Advances in Transformer Architectures: Integrating BERT, GPT, and Vision Models

Spandan V. Nagale

Abstract

The *Transformer* architecture (Vaswani *et al.*, 2017) introduced a novel approach to sequence modeling by dispensing with recurrence and convolution in favor of self-attention mechanisms [1]. In this paper, we survey major advancements in Transformer-based models since this original work, focusing on modern variants such as BERT, GPT, and the Vision Transformer (ViT). We outline the core Transformer architecture, describe the pre-training and fine-tuning methodology of BERT (Bidirectional Encoder Representations from Transformers) [2] and GPT (Generative Pre-trained Transformer) [3] models, and explain how self-attention has been adapted to vision tasks in the ViT model [4]. Key results from each model family are presented, illustrating how large-scale pre-training has driven state-of-the-art performance on diverse NLP and vision benchmarks. Through this analysis, we emphasize how Transformer-based models have achieved broad applicability across domains [5]. The paper follows a standard research structure with contributions including a synthesis of core Transformer concepts, a description of recent innovations, and a discussion of ongoing challenges.

Introduction

Traditional sequence transduction models relied on recurrent neural networks (RNNs) or convolutional neural networks (CNNs) in encoder-decoder configurations. However, RNNs process inputs sequentially, limiting parallelism, and both RNNs and CNNs often struggle to model long-range dependencies efficiently. The Transformer model of Vaswani *et al.* (2017) overcame these limitations by using a purely attention-based architecture [1] [6]. Instead of sequential recurrence, the Transformer uses self-attention to connect any two positions in the input (or output) directly. This design allows for far greater parallelization during training and better handling of long-range dependencies [7] [6]. The original Transformer achieved new state-of-the-art results on machine translation tasks (e.g. 28.4 BLEU on WMT14 English→German and 41.8 BLEU on English→French) while training much faster than prior models [1].

Since 2017, the Transformer architecture has become the foundation for numerous advances. Notably, researchers have developed specialized variants for different tasks: - **BERT** (Devlin *et al.*, 2019) uses a multi-layer Transformer *encoder* to pre-train deep bidirectional language representations from unlabeled text [2]. BERT demonstrated that pre-training a bidirectional Transformer on large text corpora yields powerful contextual embeddings that can be fine-tuned for many tasks (question answering, classification, etc.) [8]. - **GPT** (Radford *et al.*, 2019/2020) employs a Transformer *decoder* architecture trained autoregressively as a language model. GPT models scale up the number of parameters dramatically (e.g. GPT-3 with 175 billion parameters [3]), and show that very large models can perform many tasks (translation, reasoning, question answering) in a zero-shot or few-shot setting without task-specific training [3] [9]. - **Vision Transformers**

(ViT) (Dosovitskiy *et al.*, 2020) apply the Transformer to image recognition by splitting an image into a sequence of patches treated as tokens. A pure Transformer encoder (without convolution) can achieve competitive or superior results to CNNs on image benchmarks when pre-trained on enough data [4].

These developments illustrate that transformer-based architectures excel across domains. As Islam *et al.* (2023) observe, "transformer models have attracted substantial interest… not only in NLP tasks but also in… computer vision, audio and speech processing" [5]. This paper provides a cohesive overview of these models, connecting the original Transformer architecture to its modern incarnations. We first review the core Transformer mechanism (Section 2), then detail the model architectures and training methodologies of BERT (2.1), GPT (2.2), and ViT (2.3). In Section 3, we summarize key results for each model family. Section 4 discusses the implications of these advances and outlines current challenges. Finally, Section 5 concludes the survey.

Methodology

In this section, we outline the Transformer model architecture and describe how recent variants specialize it for different tasks.

2.1 Transformer Architecture

The original Transformer consists of an encoder-decoder stack of layers using self-attention and pointwise feed-forward sublayers [6]. A visual overview is shown in Figure 1.

Figure 1: The Transformer architecture with encoder (left) and decoder (right) stacks, each containing multi-head self-attention and feed-forward sublayers [6].

The encoder is composed of N identical layers. Each layer has two sub-layers: a multi-head self-attention mechanism, and a position-wise fully connected feed-forward network. Residual connections and layer normalization are applied around each sub-layer [6] [10]. The decoder similarly consists of N layers, but each decoder layer has an additional sub-layer that performs multi-head attention over the encoder's output. The decoder's self-attention is masked to prevent access to future positions (so that predictions for a given position only depend on earlier outputs).

