Программные системы и вычислительные методы
Правильная ссылка на статью:
Корчагин В.Д. — Анализ современных SOTA-архитектур искусственных нейронных сетей для решения задач классификации изображений и детекции объектов // Программные системы и вычислительные методы. – 2023. – № 4. – С. 73 - 87. DOI: 10.7256/2454-0714.2023.4.69306 EDN: MZLZMK URL: https://nbpublish.com/library_read_article.php?id=69306

Анализ современных SOTA-архитектур искусственных нейронных сетей для решения задач классификации изображений и детекции объектов

Корчагин Валерий Дмитриевич

ORCID: 0009-0003-1773-0085

аспирант, кафедра ПИШ ХИМ, образовательной организации федеральное государственное бюджетное образовательное учреждение высшего образования "Российский химико-технологический университет имени Д.И. Менделеева"

127253, Россия, г. Москва, ул. Псковская, д.12, к.1, кв. 159

valerak249@gmail.com

Статья из рубрики "Языки программирования"

DOI:

10.7256/2454-0714.2023.4.69306

EDN:

MZLZMK

Дата направления статьи в редакцию:

12-12-2023

Дата публикации:

31-12-2023

Аннотация: Научное исследование сфокусировано на проведении анализа наиболее эффективных архитектур искусственных нейронных сетей для решения задач классификации изображений и детекции объектов, согласно данным, полученным с открытого портала для публикации результатов проведения эмпирического исследования собственного алгоритма или применения существующих решений для решения альтернативного перечня задач. Актуальность исследования опирается на растущий интерес к технологиям машинного обучения и регулярного улучшения существующих и разработке инновационных алгоритмов компьютерного зрения. Предметом анализа выступают структурные особенности существующих архитектур

нейронных сетей. В частности, наиболее эффективные подходы, используемые в современных архитектурах, позволяющие достигать рекордных показателей в рамках используемых метрик качества, а также ключевые недостатки существующих подходов. Исследуется временной интервал, затрачиваемый как на обучение модели, так и на получение итогового результата. В рамках данной статьи было проведено аналитическое исследование преимуществ и недостатков существующих решений, рассмотрены передовые SOTA архитектурные решения. Изучены наиболее эффективные подходы, обеспечивающие повышение точности базовых моделей. Определено количество используемых параметров, величина обучающей выборки, точность модели, её размер, адаптивность, сложность и требуемые вычислительные ресурсы для обучения отдельно взятой архитектуры. В рамках настоящей исследовательской работы была осуществлена детальная аналитика внутренней структуры наиболее эффективных архитектур нейронных сетей путем сравнительного анализа пяти перспективных решений, извлеченных из каждого анализируемого датасета, ориентированных на классификацию изображения и детекцию объектов. Построены графики зависимости точности от количества используемых параметров в модели и величины обучающей выборки. Проведенный сравнительный анализ эффективности рассматриваемых решений позволил выделить наиболее действенные методы и технологии для проектирования архитектур искусственных нейронных сетей. Дополнительно, были идентифицированы перспективы для последующих исследований, сфокусированных на гибридизации сверточных нейронных сетей с визуальными трансформерами. Предложен новый метод, ориентированный на создание комплексной адаптивной архитектуры модели, которая может динамически настраиваться в зависимости от входного набора параметров, что представляет собой потенциально значимый вклад в область построения адаптивных нейронных сетей.

Ключевые слова:

визуальный трансформеры, сверточные нейронные сети, машинное обучение, анализ, гибридные нейронные сети, искусственный интеллект, компьютерное зрение, классификация, детекция, новый метод

Введение

В настоящее время наблюдается активная популяризация исследований в области развития и интеграции AI-технологий в различных сферах человеческой деятельности. Искусственные нейронные сети (далее – ИНС) становятся фундаментальным инструментом, оказывающим революционное воздействие на многие аспекты человеческой жизни. Они обладают способностью обучаться на обширных объемах данных, распознавать сложные образы, проводить анализ и формулировать прогнозы, которые ранее казались недостижимыми и предоставляют возможность оптимизировать выполнение повседневных задач путем частичной или полной автоматизации определенных операций. С каждым годом ИНС становятся все более мощными и сложными, открывая новые возможности в таких областях, как CV, NLP (от англ. Natural Language Processing, NLP - обработка естественного языка. Терминология), автономная навигация, прогнозирование и многое другое. Повышение эффективности нейросетевой модели становится приоритетным направлением исследований в различных ML-областях, позволяя достигать аналогичных результатов при более экономичном потреблении вычислительных ресурсов.

Целью настоящей статьи является проведение исследования современных архитектур ИНС, применяемых для решения задач классификации и детекции объектов. Основной фокус исследования направлен на выявление наиболее эффективных методов в структуре ИНС путем сравнительного анализа их преимуществ и недостатков.

В статье будет рассмотрено историческое развитие последних архитектурных достижений, приведено детальное описание ключевых аспектов, лежащих в основе современных подходов в области CV, проведен обзор актуальных тенденций и вызовов, с которыми сталкиваются исследователи и инженеры, работающие с AI-технологиями, представлены сравнительные графики, демонстрирующие основные показатели производительности современных моделей ИНС. Полученные результаты позволят выявить общие преимущества и недостатки современных подходов к построению нейронных сетей, определить перспективы и направления будущих исследований.

