АНАЛИЗ ПРИМЕНЕНИЯ АРХИТЕКТУРЫ TRANSFORMER СЕГОДНЯ

Е. А. Артемьев, email: artemev.15.01.1998@list.ru

Казанский национальный исследовательский технический университет

Аннотация. *Сегодня трудно представить жизнь без компьютерного зрения. Оно встречается повсюду: начиная с распознавания вывесок на улице и заканчивая внедрением в беспилотные автомобили и постановкой диагноза в больнице. Что раньше было описано в книгах и что называли фантастикой, теперь осязаемо. Каждый человек использует FaceId, проезжает под камерами дорожного движения, не задумываясь над тем, что встроенные алгоритмы в реальном времени могут определить сотни объектов, предсказать события. Это обширная область исследования, в которой каждый день ведутся работы над созданием более точных алгоритмов решения задач. Одним из таких алгоритмов является «Transformer», возникший относительно недавно и о котором пойдёт речь в статье.*

Ключевые слова: *распознавание и детектирование изображений; архитектура "Transformers", обнаружение, объект.*

Введение

Компьютерное зрение предстаёт перед нами такими разделами как детектирование, распознавание, сегментация, классификация и исследование аномальных зон. Всё это находит применение в реальной жизни, что является стимулом для дальнейших разработок.

Данный раздел машинного обучения прошёл долгий путь, на котором появлялись перцептрон, RNN, CNN, R-CNN, Fast-CNN и Faster-CNN и другие архитектуры. Всё перечисленное довольно глубоко изучено, но вращается вокруг прироста скорости в обучении, точности, размера выборки на давно известных алгоритмах. Поэтому революционные решения появляются довольно редко.

И всё же стоит обратить внимание на такую архитектуру как "Transformer", которая смогла оживить застой в компьютерном зрении.

"Transformers" не один год применялись в интеллектуальных системах, но в детектировании их использовать начали относительно недавно. За счёт своего понятного устройства, поддержки параллельных вычислений и алгоритма «самовнимания» архитектура стала мощным инструментом в работе с текстом и изображениями.

Степень изученности исследуемой проблемы: теоретическим и методологическим проблемам повышения эффективности работы с текстом и с изображениями посвящены труды зарубежных учёных и специалистов, среди которых B.Wu, C. Xu, X. Dai, A.Wan, P.Zhang, Z.Yan, M.Tomizuka, J.Gonzalez, K.Keutzer, P.Vajda.

Большой вклад в разработку алгоритмов с использованием “Transormers” внесли N.Carion, F.Massa, G.Synnaeve, N.Usunier, A.Kirillov, S.Zagoruyko, H.Chen, Y.Wang, T.Guo, Y.Deng, Z.Liu, S.Ma, W.Gao и другие.

1. Литературный обзор

В статье “End-to-End Object Detection with Transformers” [1] описана архитектура на базе “Transformers”, которая эффективно позволяет детектировать объекты. Они утверждают, что “внимание”, описанное в статье «Attention Is All You Need» и применяемое в анализе текста, подходит и для анализа изображения. Авторы добавили ResNet50 для получения характеристик со следующим переходом в трансформер и вычислением потерь при помощи венгерского алгоритма.

В статье «Attention Is All You Need» [2] представлена архитектура, применяемая в анализе текста. Она отличалась от своих предшественников хорошим распараллеливанием, разработкой “Multi-head attention”. «Мультиголовое внимание» позволило входному вектору взаимодействовать с другими словами через attention mechanism.

В статье «Rethinking Transformer-based Set Prediction for Object Detection» [3] описано, почему архитектура “DETR” от специалистов Facebook AI медленно обучается и как можно исправить данный момент. В статье приводятся результаты исследования и предложены методы на замену венгерскому алгоритму.

В статье «Attention is not all you need: pure attention loses rank doubly exponentially with depth» [4] описано, как работает “внимание” изнутри и описываются минусы его использования.

