Exploring Modulated Detection Transformer as a Tool for Action Recognition in Videos

Tomás Crisol[1⋆], Joel Ermantraut[1⋆⋆], Adrián Rostagno[1], Santiago L. Aggio[1,2], and Javier Iparraguirre[1]

[1] Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca, Argentina
tomascrisol12,joelermantraut@gmail.com
arostag@frbb.utn.edu.ar
j.iparraguirre@computer.org
[2] CONICET, Bahía Blanca, Argentina
slaggio@criba.edu.ar

Abstract. During recent years transformers architectures have been growing in popularity. Modulated Detection Transformer (MDETR) is an end-to-end multi-modal understanding model that performs tasks such as phase grounding, referring expression comprehension, referring expression segmentation, and visual question answering. One remarkable aspect of the model is the capacity to infer over classes that it was not previously trained for. In this work we explore the use of MDETR in a new task, action detection, without any previous training. We obtain quantitative results using the Atomic Visual Actions dataset. Although the model does not report the best performance in the task, we believe that it is an interesting finding. We show that it is possible to use a multi-modal model to tackle a task that it was not designed for. Finally, we believe that this line of research may lead into the generalization of MDETR in additional downstream tasks.

Keywords: Multi-modal transformers · Action detection · Model generalization.

1 Introduction

Transformers architectures have been increasing in popularity among the machine learning community [5]. Initially, this type of architecture emerged in the natural language processing space [9]. However, it is possible to observe a rapid expansion in other modalities such as computer vision [5]. Recently, multi-modal transformers enabled the possibility to process images and text using a single model. Additionally, video understanding tasks were tackled by transformers models such as the work proposed by Wang *et al*. [10].

Modulated Detection Transformer (MDETR) [4] is a multi-modal transformer. The architecture accepts an image and text as input, and it can be trained on multiple downstream tasks. One particular task is visual question answering. Although the initial design of the model does not target video understanding, we used MDETR as an

⋆ These authors contributed equally
⋆⋆ These authors contributed equally

action recognition model. Without any additional training, we evaluated the performance of the model on an action recognition dataset. Naturally, the results are not the best reported. However, we found it valuable to assess the use of a multi-modal transformer in tasks that it was not designed for. It is important to highlight that no previous training was performed before the evaluation.

The Atomic Visual Actions (AVA) [2] dataset consists of a collection of 430 videos annotated with 80 visual actions. It contains 1.58M action labels associated with a bounding box. Since MDETR provides coordinates that are related to the output, it is possible to ask a question and get an answer with the related area of interest. We ran experiments on AVA and we obtained quantitative results. Additionally, all reported findings were published in an open repository [3]. Next section explores the related work. In section 3 quantitative results are presented. Finally, conclusions are stated in section 4.

2 Related Work

2.1 MDETR

Modulated Detection Transformer (MDETR) [4], performs object detection in conjunction with language understanding. The concept enables end-to-end multi-modal understanding. The model relies only on text and the aligned boxes in an image. Unlike previous detection methods, MDETR detects concepts from free text and generalizes to unseen combinations of categories and attributes. Quantitative results reported by MDETR authors are outstanding in four tasks. Reported categories were phase grounding, referring expression comprehension, referring expression segmentation, and visual question answering. Given a collection of videos, we sampled the clips and extracted 1 frame per second. Using visual question answering, we asked for actions and measured the output of the system.

As any transformer, MDETR was trained in two stages, the pre-training and the downstream tasks. During pre-training, the model ingested a combination of Flickr30k [8], MS COCO [7] and Visual Genome (VG) [6] datasets. In the case of the visual question answering task, GQA [3] was the selected dataset. During inference, the model reads a tensor of linear image features and a tensor of linear text features. Then, the input is concatenated and fed into an encoder. Depending on the task, the decoder presents some variation. In the case of question answering, object queries and specific queries are fed into the decoder. As a result, the decoder provides new object positions and answers to the queries.

