Tools for Research on Multimodal Perception and Interaction with Transformers

James Crowley
Professor, Grenoble Institut Polytechnique

1. Introduction

Recent revolutionary progress in Machine Learning, Natural Language Processing, Computer Vision, and Spoken Language Understanding has been driven by three phenomena:

1) The continued growth of available computing power made possible with GPUs and ASICs freely available via cloud computing.
2) The availability of extremely large-scale data-sets of images, video, speech and natural language for training and testing freely available over the internet
3) A movement toward challenge-based research on machine learning.

Challenge based research, in which a community publishes a data set and invites researchers to compete in writing code for tasks on this data, has been particularly important in stimulating the empirical scientific research driving much of this progress.

The synergistic effects of these three advances has been amplified by the obligation to publish the computer code and data used for experiments as a requirement for publication in major scientific conferences, and the movement to open access scientific papers with web sites such as arXiv and HAL. The result has dramatically lowered barriers to research in machine learning and related areas while opening scientific research in these areas to the entire planet.

This movement has dramatically amplified the impact of the discovery of the power of Self-Supervised learning with Transformers. Since the initial publication in 2017, Transformers and self-attention [Vaswani et al., 2017], have become the dominant approach for natural language processing (NLP) with systems such as BERT [Devlin et al., 2019] and GPT-3 [Brown et al., 2020] rapidly displacing more established RNN and CNN structures with an architecture composed of stacked encoder-decoder modules using self-attention. More recently, Transformers have had a similar revolutionary impact on Computer Vision, with a flood of transformer inspired computer vision techniques that have come to dominate the major conferences and journals. Similar impact has been observed in Speech Recognition and recent results have shown that transformers are well suited for multi-modal perception combining language and computer vision [Sun et al, 2019]. The result is a widespread availability of source code and data for research on multimodal perception with transformers. The web sites arXiv and "Papers with Code" and Code examples in Keras present the most salient examples of this movement and an excellent educational and tutorial resource. In the following, we summarize and provide pointers to some of the more salient examples of such resources.

2. Seminal Transformers

2.1 Attention is all you need

While attention has long been recognized as fundamental in biological and cognitive vision, its potential for machine learning was widely ignored until the appearance of the 2017 NeurIPS paper by Vaswani et al [Vaswani et al., 2017].

A Transformer is a model architecture that relies entirely on an attention mechanism to draw global dependencies between input and output. Before Transformers, the dominant sequence transduction models were based on complex recurrent or convolutional neural networks that include an encoder

and a decoder. The Transformer also employs an encoder and decoder, but removing recurrence in favor of attention mechanisms allows for significantly more parallelization than methods like RNNs and CNNs.

Code at:
https://github.com/tunz/transformer-pytorch/blob/e7266679f0b32fd99135ea617213f986ceede056/model/transformer.py#L201

https://paperswithcode.com/method/transformer

2.2 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT [Devlin 2018], developed by a team at Google, Research is an open source machine learning framework for natural language processing (NLP) that interprets ambiguous language in text by using surrounding text to establish context. The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with question and answer datasets.

BERT stands for Bidirectional Encoder Representations from Transformers. BERT is a conceptually simple and empirically powerful architecture that employs pre-trained deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT demonstrated the potential of Self-supervise learning using missing token replacement and next token prediction. The over all architecture is extremely flexible and is well suited for multimodal research.

Official Code
https://github.com/google-research/bert

3. Code Examples in Keras

Researchers at Google maintain an extensive repository of several hundred examples of AI systems coded in the language Keras using tensor flow. These are an excellent educational and tutorial resource for many areas of machine learning, with focused demonstrations of vertical deep learning workflows, typically expressed with less than 300 lines of code. Examples are written as Jupyter notebooks and can be run in one click in Google Colab, a hosted notebook environment that requires no setup and runs on remote cloud-based computing resources. Google Colab includes GPU and TPU runtimes programs.

The "Code in Keras" repository contains over 100 examples of Keras code for Computer Vision, Natural Language Processing, Structured Data, Times Series, audio Data, Generative Deep Learning, Reinforecment Learning, Graph Data, and Quick Keras Recipes. At the time of this writing, approximately 25% of these example concerned the use of Transformers for vision, speech or natural language processing, including demo code for BERT, ViT, SWIN, and many other classic systems.