В статье “Visual Transformers: Token-based Image Representation and Processing for Computer Vision” [5] описано, как при помощи архитектуры “Transformer” классифицируются изображения. Авторы делят изображение на несколько токенов и с помощью внимания определяют, как связаны пиксели токенов между собой

Работы, приведённые выше рассматривают архитектуру «Transformer» с разных сторон относительно недавно, что свидетельствует о её актуальности и разноплановом применении. Но данная модель не является идеальной, о чём пишут в статьях выше. У неё есть ряд задач, с которыми она справляется и ряд условий, при которых результаты будут впечатляющими.

2. Задачи с использованием “Transformer” сегодня

Архитектура применяется в такой задаче как понимание трёхмерных сцен для различных приложений. Суть в том, что модель способна определить границы комнат и объектов внутри них. В статье “PQ-Transformer: Jointly parsing 3d Objects and Layouts From Point Clouds” авторы применяют декодер разбираемой архитектуры [6].

Также применяется для оценки эмоционального состояния спикеров. Резкие переходы из одного состояния в другое как описано в статье “Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer” [7]. Прогресс в оценке настроения собеседника поможет умным ассистентам, чат-ботам понимать тон отвечающего, его интонацию и отвечать более гибко уже исходя не только из контекста предложения, но и из эмоционального контекста.

В работе “Graph transformer network with temporal kernel attention for skeleton-based action recognition” описывается распознавание действий человека на основе скелета при помощи графового трансформера для предотвращения игнорирования сложных зависимостей между действиями [8].

Распознавание жестов рук является сложной задачей. Авторы “Content-Adaptive and Attention-Based Network for Hand Gesture Recognition, 3D interacting hand pose and shape estimation from a single RGB image” использовали “Transformer”, чтобы решить данную проблему и получить внушительные результаты, не использовав свёртки и повторяющиеся слои. [9].

Диагностика лёгких (“Hybridizing Convolutional Neural Network for Classification of Lung Diseases”), эпилепсии (“Epileptic Seizure Prediction Using Deep Transformer Model”), лапароскопическая сакроколпопексия (“Large-scale surgical workflow segmentation for laparoscopic sacrocolpopexy”) это задачи, решение которых поможет врачам ставить более точные диагнозы и проводить точные операции [10].

По мнению авторов “UTRAD: Anomaly detection and localization with U-Transformer” обнаружение аномалий в промышленности или медицине является активной областью исследования. Они предложили “U-Transformer” для более точного результата [11].

Идею трансформеров используют для анализа трафика на дорогах в “Spatial-Temporal Convolutional Transformer Network for Multivariate Time Series Forecasting”, чтобы предотвращать пробки и рационально использовать всю площадь дорог[12].

В “Cross-Modal Object Detection Based on a Knowledge Update” используют 3 кодировщика, чтобы донести до машины способность

узнавать объекты по описанию и рассуждать [13]. Это приближает способ мышления машины к тому, как мыслит человек, который описывает то, что видит или по описанию объекта может сделать выводы о его принадлежности к определённому классу.

По перечисленным статьям заметна главная черта, по которой выбирают “Transformer” в том или ином виде – это способность анализировать сложные меняющиеся ситуации и сложные связи внутри объекта.

3. Метрики

Для оценки точности алгоритма используют такие метрики как MAE, GIOU, MSE, RMSE. MAE – в статистике, средняя абсолютная ошибка, является мерой ошибок между парными наблюдениями, выражающими то же явление. Вычисляется по формуле:

$$MAE = \sum_{i=1}^{n} \left| y_i - x_i \right| / n = \sum_{i=1}^{n} \left| e_i \right| / n ,$$

- n – число прогнозов;
- y – спрогнозированное значение;
- x – изначальные значения.

GIOU известна как generalized intersection over union – метрика и потеря для регрессии ограничивающей рамки:

$$GIOU = \frac{\left| A \cap B \right|}{\left| A \cup B \right|} - \frac{\left| C \setminus (A \cup B) \right|}{\left| C \right|} = IOU - \frac{\left| C \setminus (A \cup B) \right|}{\left| C \right|} ,$$

- A, B– ограничивающие рамки предсказанные и данные изначально соответственно;
- C – наименьшая выпуклая оболочка, которая охватывает A и B;

cross_entropy (перекрёстная энтропия) – этот критерий вычисляет потерю перекрёстной энтропии между входом и целью.

