Transformers Beyond NLP: A Survey on Vision, Speech, and Multimodal Applications

Harsh Gupta

July 2025

Contents

Abstract

This survey explores the expansion of Transformer architectures beyond natural language processing (NLP), examining their applications in computer vision, speech processing, and multimodal domains. We analyze the core adaptations, representative models, performance benchmarks, training methodologies, and emerging trends that highlight the versatility of transformers as general-purpose learners across modalities.

1 Introduction

Transformers, first introduced in 2017 through the seminal paper "Attention Is All You Need", revolutionized natural language processing (NLP) by replacing recurrent architectures with self-attention mechanisms. This shift enabled superior performance across a variety of NLP tasks such as machine translation, text classification, question answering, and language modeling. Models like BERT, GPT, and T5 have since become the foundation of modern NLP systems.

While originally designed for sequential language data, the core principles of transformers—such as self-attention, parallelizable architecture, and scalability—have prompted researchers to explore their applicability in other domains. In recent years, transformers have gained significant traction in areas such as computer vision, speech processing, and multimodal learning, where they have outperformed or complemented traditional models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

The motivations behind this cross-domain shift are manifold:

- **Unified modeling framework**: Transformers offer a consistent architectural design that can be adapted to various input modalities, facilitating joint training and modular integration.
- **Long-range dependency handling**: Self-attention enables transformers to capture global context more effectively than local operations used in CNNs or RNNs.
- **Scalability with data and compute**: With sufficient data and hardware, transformers continue to scale in performance, as observed in large models like GPT-4 and Google Gemini.

For instance, in computer vision, the Vision Transformer (ViT) demonstrated that pure transformer architectures could rival state-of-the-art CNNs on image classification tasks. Similarly, in speech, models like wav2vec 2.0 have showcased the effectiveness of self-supervised pretraining for automatic speech recognition. In the multimodal space, OpenAI's CLIP and DeepMind's Flamingo have shown how transformers can align and jointly model text and visual inputs with remarkable generalization capabilities.

This survey provides a comprehensive overview of the application of transformers beyond NLP. We organize the paper into several key areas:

- A review of transformer architectures and their foundational design.
- An in-depth look at transformer applications in vision and speech domains, including representative models and benchmarks.
- A detailed discussion of multimodal transformers that combine multiple modalities such as text, image, and audio.
- Identification of existing challenges, limitations, and potential research directions for cross-domain transformer models.

By consolidating recent advancements, this paper aims to highlight the versatility of transformers as general-purpose learners and their potential to serve as a unifying framework across diverse machine learning tasks.

2 Transformer Foundations: A Brief Recap

The Transformer architecture was introduced as a novel sequence transduction model, replacing the need for recurrence and convolution in sequence modeling. At its core, the transformer leverages a self-attention mechanism that enables the model to weigh the importance of different parts of the input sequence when generating representations.

Encoder-Decoder Architecture

The original transformer follows an encoder-decoder architecture. The encoder processes the input sequence and generates contextualized embeddings, while the decoder uses these embeddings to produce the target output sequence step-by-step. Each encoder and decoder block consists of a stack of layers, including multi-head self-attention, position-wise feedforward networks, residual connections, and layer normalization.

In tasks like machine translation (e.g., English to German), the encoder consumes the source sentence, and the decoder generates the translation by attending to both previously generated tokens and the encoder output.