Как известно [1,2,3], внутренняя структура ИНС представляет собой сложную математическую модель, содержащую множество переменных параметров и динамически изменяемых вычислительных блоков. Независимо от глубины модели и сложности применяемых подходов, общее архитектурное представление ИНС может быть проиллюстрировано следующим образом (Рис 1).

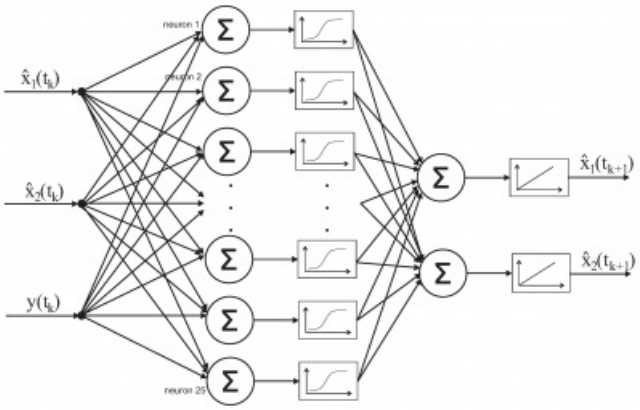

Рис. 1 – Схема архитектуры нейронной сети в общем виде

где *x* и *y* - нейроны или признаки входного слоя, Σ - сумматор весовых коэффициентов, полученных от входного набора признаков, график – отражает функцию активации для поддержания линейности или нелинейности выходных результатов

Кроме того, ИНС классифицируются по различным критериям, включая количество вычислительных слоев и тип построения внутренних связей между нейронами: однослойные, многослойные и с обратными связями. Однако, независимо от используемого класса, сохраняются основные структурные элементы, присущие любой модели ИНС.

В качестве функции активации может быть применено любое дифференцируемое решение [4,5]. Определение наиболее подходящей зависимости осуществляется инженером по работе с данными на основе: глубокого анализа предметной области и выбора активационной функции с оптимальным характером распределения, имеющегося опыта решения аналогичного набора задач, метода подбора для достижения наиболее точных показателей сходимости модели.

После выбора функции активации следует этап вычислительной обработки. Входные данные подаются в нейронную сеть в виде тензора, содержащего определенный набор

параметров [6]. Далее, в зависимости от текущего этапа построения модели, происходит либо однонаправленное прямое распространение ошибки, в случае готовности модели для применения в реальных задачах, либо двунаправленное (прямое и обратное) распространение ошибки, в случае нахождения модели в процессе обучения.

В процессе прямого распространения вычисляется сумма входных сигналов с учетом весовых коэффициентов для каждой связи, затем применяется активационная функция для получения выходных значений.

В процессе обучения модели, конечный результат подвергается анализу в соответствии с выбранным методом обучения - с учителем, без учителя или с подкреплением. Различия методов заключаются в способе получения ответа о корректности полученного предсказания.

Целью обучения ИНС является минимизация функции потерь путем обновления весовых коэффициентов в направлении антиградиента на каждом шаге обучения с помощью вариативных алгоритмов градиентного спуска.

Величина градиента вычисляется путем определения частных производных весовых коэффициентов.

Процесс обучения с помощью обновления весовых коэффициентов в соответствии с установленным шагом по направлению вектора антиградиента называется бэкпропом (от англ. backpropagation, backprop – обратное распространение ошибки. Терминология) и продолжается до момента достижения порогового значения величины точности, либо до момента прохождения заданного количества эпох обучения.

Основное отличие более сложных моделей заключаются в масштабе используемого пространства входных параметров, что позволяет создавать более сложные зависимости, агрегируя более обширный набор признаков и способствуя обработке как структурированных, так и неструктурированных данных большего объема.

Кроме того, современные прогрессивные модели ИНС используют многослойную структуру бэкбоуна (от англ. backbone – скелет, позвоночник. Терминология), что позволяет выделять как общие, так и более локальные зависимости в рамках одного датасета. Например, в контексте распознавания изображений многослойная архитектура может использоваться для выделения и конкатенации различных паттернов изображения.

Передовые ИНС имеют ряд общих проблем, решение которых способствует распространению исследований и интеграции новых подходов в области ML и AI-технологий. К ним относятся:

– Большое число входных параметров. В контексте обработки и анализа изображений наблюдается прямопропорциональное увеличение размерам модели в зависимости от разрешения и количества цветовых каналов. В следствии этого, использование базовой архитектуры ИНС для анализа современных изображений высокого качества становится неприменимым по причине существенного роста как времени обучения, так и обработки входного набора данных;

– Переобучение. Представляет собой альтернативную проблему, связанную с размером модели. Увеличение количества нейронов в ИНС способствует не только повышению временной продолжительности обучения, но и улучшению способности модели запоминать признаки. При обучении на ограниченном объеме данных ИНС может достичь

минимальных значений функции потерь, что свидетельствует о идеальном предсказании в рамках обучающей выборки. Однако, при проведении тестирования на реальном наборе тестовых данных, сеть может демонстрировать существенное увеличение доли ошибочных значений по сравнению с результатами на обучающей выборке. Данное явление характеризует непригодность полученной модели для эффективного решения поставленной задачи, поскольку, вместо обобщения зависимостей в данных, происходит запоминание конкретного обучающего набора, включая как структурированные и "чистые" данные, так и случайные, зашумленные или аномальные образцы. В итоге, модель может демонстрировать слабую способность к предсказанию достоверных результатов на практике;

– Затухание градиентов. Представляет собой распространенное явление в контексте обучения ИНС. Эта проблема связана с тем, что величины частных производных функции потерь по параметрам модели. Вычисляемые в процессе бэкпропа, с течением времени, могут принимать значения антиградиента на уровне погрешности. В результате, ИНС демонстрирует резкое снижение скорости обучения и обновления весовых коэффициентов, что, также, характеризует непригодность модели к использованию в реальных задачах.

