AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "TRANSFORMERS FOR COMPUTER VISION"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu

Authors	**Ioannis Pitas** (Aristotle University of Thessaloniki) **Vasileios Mygdalis** (Aristotle University of Thessaloniki)

This report is part of the deliverable D2.3 - "*AI technologies and applications in media: State of Play, Foresight, and Research Directions*" of the AI4Media project.

You can site this report as follows:

F. Tsalakanidou et al., Deliverable 2.3 - AI technologies and applications in media: State of play, foresight, and research directions, AI4Media Project (Grant Agreement No 951911), 4 March 2022

This report was supported by European Union´s Horizon 2020 research and innovation programme under grant number 951911 - AI4Media (A European Excellence Centre for Media, Society and Democracy).

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Copyright

© Copyright 2022 AI4Media Consortium

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the AI4Media Consortium. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.
All rights reserved.

Transformers for computer vision tasks

Current status

During the last decade, we have experienced the vast power of different types of deep neural networks in a variety of application tasks. Transformer models constitute a novel type of artificial neural networks that joined the family of already well-established neural network architectures such as multi-layer perceptrons (MLPs), convolutional neural networks (CNNs), recurrent neural networks (RNNs) and many more. ***Transformers***[1], originally introduced for language machine translation, demonstrated unprecedented performances in a range of NLP tasks such as text classification, text summarisation and question/answering systems. One such famous language model is the GPT model from OpenAI[2], which is capable of generating text that is rather difficult to determine whether or not it is written by a human. Following the breakthroughs in the NLP domain, a great interest emerged from the computer vision community in order to adapt Transformer models in computer vision. To this extent, during the last two years, Transformers have been applied in some standard computer vision tasks achieving superior results.

So far, CNNs have been the de-facto approach for tackling computer vision tasks. However, the convolution operation comes with certain shortcomings, including their inefficacy to capture long-range dependencies such as relations between pixels in an image that are distant. For example, an early convolution layer of a model trained to recognise faces can encode information about whether certain face features such as “eyes”, “mouth” or “nose” are present in an image, but these representations will not contain information such as “eyes are above nose” or “mouth are below nose”. The formidable power of Transformers, on the other hand, derives from the ***attention operation***. Inspired by the concept that humans tend to pay attention to certain factors when processing information, the attention mechanism was developed in order to direct neural networks to focus only on important parts of the input data (Figure 1). The attention mechanism utilises an entire image as context, as opposed to the convolution operation which is meant to operate on a small fix-sized region at a time. Moreover, the attention mechanism enhances the most important parts of an image while simultaneously discarding the rest. Despite their apparent higher complexity, Transformer models with attention have yielded state-of-the-art results in some standard computer vision tasks such as image recognition, object detection, image segmentation, image generation and much more.

[1] A. Vaswani et al., “Attention is all you need,” in Advances in neural information processing systems, pp. 5998–6008, 2017.
[2] T. B. Brown et al., “Language models are few-shot learners,” arXiv preprint arXiv:2005.14165, 2020.

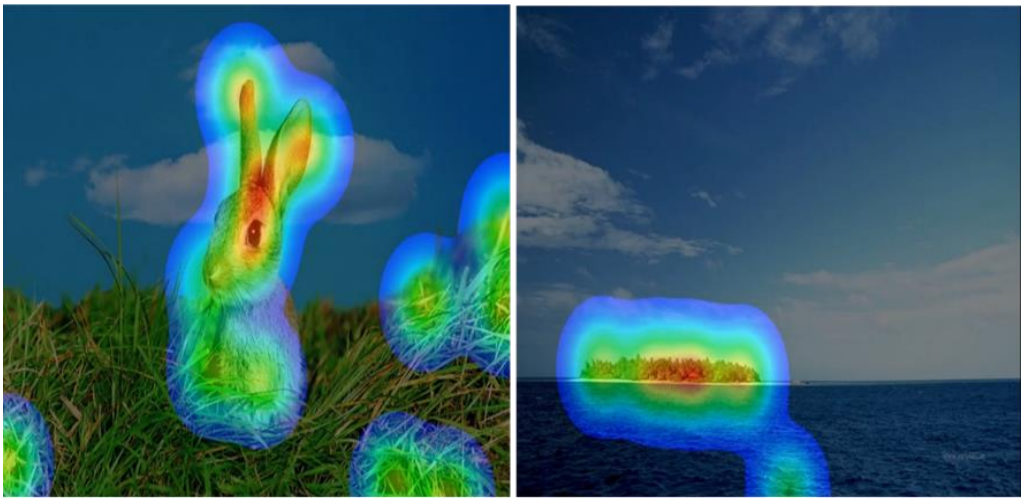

Figure 1: Visualisation of the attention operation applied on images [3].

