаче определения белков, где позволили улучшить существующие модели. Дж. Джампер с коллегами из DeepMind представили AlphaFold — первую нейросеть, способную определять последовательности белков с атомной точностью [95]. Успехи были подтверждены на конкурсе 14th Critical Assessment of protein Structure Prediction (*CASP14*). AlphaFold открыла новую эру в предсказании белковых структур, значительно ускорив процесс, который ранее занимал месяцы и годы расчётных и экспериментальных исследований.

П. Левис с коллегами в 2020 году представляют универсальный рецепт тонкой настройки моделей генерации с расширенным поиском (*RAG*), которые объединяют предварительно обученную параметрическую и непараметрическую память для генерации языка [96]. Анализируя данный метод настройки в широком спектре задач обработки естественного языка, команда делает вывод, что модели RAG генерируют более конкретный, разнообразный и основанный на фактах язык, чем базовая версия модели seq2seq, основанная только на параметрах. Данный принцип впоследствии стал использоваться многими коммерческими решениями, по крайней мере специальными версиями моделей из семейства архитектур, таких как GPT, DALL-E, Claude, Gemini, Copilot и другие.

7.2 Компьютерное зрение

А. Досовитский с соавторами в работе [97] рассуждают о том, что архитектура сетей-трансформеров стала стандартом для задач обработки естественного языка, а в компьютерном зрении применение архитектуры остается ограниченным. Механизм внимания применяется либо совместно со свёрточными сетями, либо используется для замены определённых компонентов свёрточных сетей при сохранении их общей структуры. Команда демонстрирует, что зависимость от свёрточных нейросетей необязательна, и чистый преобразователь, применяемый непосредственно к последовательностям фрагментов изображений, хорошо справляется с задачами классификации изображений.

Р. Ромбах с соавторами представляют в 2021–2022 годах скрытую диффузионную модель Stable Diffusion для генерации изображений по текстовому описанию [98]. Команда улучшает диффузионные модели, основанные на декомпозиции процесса формирования изображения на последовательное применение автоэнкодеров (*энкодер* + *декодер*) с шумоподавлением, с помощью использования скрытого пространства мощных предварительно обученных автоэнкодеров и внедрения в архитектуру модели уровней перекрёстного внимания. Архитектура стала первой высокопроизводительной моделью с открытым кодом в области генеративного ИИ.

Конкурирующей архитектурой является DALL-E 2 — генеративная модель от OpenAI, представленная в 2022 году, способная создавать реалистичные изображения и произведения искусства на основе текстовых описаний. К сожалению, OpenAI не публикует полные научные статьи о DALL-E 2 в открытом доступе, поэтому преимущества в виде более высокого качества финальных изображений, лучшего понимания контекста запроса, меньшее число дефектов при генерации и более стабильные результаты в сравнении со Stable Diffusion можно считать субъективными, однако из рассмотрения исключить саму нейросеть не позволяют. Ещё одним коммерческим конкурентом является Midjourney от одноимённой компании. Нейросети приписывают преимущество при генерации изображений в художественном стиле, однако закрытый характер разработки не позволяет полноценно провести анализ модели.

В области компьютерного зрения NVIDIA в 2022 году предлагают свою разработку StyleGAN v4 — генеративно-состязательную сеть (*GAN*), предназначенную для создания фотореалистичных изображений лиц и других объектов. Архитектура включает основанный на грамматике обучения основанной лексике (*G2L2*) подход к изучению композиционного и обоснованного представления значений языка на основе обоснованных данных, таких как парные изображения и тексты. В ходе работы сети слова сопоставляются с кортежем

синтаксического типа и нейросимволической семантической составляющей. Модель продолжает развиваться, открывая новые возможности для создания фотореалистичного контента.

В области компьютерного зрения можно отметить CLIP (*Contrastive Language-Image Pre-training*) [99]. В 2021 году А. Радфорт с коллегами демонстрируют, что предварительное обучение определения соответствия подписей и изображений является эффективным и масштабируемым способом изучения представлений изображений с нуля на основе набора данных из 400 миллионов пар (изображение, текстовое описание), собранных из Интернета. Модель нетривиально подходит для большинства задач и часто конкурирует с полностью контролируемыми базовыми показателями без необходимости какого-либо обучения для конкретного набора данных.