Web Site: https://keras.io/examples/

4 Transformers for Computer Vision

4.1 ViT: Vision Transformer

The Vision Transformer (ViT) [Dosovitskiy 2021] is a model for image classification that employs a Transformer-like architecture over patches of the image. An image is split into fixed-size patches, each of which are augmented with position embeddings and used as input to a standard Transformer encoder. As with BERT, classification is performed by the addition of an extra learnable "classification token".

Code at:
https://github.com/google-research/vision_transformer
https://paperswithcode.com/method/vision-transformer

4.2 Data-efficient image Transformers (DEIT): Training with distillation through attention

Data-efficient image Transformers (DEIT) [Touvron 2020] build on the ViT architecture from [Dosovitskiy 2021] using a new distillation procedure based on a distillation token, which plays the same role as the class token that represents the label estimated by the teacher. distillation token interacts with the class token using attention. The result is a competitive convolution-free transformer, that can be trained on a single computer in less than 3 days using only data from ImageNet. DEIT has 86M parameters and achieves top-1 accuracy of 83.1% (single-crop evaluation) on ImageNet with no external data. It employs a teacher-student strategy specific to transformers and relies on a distillation token ensuring that the student learns from the teacher through attention. The authors have shown the interest of this token-based distillation, especially when using a convnet as a teacher. DEIT provides results that are competitive with deep convolutional networks for both Imagenet and when transferring to other tasks.

Paper:
https://arxiv.org/abs/2012.12877
code:
https://github.com/facebookresearch/deit

4.2 Pyramid Vision Transformer

The Pyramid Vision Transformer (PVT) [Wang 2021], is a type of vision transformer that uses a pyramid structure to as a backbone for dense prediction tasks. This allows for more fine-grained inputs (4 x 4 pixels per patch) to be used, while simultaneously shrinking the sequence length of the Transformer as it deepens - reducing the computational cost. Additionally, a spatial-reduction attention (SRA) layer is used to further reduce the resource consumption when learning high-resolution features.

The entire model is divided into four stages, each of which is comprised of a patch embedding layer and a L-layer Transformer encoder. Following a pyramid structure, the output resolution of the four stages progressively shrinks from high (4-stride) to low (32-stride).

Tasks:
Image Classification, Instance Segmentation, Object Detection, Semantic Segmentation

Code at:
https://paperswithcode.com/method/pvt
https://paperswithcode.com/paper/pyramid-vision-transformer-a-versatile#code

4.3 Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Swin Transformer [Liu 2021], is a hierarchical Transformer whose representation is computed with shifted windows. The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection. This hierarchical architecture has the flexibility to model at various scales and has linear computational complexity with respect to image size. These qualities of Swin Transformer make it compatible with a broad range of vision tasks, including image classification and dense prediction tasks such as object detection and semantic segmentation The hierarchical design and the shifted window approach also prove beneficial for all-MLP architectures.

Tasks:
Image Classification, Instance Segmentation, Object Detection, Real Teim Object Detection, Semantica Segmentation.

The code and models are publicly available at:
https://github.com/microsoft/Swin-Transformer
https://paperswithcode.com/paper/swin-transformer-hierarchical-vision

4.4 DETR: End-to-End Object Detection with Transformers

DEtection TRansformer (DETR) [Carion 2020] streamlines the detection pipeline, effectively removing the need for many hand-designed components like a non-maximum suppression procedure or anchor generation that explicitly encode our prior knowledge about the task. The main ingredients of the DETR framework are a transformer encoder-decoder architecture, and a set-based global loss-functinos that forces unique predictions via bipartite matching. Given a fixed small set of learned object queries, DETR reasons about the relations of the objects and the global image context to directly output the final set of predictions in parallel.

The DETR model is conceptually simple and does not require a specialized library. DETR demonstrates accuracy and run-time performance on par with the well-established and highly-optimized Faster RCNN baseline on the challenging COCO object detection dataset. Moreover, DETR can be easily generalized to produce panoptic segmentation in a unified manner and significantly outperforms competitive baselines.