Self-attention computes a weighted sum of values by comparing a query to all keys. The Transformer uses *scaled dot-product attention*: given query, key, and value matrices Q,K,V, it computes
$$ \mathrm{Attention}(Q,K,V) = \mathrm{softmax}!\Bigl(\frac{QK^T}{\sqrt{d_k}}\Bigr)\,V. $$ This enables each token to attend to all other tokens in the sequence. Multi-head attention repeats this process h times with different learned linear projections of Q,K,V, allowing the model to capture information from multiple representation subspaces [10] [11]. Sinusoidal positional encodings are added to the inputs to inject information about token order.

Overall, the Transformer architecture replaces recurrence with attention. As Vaswani *et al.* explain, this design allows the model to "draw global dependencies" and achieve significantly more parallelization [12]. Because any pair of positions can be connected in a single attention step, long-range dependencies can be learned more directly. The original Transformer (with $N=6$ encoder/decoder layers, model dimension

$d_\text{model}=512$, and 8 attention heads) contains on the order of 65 million parameters, yet it achieved strong performance on translation tasks [1].

2.2 BERT: Bidirectional Pre-training

BERT (Bidirectional Encoder Representations from Transformers) uses a stack of Transformer *encoder* layers to learn contextual embeddings from unlabeled text [2]. Its key innovation is masked language modeling: during pre-training, random tokens in the input are replaced with a special [MASK] token, and the model must predict the original token from its context. This forces the model to jointly use both left and right context to learn representations. Devlin *et al.* emphasize that BERT "pre-train[s] deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers" [2].

After pre-training, BERT can be fine-tuned on specific tasks by adding a small output layer. For example, for classification or inference tasks, the hidden state corresponding to a special [CLS] token is fed into a task-specific softmax classifier. BERT achieved dramatic improvements: Devlin *et al.* report that BERT-large (24 layers) "pushes the GLUE score to 80.5%" (an absolute gain of 7.7 points) and set state-of-the-art on SQuAD question-answering benchmarks [8]. In practice, the common configurations are BERT-base (110 million parameters, 12 layers) and BERT-large (340 million parameters, 24 layers), which have become widely used for NLP tasks.

2.3 GPT: Autoregressive Pre-training

GPT (Generative Pre-trained Transformer) models use Transformer *decoder* architectures trained as large-scale language models [3]. For example, GPT-3 is a decoder-only Transformer with 175 billion parameters, trained to predict the next word in vast text corpora [3]. Unlike BERT, GPT does not mask tokens; it relies on unidirectional (left-to-right) context.

A notable finding from GPT-3 is that massive scale can yield strong performance on many tasks without additional fine-tuning. Brown *et al.* report that by scaling model size (GPT-3 has 10× the parameters of GPT-2), the model can be applied in a zero- or few-shot manner: tasks are specified via textual prompts and the model generates answers on the fly. GPT-3 "achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks" without any gradient updates [3]. In other words, GPT-3 learns to perform tasks by inference alone. The authors also show GPT-3 can generate news articles "which human evaluators have difficulty distinguishing from articles written by humans" [13].

Architecturally, GPT uses the same masked self-attention and feed-forward blocks as the Transformer decoder. The primary difference in GPT-3 is scale: with 175B parameters, the model has capacity to capture diverse linguistic patterns. GPT-3's few-shot learning behavior suggests that extremely large Transformers can serve as general-purpose language models with minimal task-specific adaptation.

2.4 Vision Transformer (ViT)

The Vision Transformer (ViT) adapts the Transformer encoder for image recognition [4]. An image is divided into fixed-size patches (e.g. 16×16 pixels). Each patch is flattened and projected to an embedding vector, and positional embeddings are added to maintain spatial structure. This sequence of patch embeddings is fed into a standard Transformer encoder (similar to BERT). The output corresponding to a special [CLS] token is then used for classification.

Dosovitskiy *et al.* demonstrate that a pure Transformer applied to image patches can "perform very well on image classification tasks." When pre-trained on large image datasets, ViT achieves excellent results on benchmarks like ImageNet and CIFAR-100, often surpassing state-of-the-art convolutional networks while requiring fewer compute resources to train [4]. This shows that self-attention can effectively capture visual information when enough data is available, unifying the modeling approach between vision and language.

Results

Key empirical results from each Transformer variant include:

- **Transformer (Vaswani *et al.*, 2017)** – New state-of-the-art on translation: BLEU scores of 28.4 (WMT14 English→German) and 41.8 (English→French) [1], with significantly faster training than prior RNN/CNN models.
- **BERT (Devlin *et al.*, 2019)** – BERT-large achieved a GLUE score of 80.5% (a +7.7 point gain) and set new records on SQuAD v1.1 (F1=93.2) and SQuAD v2.0 (F1=83.1) [8].
- **GPT-3 (Brown *et al.*, 2020)** – Without any task-specific fine-tuning, GPT-3 (175B) matched or exceeded previous models on tasks like translation and QA [3]. Its generated text (e.g. news articles) was often indistinguishable from human writing [13].
- **Vision Transformer (Dosovitskiy *et al.*, 2020)** – ViT (Base/16, ~86M parameters) achieved approximately 77–78% top-1 accuracy on ImageNet, on par with top CNNs, and even outperformed them when training data was plentiful [4].