Решение представленного ансамбля проблем способствует разработке более гибких и эффективных архитектурных подходов, позволяющих достигать аналогичных или более высоких показателей при более экономичном расходе вычислительных ресурсов. Поэтому, исследование современных архитектур ИНС является актуальной задачей, позволяющей получить сведения о наиболее удачных подходах, применяемых на практике.

Материалы и методы исследования

Для анализа последних достижений в области архитектур машинного обучения, необходимо ознакомиться с компаративным графиком наиболее эффективных моделей, применяемых в задачах классификации и детекции за последнее время. При отборе наиболее эффективных моделей учитывались только конечные итерации State-Of-The-Art (далее – SOTA) архитектурных решений. Анализ базируется на результатах, полученных при использовании следующих датасетов: ImageNet, CIFAR-10, CIFAR-100, MS COCO·minival. Сведения об использованных датасетах были взяты с открытого портала научных исследований PapersWithCode.

Для проведения анализа был сформирован набор из пяти SOTA-моделей из каждого датасета. На рисунке 2 представлены сводные показатели эффективности моделей ИНС, разделенные на группы в зависимости от используемого датасета и применяемой метрики.

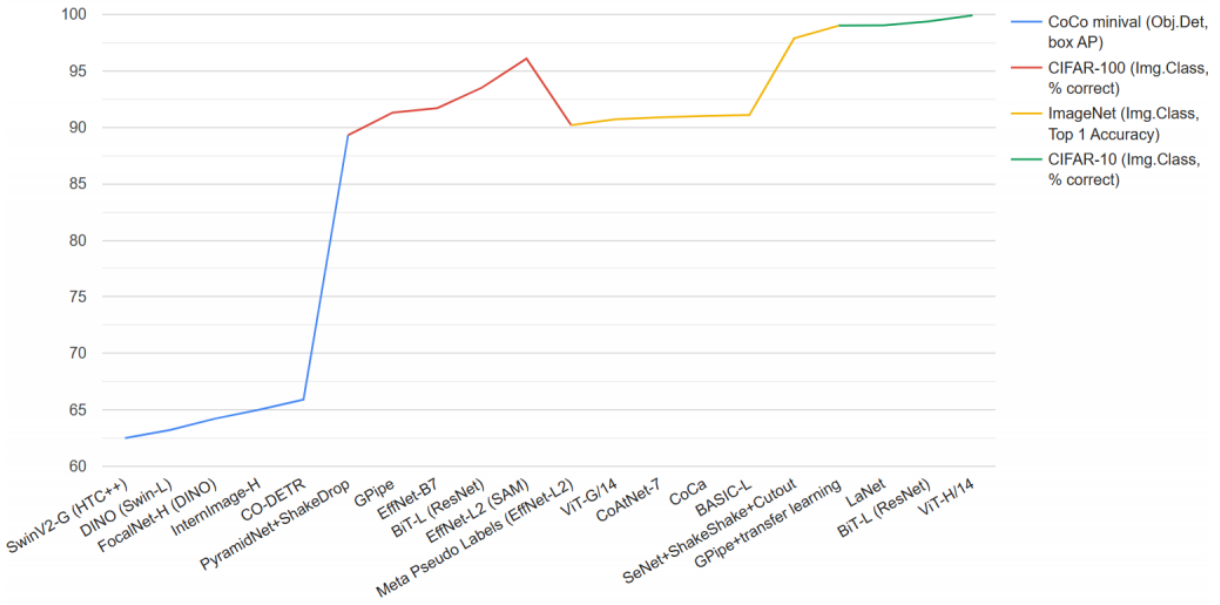

Рис. 2. Сводные результаты эффективности моделей ИНС

Сравнительный анализ эффективности ИНС будет базироваться на следующем наборе основных критериев:

– Точность классификации: способность модели корректно классифицировать представленные ей данные;

– Размер модели: количество параметров в архитектуре ИНС, влияющие на ее размер и возможность развертывания в условиях ограниченности вычислительных ресурсов;

– Адаптивность: способность модели выдавать корректный результат при введении нового, ошибочного или видоизмененного набора входных данных. Оценка данного параметра будет производиться на основе сведений о размере обучающей выборки и точности классификации.

Также, при проведении оценки эффективности наиболее удачных подходов неявно будет учитываться следующий набор параметров:

– Скорость обучения и предсказания: определяется временем, затрачиваемым на обучение модели и получение предсказаний. Является важным фактором при проведении анализа большего набора данных или решения задач в режиме реального времени;

– Устойчивость к переобучению: показатель, отражающий способность ИНС к обобщению данных обучающей выборки;

– Универсальность: способность модели работать с различными типами задач, такими как классификация, детекция, кластеризация и другие;

– Сложность модели: этот критерий связан с количеством слоев и нейронов в архитектуре, влияющий на обучаемость и производительность.