В 2023 году Meta AI представляет разработку Segment Anything Model [100], предназначенную для автоматической сегментации объектов на изображениях. Модель была обучена для обработки как текстовых запросов, так и изображений, при этом её важной особенностью является способность выполнять инструкции в условиях отсутствия примеров (*zero-shot*) для новых изображений и задач. Производительность при этих условиях впечатляет — часто результат не уступает или даже превосходит полученный при наличии полностью размеченных аналогичных примеров в обучающей выборке.

7.3 Вопросы архитектуры и масштабирования нейросетей

В 2020 году Д. Лепихин с соавторами обозначают проблему, связанную с масштабированием нейросетей [101]. Несмотря на надёжное повышение качества моделей с масштабированием, растёт сложность вычислений, программирования и эффективного распараллеливания. Для преодоления этих проблем команда использует архитектурное решение Mixture of Experts. С помощью модуля Gshard, состоящего из набора облегченных API-интерфейсов аннотаций и расширения для компилятора XLA, команда добивается расширения многоязычной модели нейронного машинного перевода архитектуры Transformer до 600 миллиардов параметров, используя автоматическое сегментирование. Модель обучалась 4 дня на 2048 тензорных процессорах и показала высокое качество перевода.

В 2019 году М. Тан и К. В. Ле публикуют работу про семейство архитектур EfficientNet [102], переосмысляющую масштабирование нейросетей. Тщательный баланс глубины, ширины и разрешения сети при масштабировании может привести к повышению производительности. Основываясь на этом наблюдении, команда предлагает новый метод,

который равномерно масштабирует все параметры глубины, ширины и разрешения с использованием простого, но высокоэффективного комплексного коэффициента. Также Тан и Ле используют поиск по нейронной архитектуре для разработки новой базовой сети и её масштабирования с целью получения нового семейства моделей, называемых EfficientNet.

В 2019 году Э. Ховард с командой представляют семейство архитектур MobileNetV3 [103]. Команда представляет новое поколение нейронных мобильных сетей, адаптированное к процессорам мобильных телефонов с помощью аппаратного обеспечения для поиска сетевой архитектуры (*NAS*), дополненного алгоритмом NetAdapt, а затем усовершенствованного за счет новых достижений в архитектуре. Данное решение представляет собой эффективное решение для задач компьютерного зрения на мобильных устройствах, предлагая баланс между точностью и производительностью.

Дж. Расли с соавторами публикуют в 2020 году статью про библиотеку DeepSpeed от Microsoft [104] для предназначенную для масштабирования и оптимизации обучения глубоких нейронных сетей на графических и тензорных процессорах. Одна из частей библиотеки — параллельный оптимизатор ZeRO, который значительно снижает ресурсы, необходимые для распараллеливания модели. При этом ZeRO увеличивает количество параметров, которые можно обучить. DeepSpeed предоставляет возможности для обучения моделей с триллионами параметров.

7.4 Обучение с подкреплением

В 2018 году DeepMind представили нейросеть AlphaZero, представляющую собой архитектуру, способную достигать сверхчеловеческого уровня игры в различные стратегические игры на основе обучения с подкреплением — метода машинного обучения, при котором нейросеть учится принимать оптимальные решения через взаимодействие с окружающей средой, получая обратную связь в виде наград или наказаний [105]. В отличие от AlphaGo [106], обучавшейся в том числе на размеченных данных игр профессионалов, AlphaZero, имея в своём арсенале только правила игры, за 24 часа достигла сверхчеловеческого уровня игры в шахматы и сёги, а также в го, и в каждом случае убедительно побеждала программы-чемпионки мира.

В 2019 году О. Винялс с соавторами публикуют работу [107] про нейросеть AlphaStar от DeepMind, достигшей уровня игры Грандмастеров в StarCraft II. Для обучения использовался многоагентный алгоритм обучения с подкреплением, который использует данные как из игр с участием людей, так и других нейросетей в рамках разнообразной лиги постоянно адаптирующихся стратегий и контрстратегий, каждая из которых представлена глубокими

нейронными сетями. Серия онлайн-игр против игроков-людей показала, что рейтинг AlphaStar был на уровне гроссмейстера для всех трех рас StarCraft и превышал рейтинг 99,8% официально зарегистрированных игроков-людей.

В 2020 году выходит статья Дж. Шритвайзера с соавторами про нейросеть MuZero от DeepMind [108], обобщающую идеи предыдущих разработок и концентрирующуюся на разработке победных стратегий без знаний про игру, полагаясь на динамику изменения окружения. MuZero при итеративном применении предсказывает величины, непосредственно относящиеся к планированию: вознаграждение, политику выбора действий и функцию ценности. При оценке в Го, шахматах и сёги, без какого-либо знания правил игры, MuZero соответствовал сверхчеловеческой производительности алгоритма AlphaZero, который был применён вместе с правилами игры.