Средняя абсолютная ошибка известна как мера точности, зависящая от масштаба и имеет вид:

$$MSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 ,$$

- n – число прогнозов;
- y – изначальные данные;
- $\hat{y}$ - предсказанное значение.

RMSE (Root Mean Square Error) - это квадратный корень из значения, полученного с помощью функции среднеквадратической ошибки.

$$RMSE = \sqrt{\sum_{i=1}^{n} (Predicted_i - Actual_i)^2 / N},$$

- $Predicted$ – число прогнозов;
- $Actual$ – изначальные данные;
- N - предсказанное значение.

Заключение

Из статей, приведённых в работе видно, что для “Transformer” существует множество областей применения. От исследования аномалий производственных и в медицине до анализа помещений и эмоций. Спектр применения будет расширяться, и архитектура поможет решить новые задачи и улучшить результаты уже решённых.

Сама идея “Transformer” интересна и несложна для понимания и интеграции в готовые алгоритмы. У неё низкий порог вхождения для новичков, а открывающиеся возможности удивляют. Несомненно, алгоритм применим не везде и даёт неплохие результаты при определённых условия и в определённых задачах, но он живёт относительно недолго и не исчерпал сферы применения.

Литература

1. End-to-End Object Detection with Transformers / J Nicolas Carion [и др.] // Lecture Notes in Computer Science. – 2020. – С. 213-239.
2. Attention Is All You Need / Ashish Vaswani [и др.] // Advances in Neural Information Processing Systems. – 2017. – С. 5998-6008
3. Zhiqing Sun / Rethinking Transformer-based Set Prediction for Object Detection / Sun Zhiqing, Cao Shengcao, Yang Yiming, Kris Kitani // IEEE/CVF International Conference on Computer Vision (ICCV). – 2021.
4. Attention is not all you need pure attention loses rank doubly exponentially with depth [Электронный ресурс] : arXiv / Yihe Dong – Режим доступа: https://arxiv.org/abs/2103.03404
5. Visual Transformers: Token-based Image Representation and Processing for Computer Vision [Электронный ресурс] : arXiv / Bichen Wu – Режим доступа: https://arxiv.org/abs/2006.03677
6. PQ-Transformer: Jointly parsing 3d Objects and Layouts From Point Clouds / Xiaoxue Chen [и др.] // IEEE Robotics and Automation Letters. – 2022. – С. 2519-2545

7. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer / Shivani Kumar [и др.] // Knowledge-Based Systems. – 2022. – № 240. – С. 108112-108136.

8. Graph transformer network with temporal kernel attention for skeleton-based action recognition / Yanan Liu [и др.] // Knowledge-Based Systems. – 2022. – № 240. – С. 108146-108157.

9. Cao, Z. Content-Adaptive and Attention-Based Network for Hand Gesture Recognition, 3D interacting hand pose and shape estimation from a single RGB image / Zongjing Cao, Yan Li, Byeong-Seok Shin // Applied Sciences. – 2022. – Т. 12. – № 4. – С. 2041-2056.

10. Hybridizing Convolutional Neural Network for Classification of Lung Diseases/ Mukesh Soni [и др.] // International Journal of Swarm Intelligence Research (IJSIR). – 2022. – Т. 13. – № 2. – С. 1-15.

11. UTRAD: Anomaly detection and localization with U-Transformer / Liang Chen [и др.] // Neural Networks. – 2022. – Т. 147. – С. 53-62.

12. Spatial-Temporal Convolutional Transformer Network for Multivariate Time Series Forecasting/ Lei Huang [и др.] // Sensors. – 2022. – Т. 22. – № 3. – С. 841-859.

13. Cross-Modal Object Detection Based on a Knowledge Update / Yueqing Gao Huang [и др.] // Sensors. – 2022. – Т. 22. – № 3. – С. 1338-1353.