
OBJECT DETECTION VIA ALTERNATIVE TRANSFORMER

UDC 004.855.5

DOI: 10.56246/18294480-2023.15-12

NEYCHEV RADOSLAV

PhD in Ttechnical Sciences,

Moscow institute of Physics and Technology

Deputy head of the department of Machine Larning

e-mail: radoslav.neychev@gmail.com

STEPANYAN ARMAN

PhD student of the Department of Machine Learning and Digital Humanities,

Phystech School of Applied Mathematics and Computer Science, MIPT

Slash.Digital GmbH, Chief Technical Officer

e-mail: stepanyan.aa@phystech.edu

Transformer is a great building block for state-of-the-art models for every direction in artificial intelligence (AI). In natural language processing, GPT3 is one of the leading language models, and its fine-tuning leads to the creation of automation products in various fields (chatting, analysis of data, content generation, etc.). In computer vision, it is DALL-E-2 with its fantastic capability to generate realistic art images with the desired style. In reinforcement learning (RL) decision transformers [1] are widely used and achieve great results in such RL baseline games as ATARI, Key-To-Door tasks, etc. Even though modern transformer blocks for end-to-end object detection tasks converge very slowly, which makes the training process computationally hard. We introduce alternative transformers improving architecture and training process, reducing convergence and training time with achieving same results in object detection tasks. Training process is parallelized, and a loss function is modified to increase model's capability in multiple tasks. Finally, this architecture can be used as a building block in other models, improving their performance.

Keywords: *Transformer, controlled training, reinforcement learning, loss function, deep learning, encoder, decoder, feature map, decision making, fine-tuning.*

Introduction: New advances in artificial intelligence (AI) based object detection provide many components and frameworks to automate the full pipeline, but they are still not end-to-end. Recently, Xizhou et al. ([1]) proposed Deformable transformers (DETRs) and first provided an end-to-end model fully

getting rid of hand-crafted components achieving state-of-the-art results. DETR's architecture combines convolutional neural networks (CNNs) for feature map construction and transformer-based encoders and decoders. Despite the powerful idea behind its architecture, DETR-s seek learning convergence speed, which results in spending a large amount of computational resources. For example, on ImageNet dataset, Faster R-CNN converges 30 times faster on average than DETR ([2]). Such an issue mainly appears because of the number of Transformer building components processing feature maps and initialization way of attention weights, which are uniformly based. Because of that, more training epochs are needed to learn the sparse meanings of different locations on the image. We propose *alternative transformers'* (ATs) idea with an additional similar operation like deformable convolution (defConv in [3]) block, which avoids the abovementioned issue of slow training. The most important note here is that DETR solves the element relation mechanism that defConv has.

Method: To solve the problem of interactions of the transformer with each part of the feature map, we need to choose another building block that may get rid of or soften it. For that, we propose *an alternative attention module* - it interacts with only a tiny subset of crucial sampling points around our current point (region) in the feature map. Architecture is provided in Fig. 1. Not formally, by assigning small-sized keys to each query, we solve the problem with convergence.

Formally, we have given a feature map $x \in R^{C \times H \times W}$, query element q , content feature z_q and reference point p_q . We define alternative attention function as:

$$AT(z_q, p_q, x) = \sum_{m=1 \dots N} W_m \left(\sum_{q,k} A_{mqk} W_m' x(p_q + \Delta p_{mqk}) \right)$$

Where A_{mqk} denotes normalized attention weight, Δp_{mqk} – sampling offset, $\Delta p_{mqk} \in R^2$ are of 2-d real numbers. Both $\Delta p_{mqk}, A_{mqk}$ are linear projections over z_q . There are $2MK$ channels are used to encode Δp_{mqk} , and remaining channels are fed to *softmax* layer to get A_{mqk}

To achieve higher accuracy from our model, we will use multi-scale feature maps as other object detection frameworks use. AT can be applied to multi-scale feature map $\{x^l\}_{l=1}^L$ in a straightforward way with rescaling function to the input feature map of the l -th level $\phi_l(p_q^*)$, where p_q^* are normalized coordinates.

$$MSAT(z_q, p_q^*, \{x^l\}_{l=1}^L) = \sum_{m=1}^M W_m \left[\sum_{l=1, k=1}^{L, K} A_{mlqk} * W'_m x^l (\phi_l(p_q^*) + \Delta p_{mlqk}) \right]$$