В 2022 году DeepMind представляет нейросеть AlphaTensor [109], использующую методы обучения с подкреплением для автоматического открытия новых алгоритмов умножения матриц. Архитектура основана на AlphaZero и предлагает существенные улучшения в эффективности умножения матриц для различных размеров, в том числе улучшение алгоритма Штрассена для матриц размера 4×4.

7.5 Мультимодальные модели

Мультимодальными нейросетями называются системы искусственного интеллекта, способные обрабатывать несколько типов данных одновременно: текст, изображения, аудио, видео и другие форматы информации. Формально к ним можно отнести много нейросетей из больших языковых моделей, сетей для компьютерного зрения и других, но акцент на мультимодальности стали делать с 2022–2023 года.

В 2022 году статья С. Рида с соавторами представляет разработку DeepMind мультимодального агента Gato [110], работа которого выходит за рамки текстового вывода. Gato работает как мультимодальная, многозадачная и многоцелевая универсальная нейросеть. Одна и та же сеть с одинаковыми весами может играть в игры, подписывать изображения, общаться в чате, складывать блоки с помощью настоящей руки робота и многое другое в зависимости от контекста.

Нейросеть Flamingo — ещё одна разработка DeepMind — освещена в статье Дж.-Б. Алайрака с соавторами [111]. Flamingo представляет собой семейство моделей визуального языка, обладающих возможностью быстро адаптироваться к новым задачам, используя всего несколько примеров с аннотациями. Предлагаются архитектурные инновации, позволяющие объединить мощные предварительно обученные визуальные и

языковые модели, обрабатывать последовательности произвольно чередующихся визуальных и текстовых данных и легко использовать изображения или видео в качестве входных данных. Благодаря своей гибкости сети Flamingo могут обучаться на крупномасштабных мультимодальных веб-ресурсах, содержащих произвольно чередующийся текст и изображения, что является ключевым фактором для обеспечения их возможностями обучения в режиме реального времени.

В 2023 году DeepMind представляет семейство сетей Gemini [112]. Семейство мультимодальных моделей Gemini обладает замечательными возможностями для понимания изображений, аудио, видео и текста. Новые возможности семейства Gemini в области кросс-модального мышления и понимания языка позволят использовать их в самых разнообразных случаях.

Р. Сан с коллегами публикует обзор на основе разработки OpenAI генеративной модели Sora, предназначенная для создания высококачественных видео на основе текстовых описаний [113]. Команда делает вывод, что продукт является важной вехой на пути к созданию общего искусственного интеллекта.

Развитие нейросетей-трансформеров и других современных архитектур заслуживает отдельного обзора [30]. В данной работе мы не будем подробно рассматривать их развитие, поскольку оно связано в большей степени с их масштабируемостью, нежели с новыми концептуальными идеями.

7.6 Сети Колмогорова-Арнольда

Теорема суперпозиции Колмогорова-Арнольда легла в основу так называемых сетей Колмогорова-Арнольда. В отличие от многослойных персептронов, лежащих в основе всех современных ИНС, KAN имеют фиксированные функции активации на узлах и обучаемые функции активации на ребрах (рис. 8). Каждый функция активации является одномерной функцией, параметризованной в виде сплайна. Это изменение позволяет KAN превосходить MLP с точки зрения точности и интерпретируемости в ряде задач [31]. KAN являются многообещающей альтернативой MLP, открывая возможности для дальнейшего совершенствования современных моделей глубокого обучения, которые в значительной степени зависят от MLP.