Research challenges

Despite their excellent performance in the aforementioned visual tasks, Transformer models suffer from certain challenges and limitations associated with their applicability to practical settings. Perhaps their most prominent bottlenecks include a) the requirement of a vast amount of data (either labelled or unlabelled) for training purposes, and b) the associated high computational complexities derived from the attention operation.

As far as the first bottleneck is concerned, it has been shown that, depending on the visual task at hand a Transformer model requires from a few hundred thousand to hundreds of millions training images. For example, the ViT model[4] requires over 300 million of image examples in order to achieve comparable performance on the ImageNet benchmark dataset. It is obvious that for many specific applications (e.g., the recognition or detection of a rare animal species) it is not possible to obtain the required amount of training images. In such a scenario a common trick is to utilise information from a pre-trained model that has been trained with adequate training samples in a different domain (see section on "*Learning with scarce data*").

The second bottleneck is strictly related to the high time and memory costs imposed by the attention operation. As already discussed, attention utilises the entire images in order to extract meaningful information regarding their relations. Furthermore, it is obvious that when dealing with images the working dimensional space can grow exponentially and so do the aforementioned costs. This can lead to high training and inference times that in certain cases become unacceptable. However, the questions of designing efficient, low-complexity Transformers that can moreover work in a data-efficient manner are open research problems and recent works report encouraging steps towards their resolution.

[3] Image taken from: X. Hou et al., "Saliency detection: a spectral residual approach," Proc. IEEE CVPR, p. 4270292, June 2007.
[4] L. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

Societal and media industry drivers

Vignette: Visual content retrieval/manipulation in large video archives & productivity enhancement

Albert is working for a televised news broadcaster and is responsible for creating short video clips/segments that will thereby be shown during the typical evening news report. It is Friday noon, and there was a devastating flood in Thailand that took place just a few hours ago that demolished 50 buildings. His manager Mary asked him to edit a short 45-second video clip to cover the story for today's evening live broadcast. At the time, Albert had access only to a 10-second short video clip obtained by a mobile phone at the scene, and he should enrich it by incorporating content from a wealth of previously covered clips covering the theme "flood" from all over the world, stored in the broadcaster archives. Unfortunately for him though, the archive was not well documented thus he did not know which of the archived video material included floods with building demolitions. To retrieve similar video content, he runs a Transformer-based computer vision model where he employs the short clip of the mobile phone as a query. During his lunch break, the system ran through the broadcaster's rich archive and retrieved the most visually similar video segments. The retrieval was accurate because it recognised semantic information from the mobile phone clip (i.e., building, water) and matched it with the one from the stored material. However, the archived video segments were in a different resolution compared to the one captured by the mobile phone. This is not a problem for Albert though, because another model smooths out visual inconsistencies and creates a visually pleasing and coherent video result, in terms of automatic brightness adjustment, colour restoration and image upsampling. Within a short time-frame, Albert carefully stitches the video segments, adds some visual effects and his work is effectively done.

Future trends for the media sector

It becomes apparent that Transformer networks are all geared up to dominate the world of computer vision as more and more institutes and companies are attracting their attention. Since their introduction for simple computer vision tasks, such as image recognition, Transformers have been applied to some more advanced and complex tasks such as object detection[5], object tracking[6], image/instance segmentation[7] and depth estimation[8]. All of the aforementioned tasks can be applied to a wide range of real-life applications, from agriculture and road inspection to traffic surveillance and search and rescue missions. The complete or even partial automation of such applications enabled by Transformers can eliminate or minimise any associated logistic costs, complicated requirements and personnel involved.