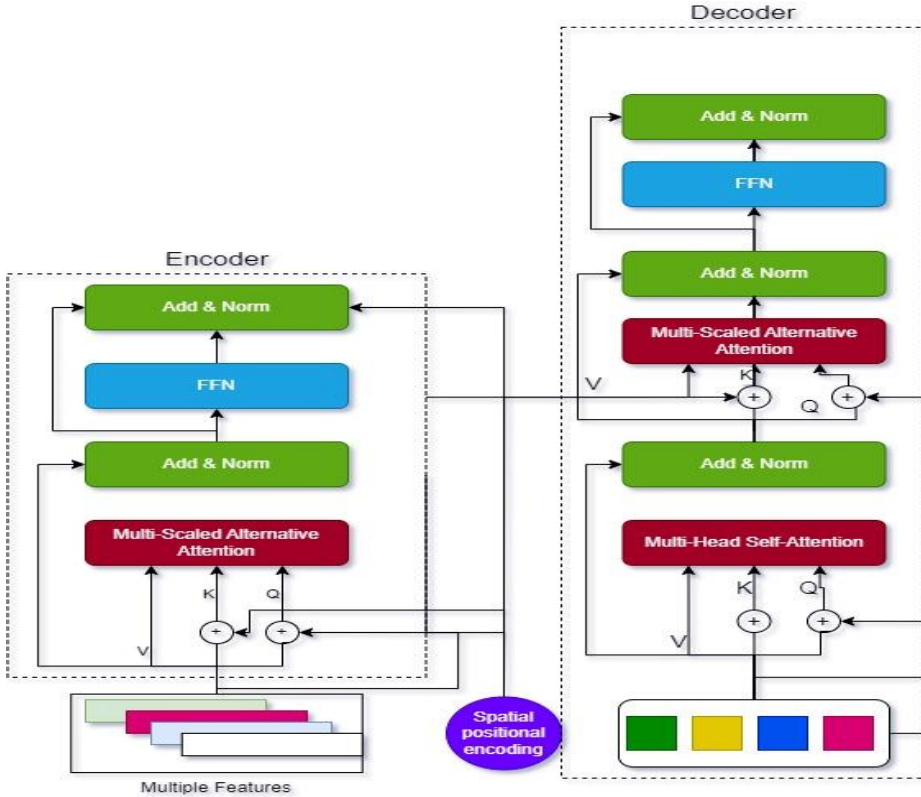
Where M is the total number of attention heads, L -number of input feature levels, K -sampling size, W'_m is fixed size identity matrix. It is important to mention that scalar attention weight A_{mlqk} is normalized. Also, it is an intuitive definition for an operation because in case $L = 1, K = 1$ it is a defConv layer, which functionality we were trying to get as a result to get rid of the convergence problem.

Encoder: The encoder is similar to DETR's encoder with one big difference - we replace the multi-headed attention module with MSAT to process multiple feature maps the way that both input and output matrices of the encoder are equal in size. We use C_5 and C_6 of ResNet ([4]). After getting our result feature map with the lowest resolution, it is processed to decoder. The number of channels in each feature is 256. The output of encoder as a result, are multiscale feature maps with the input image size. Key and query elements are from these maps too. Along positional encoding, here we use scale-level embedding, which is randomly initialized and trained along with other components of the network.

Decoder: In decoder, there are 2 types of attention used – cross and self-attention blocks. The query elements for both cases are taken from object queries. In self-attention, queries interact with each other. In the cross-attention layer, they get the information from feature maps. As long as MSAT was designed to interact with feature maps, we are changing cross-attention layers to it. Important to note that because MSAT extracts image features around goal point, then we can let our model to predict boxes around these features to let it reduce the optimization difficulty. More formally, let goal point p_q^* have coordinates (p_{qx}^*, p_{qy}^*) . We use linear projection b_q with sigmoid σ , so box coordinates will be $b_q^* = \{\sigma(b_{qx} + \sigma^{-1}(p_{qx}^*)), \sigma(b_{qy} + \sigma^{-1}(p_{qy}^*)), \sigma(b_{qw}), \sigma(b_{qh})\}$. The use of sigmoid and inverse of it is that learned decoder will strongly correlate with predicted boxes. The architecture is of MSAT is illustrated in Fig. 1.

Improvement ideas: Alternative transformers provide a variety of opportunities for end-to-end detection models due to their fast convergence.

First idea is inspired from [5] and provides a simple and iterative bounding box mechanism (finding bounding boxes based on previous layer's prediction) to



increase performance of detection.