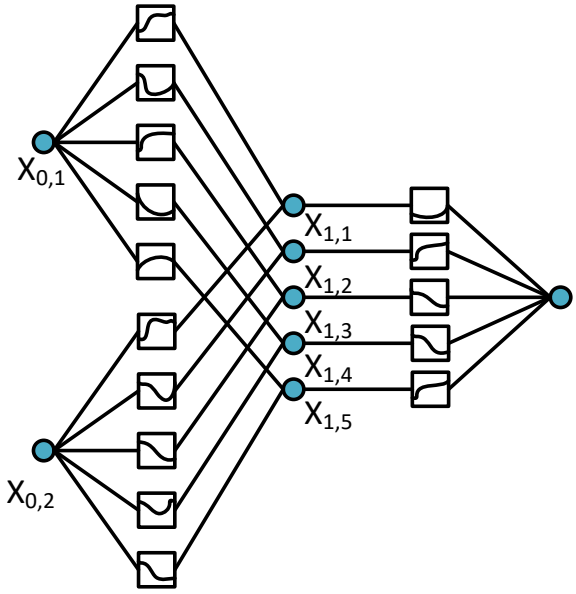

Рисунок 8. Схема архитектуры сети Колмогорова-Арнольда. В отличие от MLP, активационные функции присутствуют как в нейронах, так и в рёбрах, их связывающих.

Для задач аппроксимации при помощи KAN существует теорема Майорова-Пинкуса [114]:

Существует аналитическая вещественная строго монотонно возрастающая функция активации σ, удовлетворяющая следующему свойству. $\forall f \in C[0,\ 1]^d$и $\varepsilon > 0$ существуют вещественные константы $d_i,\ c_{ij},\ \theta_{ij},\ \gamma_i$ и вектора $\boldsymbol{w}^{ij} \in \mathbb{R}$, для которых

$$\left| f(x) - \sum_{i=1}^{6d+3} d_i \sigma \left(\sum_{j=1}^{3d} c_{ij} \sigma\left(w^{ij} \cdot x + \theta_{ij}\right) + \gamma_i \right) \right| < \varepsilon$$

$\forall x \in [0,\ 1]^d$.

Согласно теореме Майорова-Пинкуса, существует класс функций, требующих большого размера нейронной сети, причём с ростом точности масштабируемость растёт экспоненциально. Также теорема определяет нижнюю границу сложности аппроксимации.

С. А. Немковым [115] было обосновано, что для задач со сложными нелинейными зависимостями KAN являются предпочтительной архитектурой в сравнении с MLP. Это связано с тем, что MLP аппроксимируют результат кусочно-линейными функциями, что приводит к необоснованному росту числа параметров от каждого сегмента. В то же время KAN обладают индуктивным сдвигом в сторону разложения сложных зависимостей на

одномерные гладкие функции. В ряде задач это помогает не только достигнуть результата, но и получить дополнительную информацию для анализа.

8. Заключение

Основой современных нейронных сетей является многослойный персептрон. Входные сигналы могут иметь любое числовое значение, а итоговое значение в узле получается суммированием произведений входных сигналов с весами рёбер. Также в каждом узле присутствует активационная функция, которая превращает линейный функционал в произвольный, добавляя элементы нелинейности.

Структура организации искусственных нейронов, связей между ними и способы обработки информации внутри сети может быть гораздо сложнее, чем полносвязные слои нейронов. Для разных задач лучше подходят разные архитектуры и, зачастую, выбор архитектуры обусловлен какой-то эвристикой или даже просто эмпирически.

Среди современных тенденций развития наибольший интерес представляют большие языковые модели, сети-трансформеры, мультимодальные модели, среди представителей которых наиболее известны chatGPT, BERT, Cursor, DeepSeek и другие.

Также развивается альтернатива многослойным персептронам — сети Колмогорова-Арнольда, содержащие не только фиксированные активационные функции в узлах, но и обучаемые активационные функции на рёбрах.

Список литературы

1. McCarthy J., Minsky M., Rochester N., Shannon C. A proposal for the Dartmouth summer research project on artificial intelligence : technical report. Hanover : Dartmouth College, 1956. 13 p.
2. von Neumann J., Burks A. W. Theory of Self-Reproducing Automata. Urbana : University of Illinois Press, 1966. 322 p.
3. Müller V. C., Bostrom N. Future progress in artificial intelligence: a survey of expert opinion // Fundamental Issues of Artificial Intelligence. Cham : Springer International Publishing, 2016. P. 555–572.
4. Hebb D. O. The Organization of Behavior. New York : John Wiley & Sons, 1949. 335 p.
5. Minsky M., Papert S. Perceptrons: An Introduction to Computational Geometry. Cambridge, MA : MIT Press, 1969. 258 p.
6. Artificial General Intelligence / ed. by B. Goertzel, C. Pennachin. Berlin ; Heidelberg : Springer, 2007. 445 p.
7. Garibotto G. A binocular license plate reader for high precision speed measurement // Journal of Intelligent Transportation Systems. 2001. Vol. 6, no. 1. P. 35–48.