Проведение фактического сравнения указанных параметров не представляется возможным по причине отсутствия зафиксированных сведений в большинстве существующих решений.

Агрегация вышеперечисленных критериев формирует комплексную оценку

эффективности ИНС, имеющей важное значение при выборе модели для решения специфического круга задач.

В настоящей статье приводится анализ наиболее эффективных подходов, демонстрирующих значительное увеличение эффективности по сравнению с предшествующими архитектурами на основе исследования соответствующей литературы [7-33]. Рассмотрим структурные составляющие каждого из наиболее эффективных решений.

SENet[7] представляет собой инновационное решение, в рамках которого был разработан SE-блок (от англ. Squeeze and Excitation, SE – сжатие и возбуждение) с целью установления четких связей между признаковыми каналами на основе глобальной пространственной информации. Squeeze блок формируется путем применения операции GAP (от англ. Global Average Pooling, GAP – глобальное среднее объединение) к исходной карте признаков, что создает одномерный вектор, содержащий средние статистические значения для каждого канала.

Для использования информации, полученной на этапе сжатия, применяется Excitation-блок, суть которого заключается в применении механизма гейтинга (от англ. gating – стробирование, пропускание. Терминология) через функцию сигмоиды для выделения наиболее значимых элементом в промежуточном наборе данных. Для ограничения сложности модели, производится параметризация механизма гейтинга с помощью боттлнека (от англ. bottleneck – узкое место, узкое горлышко. Терминология), состоящего из двух FC-слоев (от англ. Fully Connected, FC – полносвязный. Терминология), обособленных между собой активационной функцией ReLU, являющегося слоем с понижающим размерность коэффициентом.

Дополнительными методами для повышения точности является ShakeShake [8] и Cutout[9] регуляризации. Регуляризацией называется добавление дополнительной информации к модели с целью предотвращения переобучения и уменьшения влияния «шума». Она достигается путем добавления штрафов или ограничений к функции потерь базовой модели.

Согласно результатам исследования [7], интеграция разработанного SE-блока в архитектуру ResNet [10] позволяет повысить точность ИНС, в среднем, на 10%, в зависимости от глубины модели при сравнительно небольшом увеличении вычислительных затрат. Кроме того, авторами утверждается, что подбор понижающего коэффициента следует осуществлять вручную, исходя из требований базовой архитектуры, поскольку в зависимости от величины коэффициента, снижается количество параметров, что может повлиять на точность модели, о чем свидетельствуют результаты сравнения на примере модели SE-ResNet-50 [7]. Примечательно, что вычислительная сложность наиболее глубокой модели ResNet-152 с использованием SE-блока имеет значение в 11.32GFLOPs, в то время как результат оригинальной модели составляет 11.3GFLOPs.

EfficientNet [11] (сокращенно EffNet) представляет собой архитектурное воплощение глубокой нейронной сети, основанной на технологии генеративного построения адаптивной архитектуры, аналогичной по отношению к использованной в MnasNet. Ключевая идея и отличие EffNet от MnasNet заключаются в том, что, как утверждают авторы [11], размерность сверточного слоя, извлекающего признаки из изображения, имеет линейную зависимость от размера входного изображения, что требует изменения

глубины, ширины сверточного слоя или разрешения изображения для извлечения более «тонких» паттернов. В отличие от традиционных методов [11], авторский подход заключается в комбинированным применении всех традиционных методов.

Дополнительно было выявлено, что изменение гиперпараметров модели неоднородно влияет на вычислительную сложность. Увеличение глубины модели путем добавления новых сверточных слоев имеет линейную зависимость вычислительной сложности, в то время как увеличение ширины фильтра или разрешения изображения оказывает квадратичное воздействие.

Основываясь на результатах исследований 2017 года специалистами компании Google [12], был применен подход к построению рекуррентных нейронных сетей для создания автономно генерируемой модели. Принцип работы архитектуры MnasNet заключается в предсказании наиболее эффективного блока ширины, глубины или страйда для поданного тензора (метод Compound Scaling), построения модели и последующего обучения с подкреплением посредством максимизации величины награды. На основании результатов классификации, проводится проброс градиента ошибки в начальный строительный блок ИНС и делается повторное предсказание. Авторами исследования [12] отмечается, что итеративное повторение продолжалось до достижения заданного количества эпох обучения, однако этот подход можно повторять неограниченное число раз. Используя концепцию многоцелевого поиска, оптимизирующего как точность, так и вычислительную сложность, была разработана собственная базовая архитектура EffNet-B0.

Главным элементом всей архитектуры выступает инвертированный bottleneck-блок, именующийся MBConv в сочетании с ранее упомянутой концепцией SE-блока [7].

Масштабирование ИНС происходит в 2 этапа:

– Этап 1. Авторами фиксируется параметр $\phi = 1$, предполагается, что имеется двукратное превышение свободных ресурсов, после чего выполняется поиск по малой сети α, β, γ на основании выражений 7 и 9. Для EffNet-B0 наилучшие значения коэффициентов составили $\alpha = 1.2$, $\beta = 1.1$, $\gamma = 1.15$, при учете следующего ограничения $\alpha*\beta^2*\gamma^2 \approx 2$.