Tasks: Object Detection, Panoptic Segmentation
Code: Training code and pretrained models are available at:
https://github.com/facebookresearch/detr.

5. Vision and Language

5.1 LXMERT: Learning Cross-Modality Encoder Representations from Transformers

Vision-and-language reasoning requires an understanding of visual concepts, language semantics, and, most importantly, the alignment and relationships between these two modalities. LXMERT (Learning Cross-Modality Encoder Representations from Transformers) [Tan 2019] is a framework to learn vision-and-language connections. LXMERT, builds a large-scale Transformer model that consists of three encoders: an object relationship encoder, a language encoder, and a cross-modality encoder. LXMERT has the ability to connect vision and language semantics, thanks to pretraining with a large number of image-and-sentence pairs, via five diverse representative pre-training tasks: masked language modeling, masked object prediction (feature regression and label classification), cross-modality matching, and image question answering. These tasks help in learning both intra-modality and cross-modality relationships. After fine-tuning from our pre-trained parameters, this

model achieves the state-of-the-art results on two visual question answering datasets (i.e., VQA and GQA).

Paper at
https://arxiv.org/abs/1908.07490
Code:
https://github.com/airsplay/lxmert

5.2 CLIP (Contrastive Language-Image Pre-Training)

Contrastive Language-Image Pre-Training (CLIP) [Radford 2021] is a neural network trained on a variety of (image, text) pairs. It can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing for the task, similarly to the zero-shot capabilities of GPT-2 and 3. CLIP matches the performance of the original ResNet50 on ImageNet "zero-shot" without using any of the original 1.28M labeled examples, overcoming several major challenges in computer vision.

Paper:
https://arxiv.org/abs/2103.00020
Code
https://github.com/openai/CLIP

6. Audio Transformers

6.1 AST: Audio Spectrogram Transformer

Audio Spectrogram Transformer (AST) [Gong 2021], demonstrates that neural networks purely based on attention are sufficient to obtain good performance in audio classification. AST is the first convolution-free, purely attention-based model for audio classification.

Tasks:
Audio Classification, Audio aging, General Classification, Keyword Spotting
Code:
https://github.com/YuanGongND/ast

7 Multimodal Transformers

7.1 VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

Video-Audio-Text Transformer (VATT) [Akbari 2021] is a framework for learning multimodal representations from unlabeled data using convolution-free Transformer architectures. VATT takes raw signals as inputs and extracts multi- modal representations that are rich enough to benefit a variety of downstream tasks. VATT has been trained end-to-end from scratch using multimodal contrastive losses. Performance has been evaluted with downstream tasks of video action recognition, audio event classification, image classification, and text-to-video retrieval. The result is a modality-agnostic, single-backbone Transformer that shares weights among the three modalities. VATT has been shown to outperform state-of-the-art ConvNet-based architectures in the downstream tasks.

Tasks:
Action Classification, Action Recognition, Action Recognition win Videos, Audio Classification, General Classification, Image Classification, Temporal Action Localization, Text to Video Retrieval, Video Retrieval

Paper:
https://arxiv.org/abs/2104.11178v3

Code:
https://paperswithcode.com/paper/vatt-transformers-for-multimodal-self

Bibliography

[Vaswani 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762

[Devlin 2018] 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.

[Brown 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., and Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[Sun 2019] Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. (2019). Videobert: A joint model for video and language representation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7464-7473)

[Dosovitskiy 2021] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021, arXiv preprint arXiv:2010.11929.

[Touvron 2021] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H., . Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning* (pp. 10347-10357). 2-21 PMLR. July 2021

[Wang 2021] Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P. and Shao, L., 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 568-578).

[Liu 2021] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012-10022).

[Carion 2020] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S., 2020, August. End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213-229), Springer.

[Radford 2021] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., . Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748-8763). PMLR., July. 2021.

[Gong 2021] Gong, Y., Chung, Y.A. and Glass, J., 2021. Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778.

[Akbari 2021] Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., & Gong, B. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, *34*.

[Tan 2019] Tan, H., and M. Bansal. "Lxmert: Learning cross-modality encoder representations from transformers." *arXiv preprint arXiv:1908.07490* (2019).