8. Мишинов С. В., Русских Н. Е., Строганов М. С. и др. Использование подходов машинного обучения для воссоздания утраченной части костей черепа // Российский нейрохирургический журнал им. проф. А. Л. Поленова. 2023. Т. 15, спец. вып. 1. С. 17. EDN YJJXPW.
9. Мишинов С. В., Гутт А. А., Пушкина Е. В. и др. Опыт применения подходов машинного обучения в нейрохирургии // Третий Сибирский нейрохирургический конгресс : сб. тезисов / под ред. Д. А. Рзаева. Новосибирск : ООО «Семинары, Конференции и Форумы», 2022. С. 56–57. EDN DOXLVA.
10. Дорошков А. В., Арсенина С. И., Пшеничникова Т. А. и др. Применение компьютерного анализа микроизображений листа для оценки характеристик опушения пшеницы Triticum aestivum L. // Информационный вестник ВОГиС. 2009. Т. 13, № 1. С. 218–226. EDN KUXGKZ.
11. Sidorova E. A., Akhmadeeva I. R., Kononenko I. S. et al. Argument extraction based on the indicator approach // Pattern Recognition and Image Analysis. 2023. Vol. 33, no. 3. P. 498–505.
12. Кожаринов А. С., Кириченко Ю. А., Афанасьев И. В. и др. Методы анализа когнитивных искажений и концепция автоматизированной интеллектуальной системы их детектирования // Нейрокомпьютеры: разработка, применение. 2022. Т. 24, № 4. С. 39–74.
13. Сидорова Е. А., Загорулько Ю. А., Кононенко И. С. и др. Подход к построению датасета для задачи извлечения аргументативных отношений // Двадцать первая Национальная конференция по искусственному интеллекту с международным участием КИИ-2023 (Смоленск, 16–20 октября 2023 г.) : труды конференции : в 2 т. Т. 1. Смоленск : Принт-Экспресс, 2023. С. 211–222.
14. Loukachevitch N., Manandhar S., Baral E. et al. Nerel-bio: a dataset of biomedical abstracts annotated with nested named entities // Bioinformatics. 2023. Vol. 39, no. 4.
15. Bruches E., Mezentseva A., Batura T. A system for information extraction from scientific texts in Russian // Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2021). Cham : Springer, 2022. P. 234–245. (Communications in Computer and Information Science ; vol. 1620).
16. Loukachevitch N., Artemova E., Batura T. et al. Nerel: a Russian information extraction dataset with rich annotation for nested entities, relations, and Wikidata entity links // Language Resources and Evaluation. 2023. Vol. 57, no. 3.
17. St Laurent G., Savva Y. A., Maloney R. et al. Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in Drosophila // Nature Structural & Molecular Biology. 2013. Vol. 20, no. 11. P. 1333–1339. EDN RXCSLX.
18. Вяткин Ю. В., Антонец Д. В., Шабурова Е. В. и др. Языковые модели в изучении белков // 11-я Московская конференция по вычислительной молекулярной биологии MCCMB’23 : материалы конференции. Москва : ИППИ РАН, 2023. EDN AAFHRE.
19. Borg M., Hewett D., Hagatulah N. et al. Echoes of AI: investigating the downstream effects of AI assistants on software maintainability. 2025.

