DOI 10.37539/2949-1991.2025.29.6.004
УДК 004.89

Чжао Чэньсяо, студентка,
Калужский филиал Московского государственного
технического университета имени Н. Э. Баумана
Калуга
Zhao Chenxiao
Kaluga Branch of the Bauman Moscow
State Technical University

Белов Юрий Сергеевич,
к.ф.-м.н., доцент,
Калужский филиал Московского государственного
технического университета имени Н. Э. Баумана,
Калуга
Belov Yuri Sergeevich
Kaluga Branch of the Bauman Moscow
State Technical University

ПРИМЕНЕНИЕ АРХИТЕКТУРЫ TRANSFORMER И МЕХАНИЗМА ВНИМАНИЯ В ДЕТЕКТИРОВАНИИ ОБЪЕКТОВ С МАЛЫМ КОЛИЧЕСТВОМ ОБРАЗЦОВ
APPLICATION OF THE TRANSFORMER ARCHITECTURE AND ATTENTION MECHANISM IN FEW-SHOT OBJECT DETECTION

Аннотация: Архитектура Transformer и механизм внимания благодаря своей мощной способности к представлению данных находят применение в различных задачах обработки естественного языка (NLP) и компьютерного зрения (CV). В данной работе кратко описывается состав и функции архитектуры Transformer, а также её применение в задачах детектирования объектов, включая малообразцовое детектирование объектов.

Abstract: Thanks to its powerful representation capability, the Transformer architecture and attention mechanism are widely applied in various tasks of natural language processing (NLP) and computer vision (CV). This paper briefly describes the components and functions of the Transformer architecture, as well as its application in object detection tasks, including few-shot object detection.

Ключевые слова: Transformer, механизм внимания, детектирование объектов, малообразцовая задача

Keywords: Transformer, attention mechanism, object detection, few-shot learning

Введение в архитектуру Transformer

С развитием области искусственного интеллекта и глубокого обучения все большее количество архитектур предлагаются и применяются для различных задач с целью повышения таких показателей, как эффективность и точность. Среди них архитектура Transformer была предложена как глубокая нейронная сеть, основанная в первую очередь на механизме само-внимания (self-attention) [1]. Эта архитектура обладает мощной способностью к представлению данных и изначально применялась в области обработки естественного языка (NLP), где может успешно выполнять типичные задачи, такие как машинный перевод и генерация текста, а также использоваться для построения предобученных языковых моделей, поддерживающих перенос обучения на множество задач.

В настоящее время, вдохновленные большим успехом Transformer в области NLP, исследователи в области компьютерного зрения начали рассматривать Transformer как альтернативу традиционным сверточным нейронным сетям (CNN), которые долгое время считались основными компонентами в этой области. Архитектура Transformer, являясь простой структурой, может достигать результатов, сравнимых с результатами многослойных сверточных нейронных сетей.

Базовая структура архитектуры Transformer выглядит следующим образом:

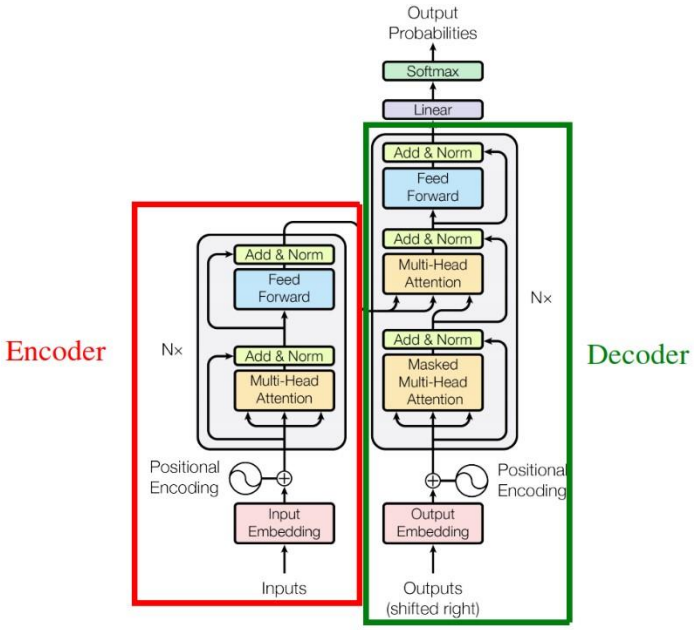

Рис. 1. Базовая архитектура Transformer

Как показано на рис. 1, общая архитектура состоит из части кодера и части декодера. Каждая из этих частей формируется путем стекирования N кодеров или декодеров. В свою очередь, каждый кодер или декодер включает в себя встраивание векторов, позиционное кодирование, много головое само-внимание (multi-head self-attention), нормализацию слоя, прямую нейронную сеть (feedforward network) и остаточное соединение (residual connection).