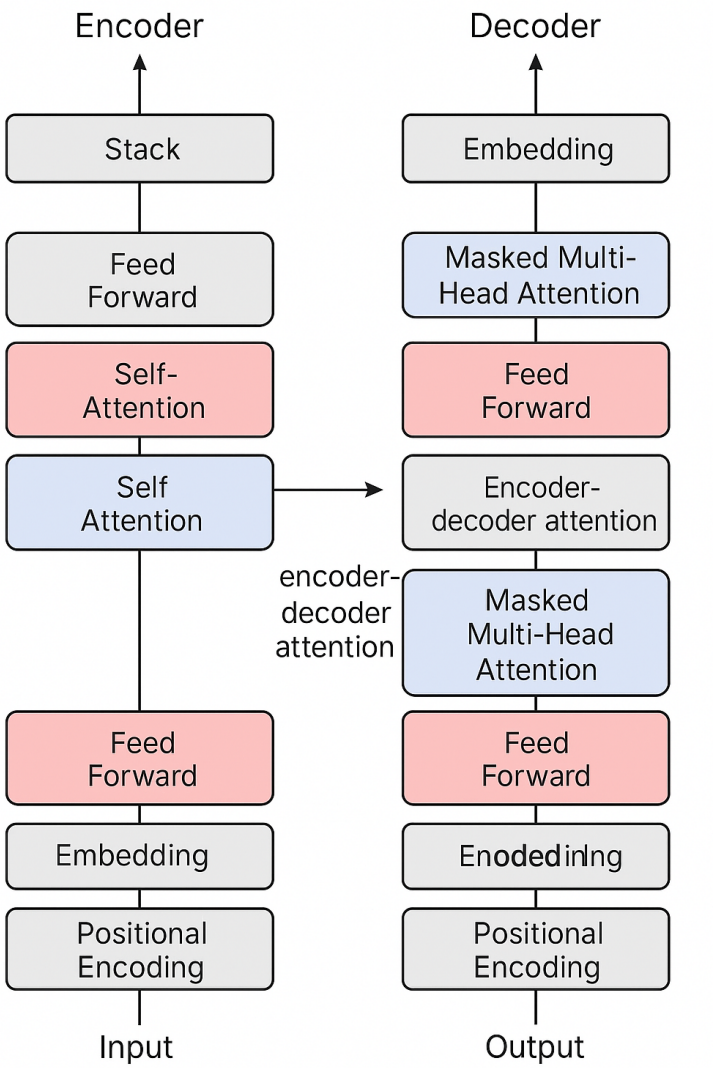

Figure 1: Transformer encoder-decoder architecture showing the flow from input tokens through the encoder stack to the decoder for sequence generation.

.

Self-Attention and Multi-Head Attention

The core innovation of transformers lies in the self-attention mechanism. Given a sequence of input tokens, self-attention computes attention weights between each pair of tokens, allowing the model to capture relationships regardless of their distance. Formally, self-attention maps input queries Q, keys K, and values V into output representations using:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Multi-head attention enhances this mechanism by learning multiple sets of projections in parallel, enabling the model to capture various types of relationships.

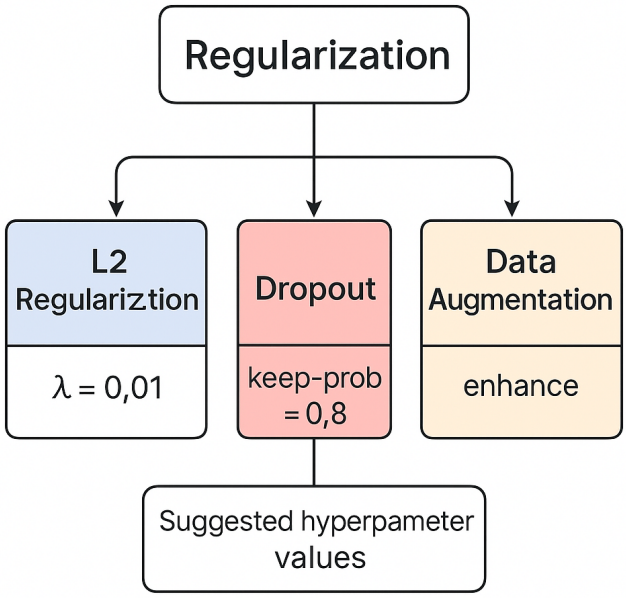

Figure 2: Visualization of the self-attention mechanism in a transformer. Each input token attends to all others using learned Query, Key, and Value projections.

.

Positional Encoding and Sequence Modeling

Since transformers lack any recurrence or convolution, they require a method to encode positional information. This is achieved through positional encodings, which are added to the input embeddings. These encodings can be either sinusoidal (as in the original paper) or learned parameters. They allow the model to differentiate between the positions of tokens, making it suitable for sequence tasks.

Transformer Variants

Several influential transformer variants have adapted the original architecture for specific purposes:

- **BERT** (Bidirectional Encoder Representations from Transformers): A masked language model that uses only the encoder for bidirectional context understanding. Widely used in classification and question answering.

- **GPT** (Generative Pre-trained Transformer): A decoder-only transformer trained with autoregressive objectives, designed for text generation.
- **T5** (Text-to-Text Transfer Transformer): A unified framework where every NLP task is cast as a text-to-text problem, using an encoder-decoder structure.

Limitations in the NLP Context

Despite their success, transformers face several limitations when applied to NLP:

- **Fixed-length context windows**: Standard transformers process input sequences of fixed length, requiring truncation or splitting of long texts. This can lead to loss of contextual information in tasks such as document summarization or long-form QA.
- **High computational cost**: The self-attention mechanism has a quadratic complexity with respect to sequence length, making transformers expensive to train and deploy, especially for long sequences.

These challenges have spurred the development of efficient transformer architectures (e.g., Longformer, Linformer), which aim to extend the applicability of transformers to broader tasks and domains.