20. Костарева И. Нейросеть успешно сдала ЕГЭ: что это значит для системы образования? [Электронный ресурс]. 2023. URL: (дата обращения: …).
21. Peng C., Yang X., Chen A. et al. A study of generative large language model for medical research and healthcare // npj Digital Medicine. 2023. Vol. 6. Art. 210.
22. Nechesov A. V., Kondratyev D. A., Sviridenko D. I. et al. Conceptual framework for trustworthy artificial intelligence: combining large language models with formal logic systems // System Informatics. 2025. No. 27. P. 93–118.
23. Kalai A. T., Nachum O., Vempala S. S. et al. Why language models hallucinate : technical report. San Francisco : OpenAI, 2025. Sept.
24. Abbasi Yadkori Y., Kuzborskij I., Stutz D. et al. Mitigating LLM hallucinations via conformal abstention. 2024.
25. Li X., Dong X. L., Lyons K. B. et al. Truth finding on the deep web: is the problem solved? // Proceedings of the VLDB Endowment. 2012. Vol. 6, no. 2. P. 97–108.
26. Yao L. et al. Online truth discovery on time series data // Proceedings of the SIAM International Conference on Data Mining. Philadelphia, PA : Society for Industrial and Applied Mathematics, 2018. P. 162–170.
27. Li Y. et al. On the discovery of evolving truth // Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’15). New York : ACM, 2015. P. 675–684.
28. Андриевский Б. Р., Балашов М. В., Бахтадзе Н. Н. и др. Теория управления: дополнительные главы : учеб. пособие. М. : ЛЕНАНД, 2019. 512 с.
29. Макаренко А. В. Глубокие нейронные сети: зарождение, становление, современное состояние // Проблемы управления. 2020. № 2. С. 3–19.
30. Xu P., Zhu X., Clifton D. A. Multimodal learning with transformers: a survey. 2023.
31. Liu Z., Wang Y., Vaidya S. et al. KAN: Kolmogorov–Arnold networks. 2025.
32. McCulloch W., Pitts W. A logical calculus of the ideas immanent in nervous activity // Bulletin of Mathematical Biophysics. 1943. Vol. 5. P. 115–133.
33. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain // Psychological Review. 1958. Vol. 65, no. 6. P. 386–408.
34. Parallel distributed processing: explorations in the microstructures of cognition / ed. by D. E. Rumelhart, J. L. McClelland. Cambridge, MA : MIT Press, 1986. Vol. 1–2.
35. Галушкин А. И. Синтез многослойных систем распознавания образов. М. : Энергия, 1974. 368 с.
36. Werbos P. J. Beyond regression: new tools for prediction and analysis in the behavioral sciences : PhD thesis. Cambridge, MA : Harvard University, 1974.
37. Барцев С. И., Охонин В. В. Адаптивные сети обработки информации : препринт № 59Б. Красноярск : Ин-т физики СО АН СССР, 1986. 45 с.

38. Rumelhart D. E., Hinton G. E., Williams R. J. Learning internal representations by error propagation // Parallel Distributed Processing. Cambridge, MA : MIT Press, 1986. Vol. 1. P. 318–362.
39. Арнольд В. И. О представлении функций нескольких переменных в виде суперпозиции функций меньшего числа переменных // Математика, её преподавание, приложения и история. 1958. Т. 3. С. 41–61.
40. Cybenko G. Approximation by superpositions of a sigmoidal function // Mathematics of Control, Signals and Systems. 1989. Vol. 2, no. 4. P. 303–314.
41. Siegelmann H., Sontag E. Turing computability with neural nets // Applied Mathematics Letters. 1991. Vol. 4, no. 6. P. 77–80.
42. Hornik K. Approximation capabilities of multilayer feedforward networks // Neural Networks. 1991. Vol. 4, no. 2. P. 251–257.
43. Funahashi K., Nakamura Y. Approximation of dynamical systems by continuous time recurrent neural networks // Neural Networks. 1993. Vol. 6, no. 6. P. 801–806.
44. Chow T., Li X. Modeling of continuous time dynamical systems with input by recurrent neural networks // IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications. 2000. Vol. 47, no. 4. P. 575–578.
45. Bengio Y., Simard P., Frasconi P. Learning long-term dependencies with gradient descent is difficult // IEEE Transactions on Neural Networks. 1994.
46. Hinton G., Salakhutdinov R. Reducing the dimensionality of data with neural networks // Science. 2006. Vol. 313, no. 5786. P. 504–507.
47. Nair V., Hinton G. Rectified linear units improve restricted Boltzmann machines // Proceedings of the International Conference on Machine Learning. 2010. P. 807–814.
48. Glorot X., Bengio Y. Understanding the difficulty of training deep feedforward neural networks // Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. 2010. Vol. 9. P. 249–256.
49. Lecun Y., Boser B., Denker J. et al. Backpropagation applied to handwritten zip code recognition // Neural Computation. 1989. Vol. 1, no. 4. P. 541–551.
50. Lecun Y., Bottou L., Bengio Y. et al. Gradient-based learning applied to document recognition // IEEE Intelligent Signal Processing. 1998. P. 306–351.
51. Ciresan D., Meier U., Gambardella L. et al. Deep big simple neural nets excel on handwritten digit recognition. arXiv preprint. 2010. arXiv:1003.0358.
52. Zeiler M. D., Krishnan D., Taylor G. W. Deconvolutional networks // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, 2010. P. 2528–2535.
53. Krizhevsky A., Sutskever I., Hinton G. Imagenet classification with deep convolutional neural networks // Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS’12). 2012. Vol. 1. P. 1097–1105.