These results indicate that when sufficient data and compute are available, transformer-based models lead performance in both NLP and vision tasks. The common factor is large-scale *pre-training* with self-attention architectures.

Model (Year)	Params	Task	Notable Result	Source
Transformer (2017)	~65M	Translation	BLEU 28.4 / 41.8	[1]
BERT (2019)	340M (large)	GLUE	80.5% score	[8]
GPT-3 (2020)	175B	Few-shot tasks	Strong zero/few-shot performance	[3]
Vision Transformer (2020)	~86M (Base)	ImageNet	≈77–78% top-1 accuracy	[4]

Table: Representative performance of Transformer-based models on benchmark tasks. (Results and sizes from cited sources.)

Discussion

Transformer models have transformed deep learning by enabling effective modeling of global context and long-range dependencies. Attention allows any part of the input to directly attend to any other part, in contrast to the sequential bottleneck of RNNs. As Islam *et al.* (2023) note, Transformers "excel in handling long dependencies" and have produced "remarkable achievements" across NLP, computer vision, audio, and

other domains [5]. This explains why they have become ubiquitous: a single Transformer framework can be adapted to many data modalities.

A major trend is **scale and pre-training**. Both BERT and GPT rely on training on huge unannotated corpora. Brown *et al.* demonstrate that increasing model size (to 175B parameters) substantially improves *few-shot* learning [9]. This paradigm shift (often called using "foundation models") means that instead of engineering task-specific networks, we train one large model and adapt it. The implication is that more data and larger models tend to yield better performance, as seen in GPT-3 and other large LMs.

However, there are challenges. Very large Transformers require enormous compute and memory, raising concerns about efficiency and carbon footprint. They also risk encoding biases present in their training data. Consequently, recent work explores more efficient variants (sparse or local attention, model distillation) and investigates fairness and interpretability of large models. Moreover, self-attention's quadratic cost with input length motivates innovations (Longformer, Performer) for long documents.

Beyond language and vision, Transformer approaches have been applied to speech, reinforcement learning, and even scientific domains (e.g. protein sequences). Researchers are also integrating multiple modalities in a single model (e.g. Vision-Language Transformers). This cross-domain success highlights the versatility of attention-based architectures. As Transformers continue to evolve, we can expect further unification of AI models under this common framework.

Conclusion

We have surveyed the evolution of Transformer-based architectures from the original encoder-decoder model [1] through modern variants like BERT [2], GPT [3], and Vision Transformer [4]. Each of these builds on the core idea of self-attention, yet specializes it: BERT leverages bidirectional encoding and masked pre-training, GPT scales up autoregressive decoding, and ViT adapts the model to image patches. Empirical results show that these models have set new state-of-the-art in their domains, demonstrating the broad impact of attention mechanisms.

In summary, the advent of the Transformer and its descendants represents a paradigm shift. By using self-attention, these models overcome the limitations of sequential architectures and harness the power of parallel computation. When combined with large-scale pre-training, they achieve unprecedented performance on NLP and vision tasks. While challenges remain (resource demands, ethical concerns), the continued success of Transformers suggests that "Attention Is All You Need" has become a foundational principle in AI. Future work will likely focus on making these models more efficient, understanding their behavior, and extending them to new applications, building on the rich legacy of Transformer research.

References

- Vaswani *et al.* (2017). *Attention Is All You Need*. NeurIPS. [1]
- Devlin *et al.* (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL. [2] [8]
- Brown *et al.* (2020). *Language Models are Few-Shot Learners*. NeurIPS. [3] [9]
- Dosovitskiy *et al.* (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. ICLR. [4]

• Islam *et al.* (2023). *A Comprehensive Survey on Applications of Transformers for Deep Learning Tasks*. arXiv. [5]

---

[1] [1706.03762] Attention Is All You Need
https://arxiv.org/abs/1706.03762

[2] [8] [1810.04805] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
https://arxiv.org/abs/1810.04805

[3] [9] [13] [2005.14165] Language Models are Few-Shot Learners
https://arxiv.org/abs/2005.14165

[4] [2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
https://arxiv.org/abs/2010.11929

[5] [2306.07303] A Comprehensive Survey on Applications of Transformers for Deep Learning Tasks
https://arxiv.org/abs/2306.07303

[6] [7] [10] [11] [12] [1706.03762] Attention Is All You Need
https://ar5iv.labs.arxiv.org/html/1706.03762v7