3 Transformers in Computer Vision

Motivation

Convolutional Neural Networks (CNNs) have dominated computer vision tasks for over a decade, due to their ability to learn spatial hierarchies via local receptive fields and shared weights. However, they also exhibit limitations: their strong inductive biases (e.g., locality, translation equivariance) restrict their ability to model global relationships, and stacking deep layers is often required to capture long-range dependencies.

Transformers, on the other hand, natively model pairwise interactions across all spatial positions using self-attention. This offers a compelling alternative to CNNs, especially in large-scale settings where learning global patterns and scalability are critical. The introduction of Vision Transformers (ViTs) challenged the long-held assumption that convolution is essential for vision tasks, setting off a wave of research into transformer-based architectures for computer vision.

Key Architectures

Several transformer models have been proposed and optimized for visual tasks. Notable ones include:

- **ViT (Vision Transformer)**: ViT splits an image into non-overlapping patches (e.g., 16x16), flattens them, and treats them as a sequence of tokens — similar to words in NLP. These patch tokens are processed by a standard transformer encoder with positional encodings. ViT achieved competitive performance on ImageNet when pretrained on large datasets like JFT-300M.
- **DeiT (Data-efficient Image Transformer)**: To reduce the data dependency, DeiT introduced training strategies such as knowledge distillation and augmentation that allowed ViTs to perform well using only ImageNet-1K without large-scale pretraining.
- **Swin Transformer**: The Swin (Shifted Window) Transformer addresses the computational inefficiency of global self-attention by introducing a hierarchical structure and local attention within shifted windows. This enables scalability to high-resolution tasks like object detection and segmentation.
- **Others**: Numerous variants have emerged:
  - **CvT (Convolutional Vision Transformer)**: Combines convolutional layers with transformers to inject locality and reduce complexity.
  - **PiT (Pooling-based Vision Transformer)**: Uses pooling to build hierarchical feature representations, akin to CNNs.
  - **NesT (Nested Transformer)**: Adopts a nested attention structure to model both local and global information hierarchically.

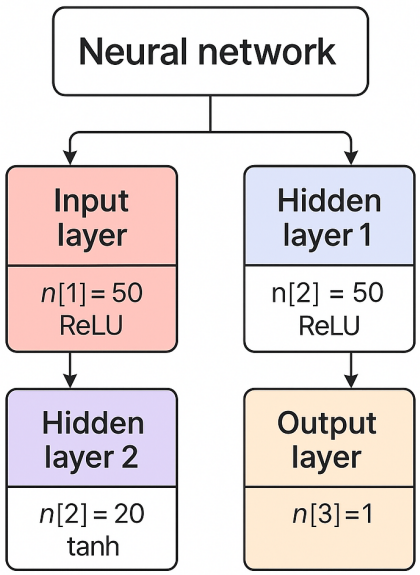

Figure 3: Vision Transformer (ViT): The image is split into fixed-size patches, embedded, and passed through transformer encoders for image classification.

Applications

Transformer-based models have achieved state-of-the-art or competitive performance across a variety of core vision tasks:

- **Image Classification**: ViT and DeiT have shown excellent performance on standard benchmarks like ImageNet, with ViT even surpassing CNNs like ResNet-152 when trained at scale.
- **Object Detection**: DETR (Detection Transformer) introduced an end-to-end object detector by formulating object detection as a direct set prediction problem using transformers.
- **Semantic Segmentation**: Architectures like SETR and Swin-based models have demonstrated strong results on segmentation tasks, leveraging the ability of transformers to aggregate global context.

Challenges

Despite their promise, applying transformers in computer vision comes with specific challenges:

- **Data Hunger**: Transformers lack the strong inductive biases of CNNs and thus require massive labeled datasets to perform well. For example, ViT requires pretraining on large datasets like JFT-300M or ImageNet-21K to match or surpass ResNet performance on ImageNet-1K.
- **Computational Cost**: The quadratic complexity of self-attention with respect to input length makes transformer-based models computationally intensive, especially for high-resolution images.
- **Lack of Built-in Locality**: Unlike CNNs, standard ViTs do not inherently model local spatial structure, potentially leading to less data-efficient learning unless locality is explicitly injected (e.g., via hybrid or hierarchical designs).

4 Transformers in Speech Processing

Motivation

Speech is a highly temporal and sequential modality, requiring models that can capture both short- and long-term dependencies over time. Traditional approaches such as Hidden Markov Models (HMMs) and Recurrent Neural Networks (RNNs) dominated speech processing for years. However, they often struggled with modeling long-range dependencies, training instability, and limited parallelism.

Transformers offer a compelling alternative due to their ability to model global dependencies through self-attention and their inherently parallelizable architecture. Their success in NLP

has motivated researchers to explore similar architectures for speech-related tasks, such as automatic speech recognition (ASR), speaker identification, and speech synthesis.