54. Sänchez J., Perronnin F. High-dimensional signature compression for large-scale image classification // 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2011. P. 1665–1672.
55. Hinton G., Srivastava N., Krizhevsky A. et al. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint. 2012. arXiv:1207.0580.
56. Lin M., Chen Q., Yan S. Network in network. arXiv preprint. 2013. arXiv:1312.4400.
57. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint. 2014. arXiv:1409.1556.
58. Szegedy C., Liu W., Jia Y. et al. Going deeper with convolutions. arXiv preprint. 2014. arXiv:1409.4842.
59. Sifre L. Rigid-motion scattering for image classification : PhD thesis. Palaiseau : École Polytechnique, CMAP, 2014.
60. Long J., Shelhamer E., Darrell T. Fully convolutional networks for semantic segmentation. arXiv preprint. 2014. arXiv:1411.4038.
61. Dumoulin V., Visin F. A guide to convolution arithmetic for deep learning. arXiv preprint. 2016. arXiv:1603.07285.
62. Odena A., Dumoulin V., Olah C. Deconvolution and checkerboard artifacts // Distill. 2016. Vol. 1.
63. Ioffe S., Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint. 2015. arXiv:1502.03167.
64. Ronneberger O., Fischer P., Brox T. U-net: convolutional networks for biomedical image segmentation. arXiv preprint. 2015. arXiv:1505.04597.
65. Yu F., Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv preprint. 2015. arXiv:1511.07122.
66. He K., Zhang X., Ren S. et al. Deep residual learning for image recognition. arXiv preprint. 2015. arXiv:1512.03385.
67. Khan H., Yener B. Learning filter widths of spectral decompositions with wavelets // Proceedings of the Neural Information Processing Systems (NIPS) Conference. 2018. P. 4601–4612.
68. Fujieda S., Takayama K., Hachisuka T. Wavelet convolutional neural networks. arXiv preprint. 2018. arXiv:1805.08620.
69. Liu P., Zhang H., Lian W. et al. Multi-level wavelet convolutional neural networks. arXiv preprint. 2019. arXiv:1907.03128.
70. Elman J. Finding structure in time // Cognitive Science. 1990. Vol. 14, no. 2. P. 179–211.
71. Jordan M. I. Serial order: a parallel distributed processing approach : report 8604. San Diego : Institute for Cognitive Science, University of California, 1986.
72. Jordan M. Serial order: a parallel distributed processing approach // Advances in Psychology. 1997. No. 121. P. 471–495.
73. Yu Y., Si X., Hu C. et al. A review of recurrent neural networks: LSTM cells and network architectures // Neural Computation. 2019. Vol. 31, no. 7. P. 1235–1270.

74. Hochreiter S., Schmidhuber J. Long-short term memory // Neural Computation. 1997. Vol. 9, no. 8. P. 1735–1780.
75. Morin F., Bengio Y. Hierarchical probabilistic neural network language model // Proceedings of the AISTATS Conference. 2005. P. 246–252.
76. Graves A., Schmidhuber J. Framewise phoneme classification with bidirectional LSTM networks // International Joint Conference on Neural Networks. 2005. P. 2047–2052.
77. Graves A. Generating sequences with recurrent neural networks. arXiv preprint. 2013. arXiv:1308.0850.
78. Shi X., Chen Z., Wang H. et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. arXiv preprint. 2015. arXiv:1506.04214.
79. Kalchbrenner N., Danihelka I., Graves A. Grid long short-term memory. arXiv preprint. 2015. arXiv:1507.01526.
80. Laurent C., Pereyra G., Brakel P. et al. Batch normalized recurrent neural networks. arXiv preprint. 2015. arXiv:1510.01378.
81. Amodei D., Anubhai R., Battenberg E. et al. Deep Speech 2: end-to-end speech recognition in English and Mandarin. arXiv preprint. 2015. arXiv:1512.02595.
82. Ba J. L., Kiros J. R., Hinton G. E. Layer normalization. arXiv preprint. 2016. arXiv:1607.06450.
83. Vaswani A., Shazeer N., Parmar N. et al. Attention is all you need. arXiv preprint. 2017. arXiv:1706.03762.
84. Dehghani M., Gouws S., Vinyals O. et al. Universal transformers. arXiv preprint. 2018. arXiv:1807.03819.
85. Devlin J., Chang M. W., Lee K. et al. BERT: pretraining of deep bidirectional transformers for language understanding. arXiv preprint. 2018.
86. Sidorova E., Akhmadeeva I., Kononenko I. et al. The role of indicators in argumentative relation prediction // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”. 2023. P. 477–485.
87. Dai Z., Yang Z., Yang Y. et al. Transformer-XL: attentive language models beyond a fixed-length context. arXiv preprint. 2019.
88. Devlin J., Chang M., Lee K. et al. BERT: pre-training of deep bidirectional transformers for language understanding // Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota : Association for Computational Linguistics, 2019. P. 4171–4186.
89. Dosovitskiy A., Beyer L., Kolesnikov A. et al. An image is worth 16×16 words: transformers for image recognition at scale // International Conference on Learning Representations. 2020.
90. Chami I., Abu-El-Haija S., Perozzi B. et al. Machine learning on graphs: a model and comprehensive taxonomy // The Journal of Machine Learning Research. 2022. Vol. 1. P. 3840–3903.