Fig. 1. Multi-scaled alternative attention (MSAT) architecture

Second idea is to use object queries from decoder at the current stage as long as they are not used in [1]. Being inspired by [6] and [7] there is a variant to generate region proposals during that stage. Afterwards, these regions will be fed to decoder to get further and more accurate refinement.

Experiment: We conduct experiment on ImageNet ([8]) dataset. For pre-trained model we chose ResNet – 50 ([9]) as a backbone. Multi-scaling is done without FPN. For alternative transformers we set $M = 8$ and $= 4$. The rest of hyperparameters' setting follow [1]. Loss function is Focal Loss [10] with loss weight 2.5 for bounding-box classification and object queries' number equals to

350. Models are trained for at least 50 epochs and learning decay is each 50th epoch with decay factor equals to 0.1. Optimization is done by Adam optimizer with the following set of parameters: $lr = 1.5 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.995$. Training is done on Nvidia Tesla V200.

Comparison with Faster R-Cnn and DETR results can be viewed in Table 1. As it can be seen, alternative transformers achieve better performance than DETR with 10x less epochs and at the same time achieve slightly worse training time than baseline Faster R-CNN.


Table 1. Comparison between AT, DETR and Faster R-CNN

Method	Epochs	FLOPs	Training GPU hours	Inference FPS
Faster R-CNN	109	180G	380	26
DETR	500	86G	2000	28
DETR v2	500	187G	7000	28
Alternative Transformer	50	173G	700	12
Iterative bounded and 2 stage Alternative Transformer	50	173G	390	19

Conclusion: Alternative transformer is an end-to-end object detection model, which is very effective and fast converging. In this article we introduced and built that model from scratch, provided the ways that it could be improved, and did an experiment, where proved that ATs are performing as well as modern baseline models at the same time outperforming DETR in both training speed and effectiveness sides.

References

1. Carion N., Massa F., Synnaeve G., Usunier N., Kirillov N., Sergey Z, End-to-end object detections with transformers, ECCV, 2020.
2. He K., Zhang X., Ren S., Sun J., Deep residual learning for image recognition, CVPR, 2016.
3. Dai J., Qi H., Xiong Y., Li Y., Zhang G., Hu H., Wei Y., Deformable convolutional networks, ICCV, 2017.
4. He K., Zhang X., Ren S., Sun J., Deep residual learning for image recognition, CVPR, 2016.
5. Teed Z., Deng J.. Raft: Recurrent all-pairs field transforms for optical flow, ECCV, 2020.

- 
6. Ren S., He K., Girshick R., Sun J., Faster R-CNN, Towards real-time object detection with region proposal networks, Advances in neural information processing systems, 2015.
 7. Lin T.Y., Dollar P. Girshick R., He K., Harihan B., Belongie S., Feature pyramid networks for object detection, 2017.
 8. Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., Imagenet: A large-scale hierarchial image database, CVPR, 2009.
 9. He K., Zhang X., Ren S., Sun, J., Deep residual learning for image recognition, CVPR, 2016.
 10. Lin T.-Y., Goyal P., Girshick R., He K., Dollar P., Focal loss for dense object detection, ICCV, 2017.

ՕԲՅԵԿՏԻ ՀԱՅՏՆԱԲԵՐՈՒՄ ԱՅԼԸՆՏՐԱՆՔԱՅԻՆ ՏՐԱՆՍՖՈՐՄԵՐԻ ՄԻՋՈՑՈՎ

ՆԵՅՉԵՎ ՌԱԴՈՍԼԱՎ

*Մոսկվայի ֆիզիկապրեխնիկական համալսարանի
մեքենայական ուսուցման ամբիոնի վարիչ,
պրեխնիկական գիտությունների թեկնածու
էլիոստ' radoslav.neychev@gmail.com*

ՍՏԵՓԱՆՅԱՆ ԱՐՄԱՆ

*«Սլեշ.Դիջիթալ» ՍՊԸ պրեխնիկական տնօրեն,
Մոսկվայի ֆիզիկապրեխնիկական ինստիտուտի կիրառական մաթեմատիկայի
և ինֆորմատիկայի դպրոց,
մեքենայական ուսուցման և թվային հումանիտար
գիտությունների ամբիոնի ասպիրանտ
էլիոստ' stepanyan.aa@phystech.edu*