2.2 AVA

AVA [2] dataset is a person centered corpus, annotated at a 1 Hz sample rate. Every person is located using a bounding box and the labels correspond to actions related to the pose, interactions with objects, and interactions with other persons. The temporal context of the annotation is centered $\pm$ 1.5 second around the keyframe. This “brief”

[3] https://github.com/BHI-Research/AVA_MDETR

time lapse gives the name to the dataset. Annotations in the dataset are precise and the number of labels reaches 1.58 million.

Multiple metrics are available in AVA. Intersection-over-union (IoU) is reported at frame level and at video level. In the case of frame level, the metric is built following the standard protocol used by the PASCAL VOC challenge [1]. Average precision (AP) is computed using an IoU threshold of 0.5. Mean Average Precision (mAP) is the average of AP over all classes. In this work, AP is reported.

3 Results

Experiments reported on this work were obtained using the original MDETR model and the AVA actions dataset v2.2. Since the model takes images and text as input, we sampled the videos at 1 Hz to obtain the frames. Regarding actions, we created a collection of questions that ask for the actions vocabulary annotated in AVA. For each frame extracted from the dataset, we asked all the questions available. The output of the model was saved in a CSV file. Afterwards, we obtained quantitative results.

Figure 1 shows a correct action detection (left) and an incorrect result (right). In this case, the frames belong to the AVA dataset and the model used is MDETR. A frame and a question about the action to detect are given to the model as input. As output, the model provides an answer, its confidence, and the location in the image where the answer was found.

(a) (b)

Fig. 1: Output of MDETR using the visual question answering used to detect and action. On the left, successful results can be observed. In this case, an image and the question “is someone sitting?” are given to the model. On the right, the image and the question “is someone dancing?” were given to the model. The example on the left shows a failure in the action detection.

State of the art results show that the action detection task is far from solved. Up to our best knowledge, the best performing model achieved 38.8 mAP [11]. Since in our experiments we are using the standard MDETR model, not new training was required. In our case, the overall performance of the model is orders of magnitude below the best reported results. This was an expected outcome.

Table 1 shows quantitative results where MDETR performed the best and where it did not detect actions. Depending on the point of view the results can be interpreted as negative or positive. The negative aspect is that the model cannot reach results as the models designed to achieve the task specifically. The positive aspect is that MDETR is detecting actions without any additional training. This is a remarkable fact considering that the original design was targeting other tasks.

Table 1: Table captions should be placed above the tables.

Pascal Boxes Categories Results			
Best Performance		**Worst Performance**	
Category	**AP@0.5IOU**	**Category**	**AP@0.5IOU**
sleep	0.0019	answer phone	0.0
sit	0.0016	kiss (a person)	0.0
stand	0.0011	throw	0.0
hand shake	0.0005	touch (an object)	0.0
dance	0.0003	write	0.0

4 Conclusions and Future Work

In this work we showed the use of an end-to-end text and image understanding transformer model in a task that it was not designed for. We obtained quantitative results using a challenging action recognition dataset and we tested the limits of the architecture. The remarkable characteristic that makes MDETR unique is that the model can infer over classes that it did not see before. For instance, it can detect a pink elephant (not present in the annotations). We wanted to push this aspect to this limit in the case of action detection. Although the model achieves poor quantitative results, it is possible to detect actions. This is an outstanding achievement considering the scenario of experiments.

As future work, we plan to train MDETR in the action detection task. We understand that there is potential in this line of research. Since the AVA dataset provides a high number of labels, the task seems feasible. We believe that there is room for generalization in the use of multi-modal transformers models.

References

1. Everingham, M., Eslami, S., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
2. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6047–6056 (2018)
3. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
4. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021)
5. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM Computing Surveys (CSUR) (2021)
6. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
7. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
8. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
10. Wang, J., Bertasius, G., Tran, D., Torresani, L.: Long-short temporal contrastive learning of video transformers. arXiv preprint arXiv:2106.09212 (2021)
11. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14668–14678 (2022)