– Этап 2. Фиксируются переменные α, β, γ как постоянные, после чего производится масштабирование сети по различным параметрам, используя выражение 9. Таким образом были получены модификации EffNet от B1 до B7

Важно отметить, что применение метода Compound Scaling эффективно лишь для базовых моделей с небольшим количеством параметров. Увеличение размера модели существенно снижает эффективность и замедляет процесс поиска оптимальных решений [11].

Общая идея последовательного генерирования строительных блоков модели ИНС получила название AutoML. Однако основным ограничением этого подхода является использование заранее заданных сверточных операций, связей и блочных структур. В связи с чем, полная автоматизация процесса поиска оптимальной архитектуры не представляется возможной в рамках AutoML подхода.

ViT. Архитектура Visual Transformer [13]. Поскольку отличительной особенностью всех моделей ViT является полное единообразие структуры, за исключением размера рецептивного поля, анализ может быть продуцирован для всех существующих моделей

одновременно.

Идея ViT возникла после успеха архитектуры Transformer в задачах NLP. Transformer [14] базируется на основе структуры энкодер-декодер, используемой для преобразования последовательностей. В своей основе для решения задач NLP трансформер использует механизм многоголового внимания (от англ. Multi-Head Attention, далее – MHA). Этот механизм опирается на метод масштабированного произведения скаляров (мультипликативного) (от англ. Scaled dot-product attention), который представляет собой взвешенную сумму сопоставления запроса с парами ключ-значение.

В качестве входных параметров в NLP используются результаты токенизации по словам. Заместо использования единой функции внимания со всеми ключами, значениями и запросами размерности модели, авторами [17] было решено применить линейное проецирование ключей и значений h раз с выученными проекциями размерностей D_k, D_k и D_V. На каждой проекции параллельно выполняется функция внимания, результатом которой являются выходные значения размерности D_V.MHA позволяет использовать информацию с разных подпространств на разных позициях.

Дополнением к слоям внимания служит FC-слой прямого распространения, применяющейся поэлементно к каждой позиции.

Для генерации вероятностного распределения последующего токена на основе выходных данных декодера применяется операция линейного преобразования в сочетании с softmax-функцией.

Поскольку в задачах CV использование явной токенизации невозможно из-за высоких требований к вычислительным ресурсам, обусловленных большим количеством параметров и квадратичной зависимостью времени работы механизма внимания от входных данных, в архитектуре ViT применяется комбинированный подход, основанный на патчинге (от англ. Patch – клочок. Терминология) и PE (от англ. Position Embedding, PE – встраивание позиции).

Механизм патчинга осуществляется путем разбиения полного набора пикселей на мелкие группы, известные как патчи. Полученный набор токенов подается в стандартный энкодер архитектуры Transformer [14] вместе с class-token, используемый для итоговой классификации изображения с последующим блоком сети прямого распространения.

Swin-V2 . Как было ранее отмечено, механизм внимания имеет квадратичную зависимости от входного набора токенов, что ставит под сомнение применимость ViT для задач сегментации или детекции из-за необходимости использования патчей меньшей размерности, что незамедлительно окажет существенное негативное воздействие на производительность сети.

Архитектура Swin [15,16] представляет собой следующий этап развития визуальных трансформеров для вышеописанных задач. На вход поступает набор патчей размерности 4*4, после чего происходит последовательное применение нескольких слоев Parch Merging (далее – PM) и Swin Transformer Block (далее – ST) через линейное преобразование. PM выполняет операцию конкатенации признаков соседних токенов в фильтре размерности 2*2, получая более высокоуровневое представление изображения. Таким образом после каждой стадии формируется карта признаков различных иерархических представлений, после чего применяется ST-блок. Исследование структуры ST-блока показывает [15,16], что он отражает классическую архитектуру

энкодер-декодер [14], но отличается применением операций skip-connection, характерной для ResNet и измененной функции внимания, которая является ключевым элементом Swin-архитектуры.

Авторами [15,16] было принято решение считать внимание не со всеми существующими в наборе токенами, а только с теми, которые находятся в определенной области фиксированного размера ((Window) Multi-Head Self Attention, далее – W-MSA)

Таким образом, внимание теперь функционирует за линейное *hw* время. Однако, применение данного метода сопровождается уменьшением репрезентативной способности сети в силу отсутствия связи между различными токенами. Для коррекции данной проблемы, авторы внесли дополнительный слой с диагональным смещением рецептивного окна после каждого блока W-MSA. Это восстановило взаимодействие между токенами, сохраняя при этом линейную вычислительную сложность.

В связи с увеличением количества вычислительных операций, при использовании данного подхода, авторами было предложено проведение предварительного циклического сдвига с целью исключения взаимодействия между несмежными токенами путем наложения маски. Такой подход позволяет сохранять количество операций внимания при применении операции паддинга изображения.

Помимо внедрения нового метода вычисления Attention, авторами введена новая калькуляция position embedding путем добавления обучаемой матрицы relative position bias.

Результатом синтеза RA и W-MSA стало возможным построение архитектуры, позволяющей извлекать признаки на разных пространственных масштабах и успешно использовать Swin в качестве бэкбоуна в задачах сегментации и детекции.