Однако в практических приложениях модель не обязательно должна содержать как кодер, так и декодер. Например, модель BERT [2], в которой сохранены только кодеры, сосредоточена на изучении и понимании языкового содержания. Предобучение модели BERT проводится на двусторонних задачах, что позволяет учитывать контекст слов и выявлять более тонкие связи между ними. Примером другой крайности являются серии моделей GPT, в которых сохранены только декодеры. Модель GPT-2 демонстрирует отличные результаты в таких языковых задачах, как перевод, генерация и завершение текста, а GPT-3 показывает, что предобученные большие языковые модели способны к обучению без примеров (zero-shot learning) [3]. Существуют также модели, в которых сохраняются и кодер, и декодер, например, модель T5, которая может выступать в роли вспомогательной модели и помогать обучению первых двух типов моделей.

Применение в детектировании объектов

Поскольку суть механизма внимания заключается в фокусировке на более важных объектах, что хорошо соответствует требованиям задачи детектирования объектов, исследователи внедрили архитектуру Transformer в рамки для детектирования объектов и предложили новую архитектуру DETR [4]. На основе небольшого набора фиксированных

обучающихся запросов object queries данная архитектура определяет взаимосвязь между объектами и глобальным контекстом изображения, чтобы параллельно напрямую вывести окончательный набор прогнозов. Было показано, что точность и время работы модели на сложном наборе данных COCO для детектирования объектов сопоставимы с показателями базовой модели Faster R-CNN.

На этой основе исследователи предложили новый метод Dynamic DETR [5], который направлен на преодоление ограничений, связанных с низким разрешением признаков и медленной сходимостью во время обучения. В рамках этого метода динамическое внимание было внедрено на этапах кодера и декодера DETR. Конкретно, для этого используется основанный на свертках динамический кодер с различными типами внимания, аппроксимирующий механизм внимания кодера Transformer, а в декодере модуль перекрестного внимания заменяется на динамический модуль внимания на основе ROI. Такой подход позволяет сосредоточиться именно на интересующих регионах и тем самым упрощает процесс обучения.

Применение в детектировании объектов с малым количеством образцов

Детектирование объектов с малым количеством образцов (few-shot object detection) – это подзадача, постепенно ставшая популярной в области моделей, объединяющих изображения и язык (VLMs), целью которой является определение, содержит ли изображение для запроса объект данного класса, при наличии лишь небольшого числа опорных образцов, а также указание его местоположения. Архитектура Transformer и механизм внимания обладают большим потенциалом применения в данной задаче.

Например, предположим, что строится модель для малообразцового детектирования объектов на изображениях дистанционного зондирования

(remote sensing). Для извлечения признаков изображений, кроме основной сети (backbone), также необходимы этапы формирования кандидатных регионов и выделения регионов интереса (ROI). Однако большинство объектов на изображениях дистанционного зондирования имеют углы поворота. В этом случае требуется строить наклонные анкеры с учетом угла поворота для выделения ROI. При этом, хотя на выходе могут быть получены наклонные кандидатные рамки, структура карты признаков остается неизменной. Объекты, которые необходимо обрабатывать, остаются наклонными, и при извлечении признаков из прямоугольных областей, ориентированных по осям, возникает несоответствие формы и направления относительно реального объекта, что может мешать окончательным результатам. Поэтому можно ввести RoI

Transformer и Self-Attention: после того, как наклонный регион будет повернут и скорректирован до горизонтального положения, использовать кодер Transformer для моделирования признаков RoI каждого proposal, чтобы усилить семантическое понимание целевого объекта.

Кроме того, модель также может быть дополнена языковой моделью для уточнения семантики, что повысит точность детектирования. Например, можно интегрировать языковую модель BERT – типичный Transformer-кодер, который кодирует каждое слово (token) в вектор фиксированной размерности (например, 768). В завершение необходимо объединить визуальные и текстовые признаки с помощью механизма cross-attention.

Список литературы:

1. Барышников, П.Н. (2024). Чем является научное знание, произведенное методами Больших языковых моделей? *Философские проблемы информационных технологий и киберпространства*, № 1 (25), 89–103. DOI: 10.17726/philIT.2024.1.6.

2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [arXiv preprint]. arXiv:1810.04805.

3. Brown, T., et al. (2020). Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901).

4. Carion, N., et al. (2020). End-to-End Object Detection with Transformers (pp. 1– 18). Springer. Lecture Notes in Computer Science, vol. 12346. European Conference on Computer Vision (ECCV).

5. Dai, X., et al. (2021). Dynamic DETR: End-to-End Object Detection with Dynamic Attention. In Proceedings of the IEEE International Conference on Computer *Vision* (pp. 2988–2997).