Տրանսֆորմերը հիանալի կառուցվածքային բլոկ է ժամանակակից մոդելների, արհեստական ինտելեկտի (AI) բոլոր ուղղությունների համար: Բնական լեզվի մշակման մեջ GPT3-ը առաջատար լեզուների մոդելներից է, և դրա ճշգրտումը հանգեցնում է տարբեր ոլորտներում ավտոմատացման արտադրանքների ստեղծմանը (չաթ, տվյալների վերլուծություն, բովանդակության ստեղծում և այլն): Համակարգչային տեսողության մեջ դա DALLEE-2-ն է՝ ցանկալի ոճով իրատեսական արվեստի պատկերներ ստեղծելու իր զարմանալի ունակությամբ: Ուժեղացման ուսուցման մեջ (RL) որոշման տրանսֆորմերները ([1]) լայնորեն օգտագործվում են և մեծ արդյունքների են հասնում այնպիսի RL ելակետային խաղերում, ինչպիսիք են ATARI-ը, Key-To-Door առաջադրանքները և այլն: Չնայած ժամանակակից տրանսֆորմատորային բլոկները ծայրից ծայր օբյեկտների հայտնաբերման առաջադրանքների համար միանում են շատ դանդաղ, որը հաշվողականորեն դժվարացնում է վերապատրաստման գործընթացը: Մենք ներկայացնում ենք այլընտրանքային տրանսֆորմերներ, որոնք բարելավում են կառուցումը և մարզելու գործընթացը՝ նվազեցնելով կոնվերգենցիան և ուսուցման ժամանակը, ձեռք բերելով նույն արդյունքները օբյեկտների հայտնաբերման առաջադրանքներում: Ուսուցման գործընթացը զուգահեռացվում է,

և կորստի ֆունկցիան փոփոխվում է՝ մեծացնելու մոդելի կարողությունը բազմաթիվ առաջադրանքներում: Ի վերջո, այս ճարտարապետությունն ինքնին կարող է օգտագործվել որպես շինանյութ այլ մոդելներում՝ բարելավելով դրանց կատարումը:

Բանալի բառեր՝ տրանսֆորմեր, վերահսկվող ուսուցում, ուժեղացման ուսուցում, կորստի ֆունկցիա, խորը ուսուցում, կոդավորիչ, ապակոդավորիչ, խաղարկային քարտեզ, որոշումների կայացում, ճշգրտում:

ОБНАРУЖЕНИЕ ОБЪЕКТОВ ЧЕРЕЗ АЛЬТЕРНАТИВНЫЙ ТРАНСФОРМЕР

НЕЙЧЕВ РАДОСЛАВ

Кандидат технических наук

Заместитель заведующего кафедрой машинного обучения

Московского физико-технического института

электронная почта: radoslav.neychev@gmail.com

СТЕПАНЯН АРМАН

Аспирант кафедры машинного обучения

и цифровой гуманитаристики физтех - школы

прикладной математики и информатики МФТИ

Технический директор ООО “Слэш.Диджитал”

электронная почта: stepanyan.aa@phystech.edu

Трансформер - отличный строительный блок для современных моделей по всем направлениям в области искусственного интеллекта (ИИ). В обработке естественного языка GPT3 является одной из ведущих языковых моделей, и ее тонкая настройка приводит к созданию продуктов автоматизации в различных областях (общение, анализ данных, генерация контента и т. д.). В области компьютерного зрения это DALL E-2 с его удивительной способностью генерировать реалистичные художественные изображения в желаемом стиле по описанию. В обучении с подкреплением (RL) трансформеры решений ([1]) широко используются и достигают отличных результатов в таких базовых играх RL, как ATARI, задачи Key-To-Door и т. д. Несмотря на это, современные блоки трансформеров для сквозных задач обнаружения объектов сходятся очень медленно, что усложняет процесс обучения с точки зрения вычислений. Мы вводим альтернативные трансформеры, улучшающие архитектуру и процесс обучения, сокращающие сходимость и время обучения с достижением тех же результатов в задачах обнаружения объектов. Процесс обучения распараллелен, а функция потерь изменена, чтобы повысить способность модели в выполнении нескольких задач. Наконец, сама эта архитектура может быть использована в качестве строительного блока в других моделях для повышения их производительности.

Ключевые слова: трансформер, контролируемое обучение, обучение с подкреплением, функция потерь, глубокое обучение, кодировщик, декодер, карта признаков, принятие решений, тонкая настройка.