CoAtNet- 7 . Сеть Convolutional Attention Network [17]представляет собой результат комбинации передовых CNN и ViT-подходов для решения CV-задач. С помощью объединения пространственной свертки и механизма самовнимания, а также последовательного наложения слоев свертки и внимания, была разработана архитектура, проявляющая высокую производительность как на больших, так и малых датасетах.

CoAtNet вводит новую структуру относительного внимания (от англ. relative attention, далее – RA), который интегрирует концепцию относительных весов или смещений в процесс кодирования-декодирования данных. Вместо присвоения фиксированных весов каждому элементу, RA позволяет модели учитывать относительные позиции элементов в последовательности при вычислении весов. Это особенно полезно, когда важность элементов зависит от их относительного расположения, а не абсолютного. Однако использование RA к исходному изображению требует значительных вычислительных затрат.

Авторами [17] было проведено исследование двух подходов к решению данной проблемы:

– Использование сверточного stem-блока 16*16, а затем блоки L-трансформеров с RA;

– Применение многоступенчатой сети с постепенным пулингом. Для исследования было выбрано 4 конфигурации сети, состоящих из C-свертки и T-трансформер.

Оценка проводилась на основе двух критериев:

– Обобщение: величина минимального разрыва между показателями по обучающей и валидационной выборке;

– Емкость: лучшая производительность за одинаковое количество эпох.

Результатом авторского исследования было получено заключение, что применение C-C-T-T конструкции является наиболее эффективной среди исследуемых.

Результаты

Для комплексной оценки эффективности имеющихся решений было необходимо провести анализ зависимости точности модели от количества параметров и объема обучающей выборки. Для этого использовались данные ранее описанных моделей, обученных на датасете ImageNet. При отсутствии данных для конкретной модели, она исключалась из результирующей выборки. Графическое представление нормализованных результатов исследования проиллюстрировано на рисунке 3.

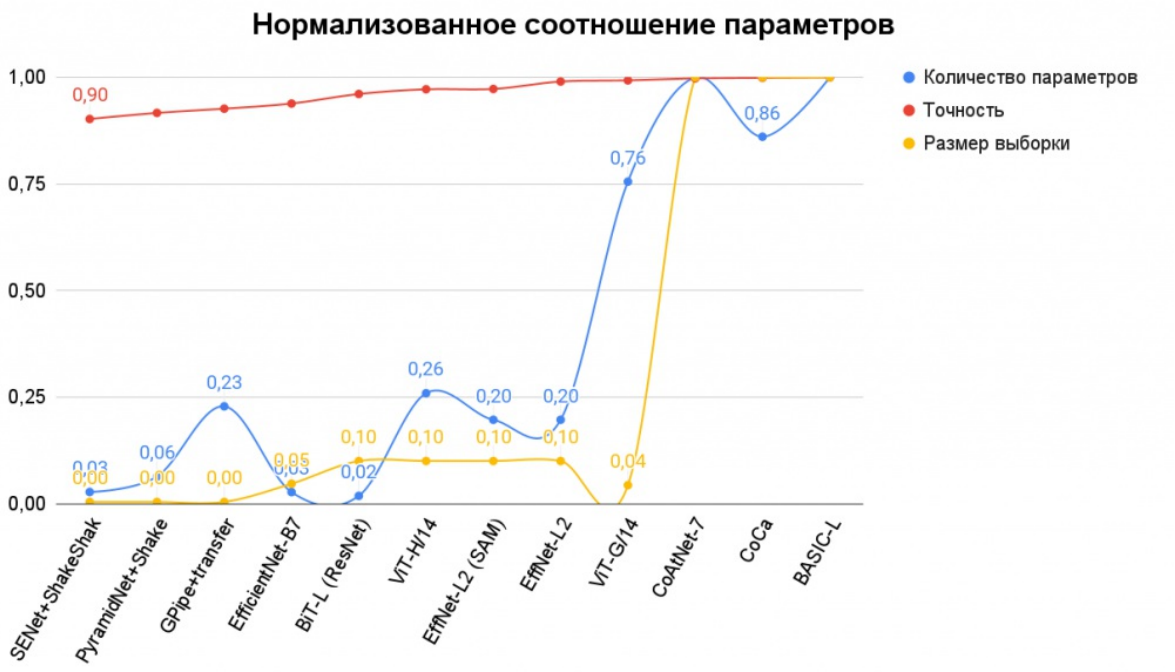

Рис. 3 – График нормализованного соотношения параметров

На основе анализа представленной информации можно утверждать, что сверточные архитектуры, на текущий момент, обладают более высокой эффективностью по сравнению с гибридными или ViT решениями. С точки зрения соотношения точности, размера обучающей выборки и количества параметров, наиболее эффективной архитектурой является EfficientNet из семейства CNN. Она проявляет высокую способность к обобщению данных в задачах классификации при малой обучающей выборке и количестве параметров модели. Кроме того, современные решения, основанные на архитектуре ViT или использующие гибридный подход, требуют большего объема данных для достижения рекордных или схожих с CNN результатов. Однако, исследование [13] свидетельствует о высокой скорости обучения ViT-решений благодаря способности патчей охватывать все рецептивное поле изображения одновременно, в отличие от последовательного сравнения выходных карт признаков в CNN-архитектурах.