[5] N. Carion et al., "End-to-end object detection with transformers," arXiv preprint arXiv:2005.12872, 2020.
[6] T. Meinhardt et al., "TrackFormer: Multi-Object Tracking with Transformers", arXiv preprint arXiv:2101.02702, 2021.
[7] L. Ye et al., "Cross-modal self-attention network for referring image segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[8] L. Zhaoshuo, et al. "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Another important computer vision field which Transformer-based neural networks have excelled is that of image generation[9]. This field consists of several interesting subfields such as image style transfer, image colourisation, image reconstructions, image super-resolution, and image synthesis (Figure 2). It has been demonstrated that Transformer models are able to generate coherent and realistic images or even artistic-like ones that can subsequently assist in the creation of a whole range of media content such as graphics for computer games or even real-action and animated movies and news videos. This increased interest in visual content generation is also amplified by the ever-expanding use of image/video-based social media platforms such as Instagram and TikTok where users are always excited and thrilled for the creation of new content.

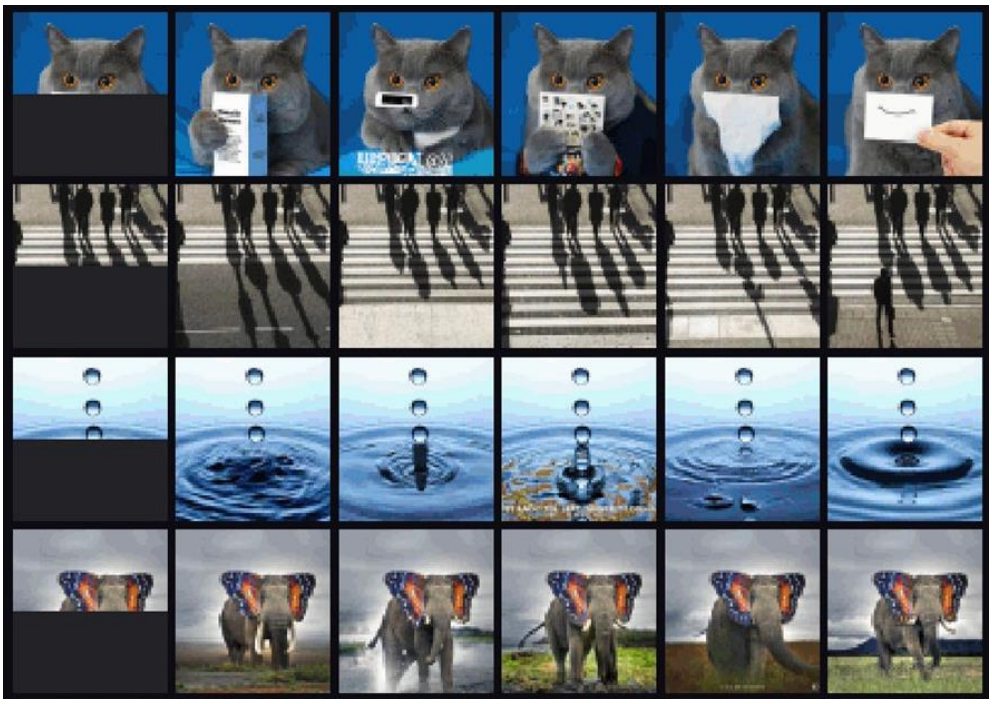

Figure 2: Image reconstruction with Transformer model[10].

Goals for next 10 or 20 years

Right now, Transformer neural networks represent one of the field's most promising present-day advances. The research on this novel type of deep neural networks is attracting more and more attention both from the academia and the industry and this attention is expected to rise significantly as, per many leading pioneers, we have only just scratched the surface. Inspired by their successes in a wide range of different individual downstream tasks, such as modelling natural language, images, proteins and behaviour amongst few the trend has shifted towards exploring ways to constitute Transformers as universal computation engines so that a single

[9] Y. Jiang et al., "TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up", arXiv preprint arXiv:2102.07074, 2021

[10] Image taken from: M. Chen, et al. "Generative pretraining from pixels." International Conference on Machine Learning. PMLR, 2020

trained model can generalise and solve a number of multiple tasks from different data domains. Transformer networks are enabling a variety of new applications and their continuous rise is a critical and important factor in the advancement of AI in the future years.

More specifically, in the next 5 years, we expect that Transformer-based architectures will become the norm in semantic metadata extraction and content retrieval, having important applications in the media industry for video archiving and indexing. In addition, within the next 10 years, due to their ability to model spatio-temporal and multi-modal relationships in a structured manner, they will become very useful in audiovisual data summarisation and will foster the automation of workflows in audiovisual content creation.

AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu