Selected Paper: *Improving Language Understanding by Generative Pre-training* by Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever

Model Type:

A task-agnostic semi-supervised Natural Language Processing (NLP) model optimised for processing long-range dependencies and transferring the acquired knowledge to solve question-answering, semantic similarity assessment, entailment determination, and text classification tasks through *generative pre-training* of long and contiguous unlabelled text corpora as well as *discriminative fine-tuning*.

Model Description:

The model was developed to overcome the environmental limitation of labelled data scarcity that challenged performance improvements of discriminatively trained NLP models. By adopting an unsupervised approach at a pre-training stage, feeding the model with a long stretch of contiguous text, the model was trained to learn initial parameters while "[conditioning] on long-range information" (pg 4). In this way, the model is equipped with the capacity to perform transfer learning across diverse NLP tasks even in the absence of a large batch of labelled data. At the fine-tuning stage, the model's initial parameters are adapted to target tasks through supervised methods. With previous models, a substantial amount of new task-specific parameters are required for fine-tuning, which essentially "re-introduces a significant amount of task-specific customisation [that does not] use transfer learning for those additional architectural components" (pg 4). However, the authors' "traversal-style approach" that transforms structured inputs into an order of token sequences allows the model's architecture to flexibly adapt to task-specific discrimination without extensive changes.

Model's Architecture:

A Transformer architecture was chosen for its provision of structured memory, which is crucial "for handling long-term dependencies in the text" and the model's transfer learning (pg 2). This architecture consists of two main frameworks: the first framework is set for the generative pre-training and the second is an adaptation of the model for discriminative tasks using labelled data(C with a label y). For the first stage of unsupervised pre-training, a variant of the Transformer model called the *multi-layer Transformer decoder* was used for the language model. As the model needs to run in parallelisation, the splitting of attention calculations in each head to produce a final attention score achieves greater efficiency and language processing power. In the supervised discriminative fine-tuning process, the initial parameters are adapted to the target task, where a sequence of input tokens from a labelled dataset gets passed through the pre-trained model to obtain the final activation function. The final transformer block's activation then goes through the model's combined linear and softmax layers with the specified parameters to predict y.

Model's Capabilities/Tasks:

The model's capabilities are evaluated based on its performance of the following NLP tasks: natural language inference (NLI), question-answering, sentence similarity detection, and text classification. Other than text classification, however, the pre-trained model requires additional modifications to the specific structured inputs. However, as explained above, the model's transfer learning achieved through the traversal-style approach minimises architectural adjustments.

NLI "involves reading a pair of sentences and judging the relationship between them from one of entailment, contradiction, or neutral"; in this task, the model was shown to excel in reasoning over long multiple sentences as well as handling ambiguous linguistic features, outperforming all the state-of-the-art methods except news articles (RTE). The model's high performance in dealing with long-range contexts was also indicated in question-answering and commonsense reasoning, where the model was set to select the correct ending of the Story Cloze test for multi-sentence stories given certain options. On this task, the model showed its clear and strong competitiveness by recording the best result in all datasets compared to the current state-of-the-art methods. Semantic similarity detection or "paraphrase detection" tasks require a comparison of two sentences and predicting their similarities. In order to execute this task, an NLP model needs to recognise the concept of rephrasing and handle linguistic ambiguity. The model achieved a significant absolute improvement in all datasets: Microsoft Paraphrase Corpus (MRPC), Quora Question Pairs (QQP), and the Semantic Textual Similarity benchmark (STS-B). The model lagged behind by 4.3% on MRPC by the TF-KLD method but outperformed other state-of-the-art results by a 1-point margin on STS-B and 4.2% on QQP. As for the text classification task, the model was tested on its capability to classify linguistic bias. In this task, the model was said to have achieved 91.3% accuracy on the Stanford Sentiment Treebank's (SST-2) standard binary classification task and 45.4 on the Corpus of Linguistic Acceptability (CoLA) task, exceeding by 10.4 from the previous best result.

Model's Limitations:

This model tried to overcome the limitations of previous models for the usage of unlabeled text, and improving the task performance. Specifically, in contrast to the discriminatively trained models that utilize architectures specifically made for each task, this general task-agnostic model significantly outperforms it in the field of 9 out of 12.

However, still there were some limitations in three different ablation studies. First, by evaluating the method's performance without the auxiliary LM objective while fine-tuning, the pattern indicates that larger datasets have a positive effect from the auxiliary objective,

but smaller datasets do not. The model should be revised to have similar benefits for datasets from the auxiliary objective regardless of the size. In addition, comparing the transformer's effect to the single layer 2048 unit LSTM, the average score of the Transformer is 5.6 high compared to the LSTM. On the other hand, for one dataset called "MRPC", which consists of the sentence pair from newswire articles, the LSTM outperforms the Transformers' performance. Lastly, a sufficient amount of training is crucial in the sense of performance. Specifically, comparing the transformer architecture directly to trained on supervised target tasks, without any pre-training. It was shown that the scarcity of pre-training negatively effect on the performance across all the assignments. Consequently, compared to the full model, a 14.8% decrease has been noticed on the performance.

Selected Paper: *An Image is Worth 16 x 16 Words: Transformers for Image Recognition at Scale* by Alexy Dosvoitskiy et al. (2021)

Model Description:

Vision Transformer (hereinafter referred to as ViT) is a high-performance model in image recognition/processing, and was created by applying Transformer, which performs well in Natural Language Processing, to Computer Vision (hereinafter referred to as CV). By adopting Transformer instead of CNN, ViT can capture global image information.

ViT divides the image into small patches, vectorizes them, and extracts contextual information in the image by using Transformer's encoder. CNN is adept at analyzing regional information while ViT can understand the situation in the image. However, ViT has a lot of memory use whereas CNN uses memory compactly and is able to build a fast image processing system.

Model's Architecture:

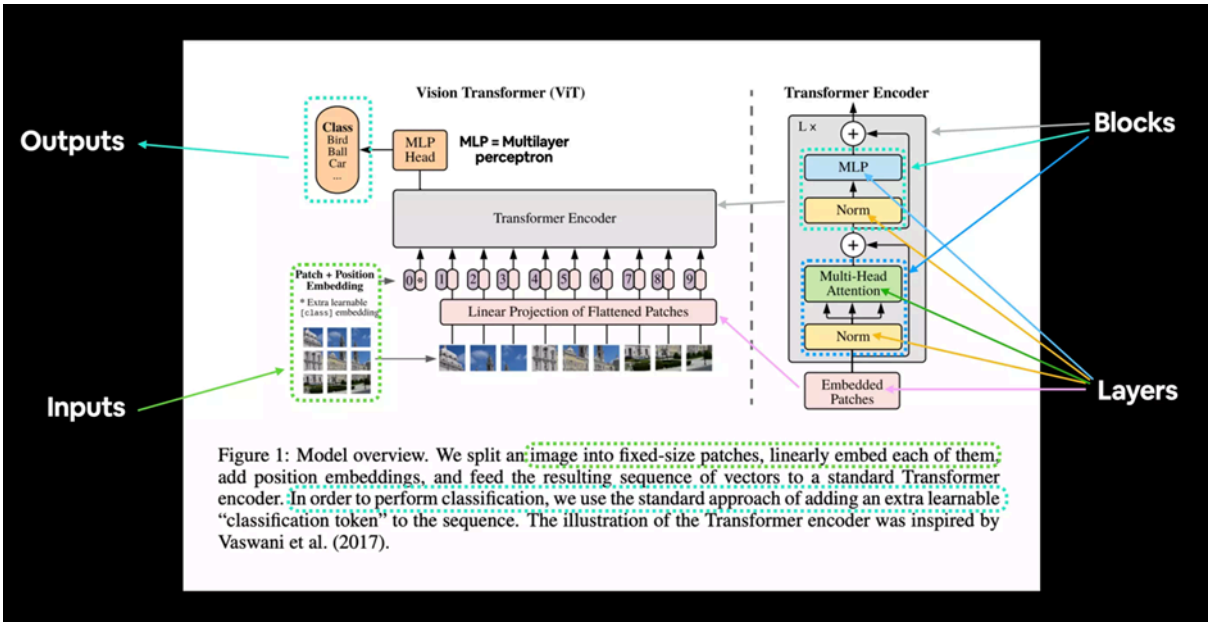


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

"Patching Images"

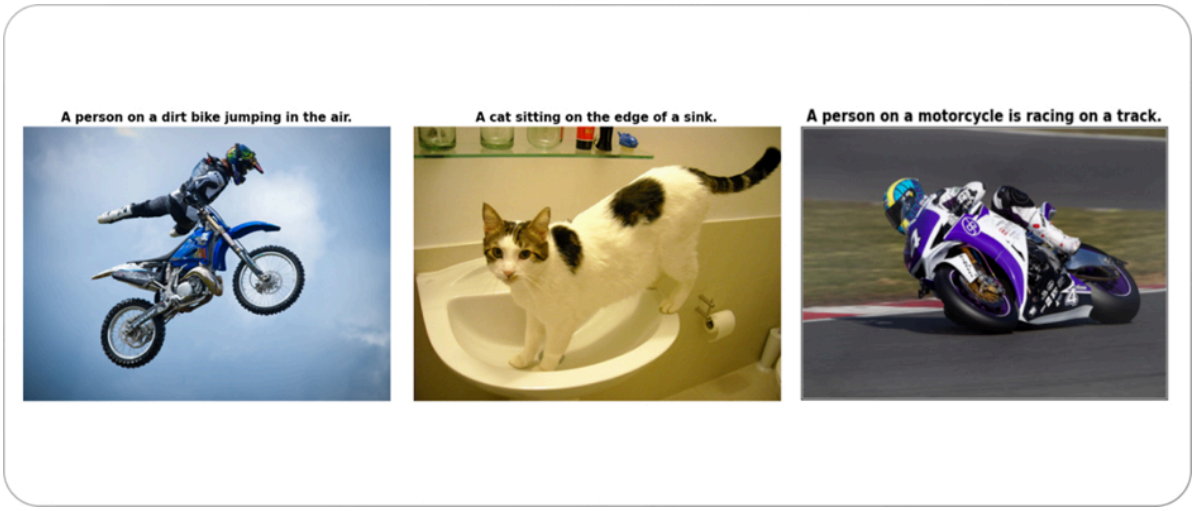
- ViT divides images into small, fixed-size, square patches.
- The image is divided into patches from 224 x 224 to 16 x 16.

1. Transform each patch into a flat one-dimensional vector, and simultaneously, add positional information of each patch. Through this, the model can understand the image information.
2. Vectorizes using Transformer encoder, and uses the patches with positional information as input to the Transformer encoder.
3. Add classification (class) tokens containing each patch's information. It will be used for classification tasks at the output of the Transformer encoder. To do this, MLP head

is used, and the vector is transformed into the final classification result with softmax activation function applied.

Model's Capabilities/Tasks:

1. Image Captioning



ViT can be used for image captioning, which is the task of receiving an image as input and generating a description/title in natural language format as the output. It is a combination of classification and Natural Language Processing tasks, and it even combines Computer Vision and Natural Language Modeling techniques.

2. Image Segmentation

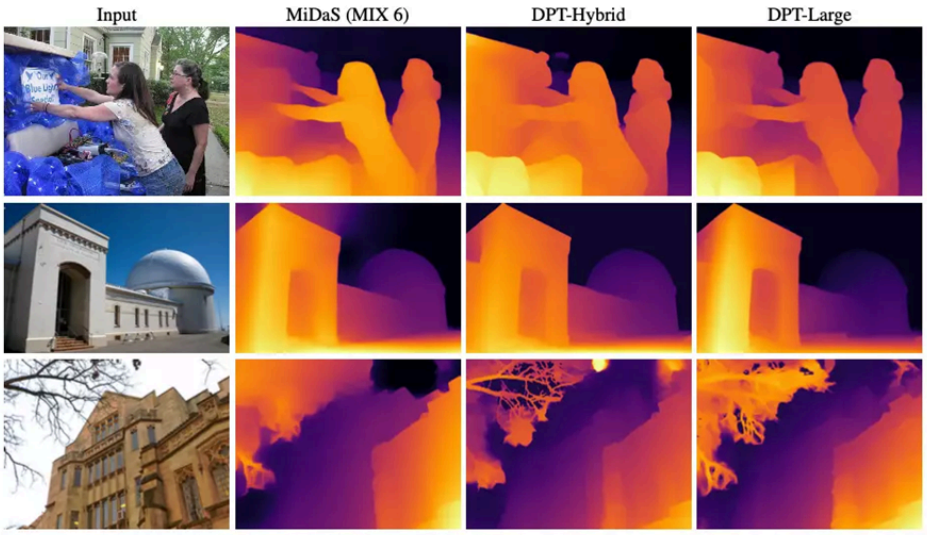
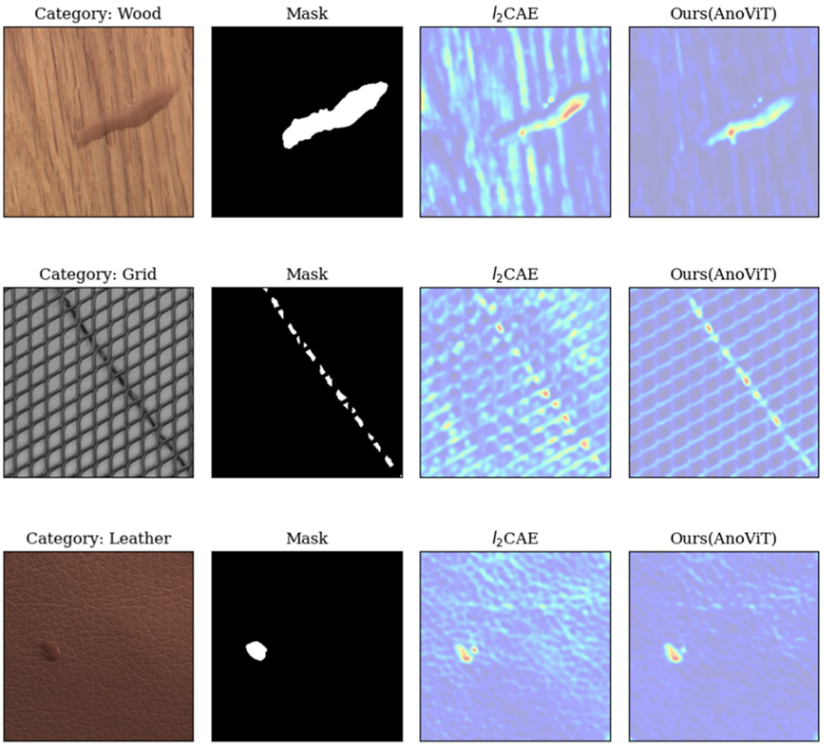


Figure 2. Sample results for monocular depth estimation. Compared to the fully-convolutional network used by MiDaS, DPT shows better global coherence (e.g., sky, second row) and finer-grained details (e.g., tree branches, last row).

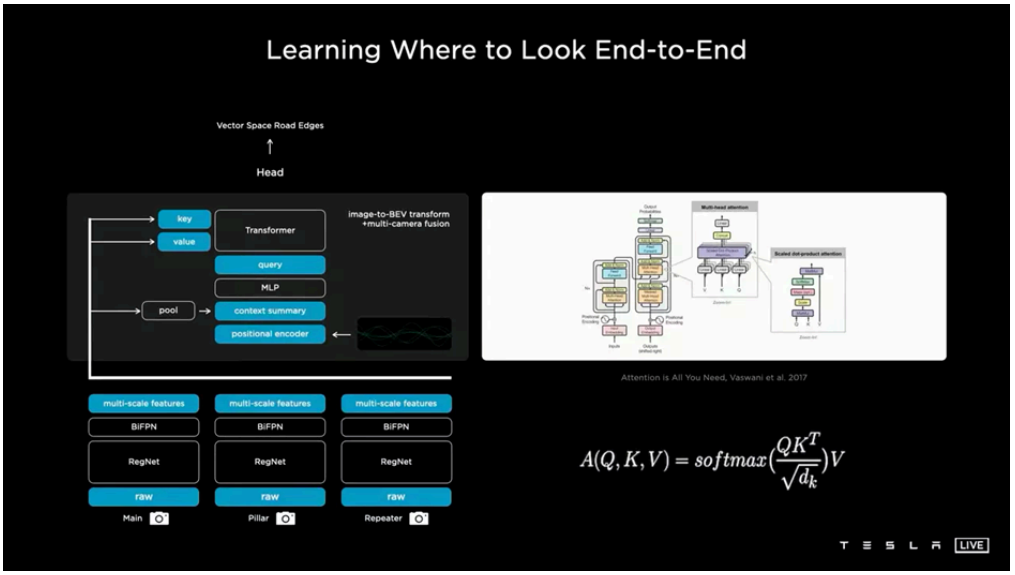
ViT even has capability of image segmentation, which is the process of analyzing images in pixel units and assigning labels containing information about boundaries of each individual object.

3. Anomaly Detection



ViT effectively incorporates local and global information using a self-attention layer that learns the relationship between image patches. After passing through the embedding of all patches, the model constructs a feature map which retains positional information of patches that have gone through multiple self-attention layers. It is also known to outperform normal CNN in Outlier detection Anomaly Detection.

4. Autonomous Driving



Transformer mode using cross-attention module l is even used as core architecture for TESLA.

In short, Vision Transformer will be the solution for Image Segmentation, Autonomous Driving, Anomaly Detection tasks rather than CNN because CNN has difficulty with understanding the context between images and capturing long range relationships. ViT has dramatically improved performance while solving these problems.

Model's Significance / Limitations:

ViT shows its importance by opening a new horizon for performance improvement and model utilization through direct application of Transformer structures, away from reliance on CNN or limited application of Attention in the field of CV. Transfer learning on small datasets using pre-trained models with huge datasets has enabled the performance comparable to existing SOTA models with less computational resources needed to pre-train. Moreover, scalability and versatility of the Transformer mechanism has led to the emergence of various architectures using ViT as backbone in CV tasks. For example, Swin Transformer, which has improved ViT models' limitation on high-resolution image processing, recorded higher performance than ResNet in object detection or segmentation tasks.

However, ViT has a noticeable limitation as its potential. First of all, learning with insufficient amounts of data can impair generalization performance. It is due to ViT's low inductive bias, which in conclusion has the disadvantage of requiring more data compared to CNN. In addition, ViT models also demonstrate limitations in applying self-supervised pre-training to CV fields, which brought many advantages in the NLP field. Even an experiment in the article proves it by showing the ViT-B/16 model achieving 79.9% accuracy on Imagenet. Of course, it's pretty great performance but it is still about 4% behind large-scale supervised pre-training. However, as various methodologies like DINO or SEER are being researched to improve self-supervised learning model's performance we can expect that it won't be a chronic limitation.

References

Lee, Yunseung, and Pilsung Kang. "Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder." *IEEE Access*, vol. 10, 2022, pp. 46717–46724, https://doi.org/10.1109/access.2022.3171559.

Liu, Ze, et al. "Swin Transformer: Hierarchical vision transformer using shifted windows." *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, https://doi.org/10.1109/iccv48922.2021.00986.

Kim, JiYeop. *[논문리뷰] Emerging Properties in Self-Supervised Vision Transformers (Dino)*, 1 Feb. 2023, kimjy99.github.io/%EB%85%BC%EB%AC%B8%EB%A6%AC%EB%B7%B0/dino/.

Moon, Sangsun. *Vision Transformer (VIT)* 란*?* - 정의*,* 원리*,* 구현*,* 응용분야 - 데이터헌트 *[What is Vision Transformer (VIT)? - Definition, Principles, implementation, applied fields - Data hunt]*, 2023, www.thedatahunt.com/trend-insight/vision-transformer.