Согласно результатам исследования [13], использование гибридизации CNN и ViT позволяет, в среднем, достигать большей точности классификации при равной вычислительной сложности, о чем свидетельствует иллюстрация на рисунке 4.

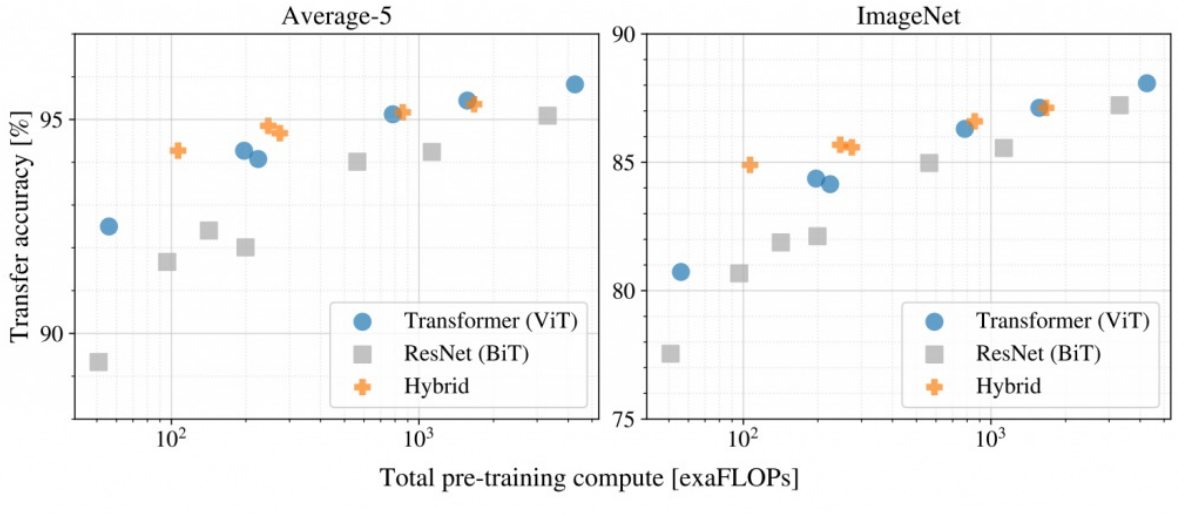

Рис. 4 - Сравнение вычислительной сложности CNN, ViT и гибридной подхода

Исходя из анализа вышеперечисленных факторов можно заключить, что в настоящее время наиболее эффективными строительными блоками для построения архитектуры ИНС являются: AutoML + Compound Scaling (EffNet, CNN), W-MSA (Swin-V2, ViT), Squeeze and Excitation (SENet, CNN), Skip-connection (residual blocks, ResNet, CNN).

Дальнейшие исследования нацелены на улучшение существующих архитектурных решений путем разработки гибридной CNN-ViT архитектуры с целью уменьшения зависимости скорости обучения и точности модели от размера обучающей выборки. Этот результат может быть достигнут с помощью комплексного применения технологий AutoML, Compound Scaling, также бэкбоуна EffNet, либо на основе разработки новейшего подхода к извлечению признаков из входного набора данных.

Полученные результаты исследования могут быть использованы для выбора оптимальной архитектуры построения ИНС, решающей специфические бизнес-задачи или разработки инновационной модели, учитывающей все вышеупомянутые критерии эффективности.

Библиография

1. Gomolka Z., Using artificial neural networks to solve the problem represented by BOD and DO indicators //Water. – 2017. – Т. 10. – №. 1. – С. 4.
2. Кадурин А., Николенко С., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей //СПб.: Питер. – 2018. – Т. 480.
3. Джабраилов Шабан Вагиф Оглы, Розалиев Владимир Леонидович, Орлова Юлия Александровна Подходы и реализации компьютерной имитации интуиции // Вестник евразийской науки. 2017. №2 (39).
4. Бабушкина, Н. Е. Выбор функции активации нейронной сети в зависимости от условий задачи / Н. Е. Бабушкина, А. А. Рачев // Инновационные технологии в машиностроении, образовании и экономике. – 2020. – Т. 27, № 2(16). – С. 12-15.
5. Соснин А. С., Суслова И. А. Функции активации нейросети: сигмоида, линейная, ступенчатая, relu, tahn. – 2019. – С. 237.
6. Бредихин Арсентий Игоревич Алгоритмы обучения сверточных нейронных сетей // Вестник ЮГУ. 2019. №1 (52).