91. Raffel C., Shazeer N., Roberts A. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. 2023.
92. Miftahova A., Pugachev A., Skiba A. et al. NamedEntityRangers at SemEval-2022 task 11: transformer-based approaches for multilingual complex named entity recognition // Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). 2022. P. 1570–1575. URL: https://aclanthology.org/2022.semeval-1.216.pdf
93. Brown T. B., Mann B., Ryder N. et al. Language models are few-shot learners. 2020.
94. Chowdhery A., Narang S., Devlin J. et al. PaLM: scaling language modeling with pathways. 2022.
95. Jumper J., Evans R., Pritzel A. et al. Highly accurate protein structure prediction with AlphaFold // Nature. 2021. Vol. 596, no. 7873. P. 583–589.
96. Lewis P., Perez E., Piktus A. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. 2021.
97. Dosovitskiy A., Beyer L., Kolesnikov A. et al. An image is worth 16×16 words: transformers for image recognition at scale. 2021.
98. Rombach R., Blattmann A., Lorenz D. et al. High-resolution image synthesis with latent diffusion models. 2022.
99. Radford A., Kim J. W., Hallacy C. et al. Learning transferable visual models from natural language supervision. 2021.
100. Kirillov A., Mintun E., Ravi N. et al. Segment anything. 2023.
101. Lepikhin D., Lee H., Xu Y. et al. GShard: scaling giant models with conditional computation and automatic sharding. 2020.
102. Tan M., Le Q. V. EfficientNet: rethinking model scaling for convolutional neural networks. 2020.
103. Howard A., Sandler M., Chu G. et al. Searching for MobileNetV3. 2019.
104. Rasley J., Rajbhandari S., Ruwase O. et al. DeepSpeed: system optimizations enable training deep learning models with over 100 billion parameters. 2020. P. 3505–3506.
105. Silver D., Hubert T., Schrittwieser J. et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. 2017.
106. Silver D., Huang A., Maddison C. J. et al. Mastering the game of Go with deep neural networks and tree search // Nature. 2016. Vol. 529, no. 7587. P. 484–489.
107. Vinyals O., Babuschkin I., Czarnecki W. M. et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning // Nature. 2019. Vol. 575, no. 7782. P. 350–354.
108. Schrittwieser J., Antonoglou I., Hubert T. et al. Mastering Atari, Go, chess and shogi by planning with a learned model // Nature. 2020. Vol. 588, no. 7839. P. 604–609.
109. Fawzi A., Balog M., Huang A. et al. Discovering faster matrix multiplication algorithms with reinforcement learning // Nature. 2022. Vol. 610, no. 7930. P. 47–53.
110. Reed S., Zolna K., Parisotto E. et al. A generalist agent. 2022.

111. Alayrac J.-B., Donahue J., Luc P. et al. Flamingo: a visual language model for few-shot learning. 2022.
112. Gemini Team, Anil R., Borgeaud S. et al. Gemini: a family of highly capable multimodal models. 2025.
113. Sun R., Zhang Y., Shah T. et al. From Sora what we can see: a survey of text-to-video generation. 2024.
114. Maiorov V., Pinkus A. Lower bounds for approximation by MLP neural networks // Neurocomputing. 1999. Vol. 25, no. 1–3. P. 81–91.
115. Немков С. А. Интерпретируемое моделирование смешения красок с помощью KAN // ИТНОУ: Информационные технологии в науке, образовании и управлении. 2025. № 1–2(24–25). С. 37–42.