7. Hu J., Shen L., Sun G. Squeeze-and-excitation networks //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – С. 7132-7141.
8. Gastaldi X. Shake-shake regularization //arXiv preprint arXiv:1705.07485. – 2017.
9. DeVries T., Taylor G. W. Improved regularization of convolutional neural networks with cutout // arXiv preprint arXiv:1708.04552. – 2017.
10. He K. et al. Deep residual learning for image recognition //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – С. 770-778.
11. Tan M., Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks //International conference on machine learning. – PMLR, 2019. – С. 6105-6114.
12. Tan M. et al. Mnasnet: Platform-aware neural architecture search for mobile //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2019. – С. 2820-2828.
13. Dosovitskiy A. et al. An image is worth 16x16 words: Transformers for image recognition at scale //arXiv preprint arXiv:2010.11929. – 2020.
14. Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – Т. 30.
15. Liu Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows // Proceedings of the IEEE/CVF international conference on computer vision. – 2021. – С. 10012-10022.
16. Liu Z. et al. Swin transformer v2: Scaling up capacity and resolution //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2022. – С. 12009-12019.
17. Dai Z. et al. Coatnet: Marrying convolution and attention for all data sizes //Advances in neural information processing systems. – 2021. – Т. 34. – С. 3965-3977.
18. Zhai X. et al. Scaling vision transformers //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2022. – С. 12104-12113.
19. Huang Y. et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism //Advances in neural information processing systems. – 2019. – Т. 32.
20. Методы аугментации обучающих выборок в задачах классификации изображений / С. О. Емельянов, А. А. Иванова, Е. А. Швец, Д. П. Николаев // Сенсорные системы. – 2018. – Т. 32, № 3. – С. 236-245. – DOI 10.1134/S0235009218030058.
21. Cubuk E. D. et al. Autoaugment: Learning augmentation strategies from data //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2019. – С. 113-123.
22. Han D., Kim J., Kim J. Deep pyramidal residual networks //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2017. – С. 5927-5935.
23. Yamada Y. et al. Shakedrop regularization for deep residual learning //IEEE Access. – 2019. – Т. 7. – С. 186126-186136.
24. Kolesnikov A. et al. Big transfer (bit): General visual representation learning //Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. – Springer International Publishing, 2020. – С. 491-507.
25. Foret P. et al. Sharpness-aware minimization for efficiently improving generalization //arXiv preprint arXiv:2010.01412. – 2020.
26. Pham H. et al. Meta pseudo labels //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2021. – С. 11557-11568.
27. Yu J. et al. Coca: Contrastive captioners are image-text foundation models //arXiv preprint arXiv:2205.01917. – 2022.

28. Chen X. et al. Symbolic discovery of optimization algorithms //arXiv preprint arXiv:2302.06675. – 2023.
29. Zhang H. et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection //arXiv preprint arXiv:2203.03605. – 2022.
30. Yang J. et al. Focal modulation networks //Advances in Neural Information Processing Systems. – 2022. – Т. 35. – С. 4203-4217.
31. Wang L. et al. Sample-efficient neural architecture search by learning actions for monte carlo tree search //IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2021. – Т. 44. – №. 9. – С. 5503-5515.
32. Wang W. et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2023. – С. 14408-14419.
33. Zong Z., Song G., Liu Y. Detrs with collaborative hybrid assignments training //Proceedings of the IEEE/CVF international conference on computer vision. – 2023. – С. 6748-6758.

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Статья глубоко анализирует современные архитектуры искусственных нейронных сетей (ИНС), акцентируя внимание на их использовании в задачах классификации и детекции изображений. Авторы эффективно обозначают исследуемую область, подчеркивая значимость ИНС в текущем научном дискурсе. Исследование базируется на комплексном подходе, включающем анализ данных из различных источников и сравнительный анализ эффективности различных архитектур ИНС. Этот подход демонстрирует глубокое понимание предмета и позволяет авторам достичь значимых выводов. Тема исследования крайне актуальна, учитывая растущее значение AI-технологий в современном мире. Исследование подчеркивает важность развития и оптимизации ИНС для улучшения аналитических возможностей в области обработки изображений. Статья представляет новый взгляд на анализ и сравнение современных ИНС, обогащая научное сообщество свежими идеями и подходами. Авторы выделяются оригинальным анализом и интерпретацией данных, что способствует продвижению научного понимания в данной области. Статья отличается ясной структурой, логичным изложением и качественным стилем написания. Текст организован так, что читатель может легко следовать логике авторов и их аргументации. Обширный и актуальный список литературы свидетельствует о тщательной работе авторов над исследованием. Использование авторитетных источников повышает достоверность и научную ценность работы. Авторы адекватно относятся к существующим взглядам и исследованиям в данной области, предоставляя уважительную и конструктивную критику, что способствует развитию диалога в научном сообществе. Выводы статьи являются убедительными и хорошо подкреплены представленными данными и анализом. Исследование представляет значительный интерес для широкого круга читателей, включая специалистов в области компьютерного зрения, искусственного интеллекта и машинного обучения. Анализ результатов исследования подчеркивает значимость и эффективность представленных архитектур ИНС, особенно в контексте обработки и анализа изображений. Авторы продемонстрировали глубокое понимание и анализ различных архитектур, включая сверточные модели, трансформеры и гибридные системы. Они эффективно сравнили их

по критериям, таким как точность классификации, размер модели, адаптивность, скорость обучения, устойчивость к переобучению, универсальность и сложность модели. Этот всесторонний анализ предоставляет ценную информацию о преимуществах и недостатках каждого подхода, что важно для исследователей и практиков в области искусственного интеллекта и машинного обучения. В целом, результаты исследования демонстрируют значительный вклад в развитие и понимание современных архитектур ИНС, предоставляя практические рекомендации для их выбора и применения в реальных задачах. Эти результаты подчеркивают актуальность и важность исследования, делая его значимым вкладом в данную область. В дополнение к положительной рецензии, хотелось бы отметить, что рисунки в статье выходят за границы HTML-страницы, что может затруднить их просмотр и восприятие. Рекомендуется представить иллюстрации более емко, чтобы обеспечить их лучшую визуальную доступность и интеграцию с текстом статьи.