Key Architectures

Several transformer-based models have emerged specifically for speech processing. These architectures often incorporate innovations like self-supervised learning, convolutional front-ends, and specialized pretraining objectives to handle raw audio or spectrogram inputs.

- **wav2vec and wav2vec 2.0**: These models introduced a self-supervised framework where the transformer learns contextualized audio representations from raw waveforms. wav2vec 2.0 improved upon the original by combining a CNN feature extractor with a transformer encoder, enabling state-of-the-art ASR performance without requiring massive labeled datasets.
- **HuBERT (Hidden-Unit BERT)**: HuBERT extends wav2vec by clustering latent acoustic units and using them as prediction targets. It further improves representation learning by iteratively refining pseudo-labels in a BERT-like masked prediction setup.
- **Conformer**: This model integrates convolutional layers into the transformer architecture, combining the strengths of local feature extraction (via CNNs) and global context modeling (via self-attention). Conformer has been particularly effective for ASR tasks, outperforming vanilla transformers in both accuracy and efficiency.
- **SpeechT5**: Inspired by the text-to-text framework of T5, SpeechT5 adopts a unified encoder-decoder transformer to support a range of speech tasks — including ASR, text-to-speech (TTS), speaker identification, and speech translation — under a single architecture.

Applications

Transformer-based models have shown promising results in a wide spectrum of speech-related applications:

- **Automatic Speech Recognition (ASR)**: Models like wav2vec 2.0 and Conformer are widely used in ASR systems, achieving competitive or state-of-the-art results on benchmarks such as LibriSpeech and CommonVoice.
- **Speaker Identification and Diarization**: Transformer-based encoders can learn speaker embeddings from speech utterances, enabling accurate speaker classification and segmentation in multi-speaker conversations.
- **Emotion Recognition from Voice**: By modeling long-term vocal patterns and prosody, transformers can capture subtle cues in pitch, tone, and cadence to infer emotional states.
- **Speech Synthesis and Enhancement**: Transformer decoders are used in text-to-speech systems (e.g., FastSpeech 2) and speech enhancement tasks like denoising or dereverberation, often outperforming conventional RNN-based approaches.

Challenges

Despite their growing popularity, applying transformers to speech still involves several key challenges:

- **Robustness to Noise and Accents**: Variability in real-world speech — such as background noise, microphone quality, and speaker accents — can degrade model performance. Robust training methods and data augmentation are required to address these issues.
- **Large Model Sizes**: Transformer-based speech models often require significant memory and computational resources, posing difficulties for deployment on mobile or edge devices.
- **Real-Time Processing**: Long sequence modeling introduces latency and hinders real-time applications such as voice assistants or live transcription. Techniques like streaming transformers or chunk-based processing are under active research.

5 Multimodal Transformers

What and Why

Multimodal learning refers to the integration and joint modeling of information from multiple input modalities, such as text, images, audio, and video. Humans process the world in a multimodal way — interpreting speech, facial expressions, text, and visual cues together — and AI systems are now advancing toward this goal.

Transformers, due to their modularity and flexible attention mechanisms, have emerged as a natural architecture for multimodal learning. They offer several key advantages:

- **Richer context understanding**: Combining signals from multiple modalities allows models to develop a deeper, more holistic understanding of input data. For instance, aligning a spoken description with a visual scene leads to better comprehension than either alone.
- **Unified learning framework**: Transformers can process sequences of modality-specific embeddings (text tokens, image patches, audio features) in parallel, enabling parameter sharing and joint representation learning across tasks.

Representative Models

Numerous transformer-based models have demonstrated the power of multimodal learning across a range of tasks:

- **CLIP (Contrastive Language-Image Pretraining)**: CLIP jointly trains a text encoder and an image encoder using contrastive loss on natural language-image pairs. It enables zero-shot classification and semantic retrieval using natural language prompts.
- **DALL·E and DALL·E 2**: These models generate images from natural language descriptions. They utilize transformer-based architectures to model the text-to-image generation process, highlighting transformers' generative capabilities across modalities.
- **BLIP and BLIP-2**: Bootstrapped Language-Image Pretraining models integrate vision and language tasks (e.g., captioning, VQA, retrieval) using vision-language transformers with vision-language pretraining and decoding stages.
- **Flamingo**: Developed by DeepMind, Flamingo uses a frozen visual backbone and a pretrained language model, allowing few-shot learning across a range of vision-language tasks by conditioning on multimodal inputs.
- **GIT, PaLI, Gemini**: These large-scale multimodal models incorporate unified architectures capable of performing multiple tasks — such as VQA, captioning, and document understanding — using massive pretraining on aligned multimodal datasets.

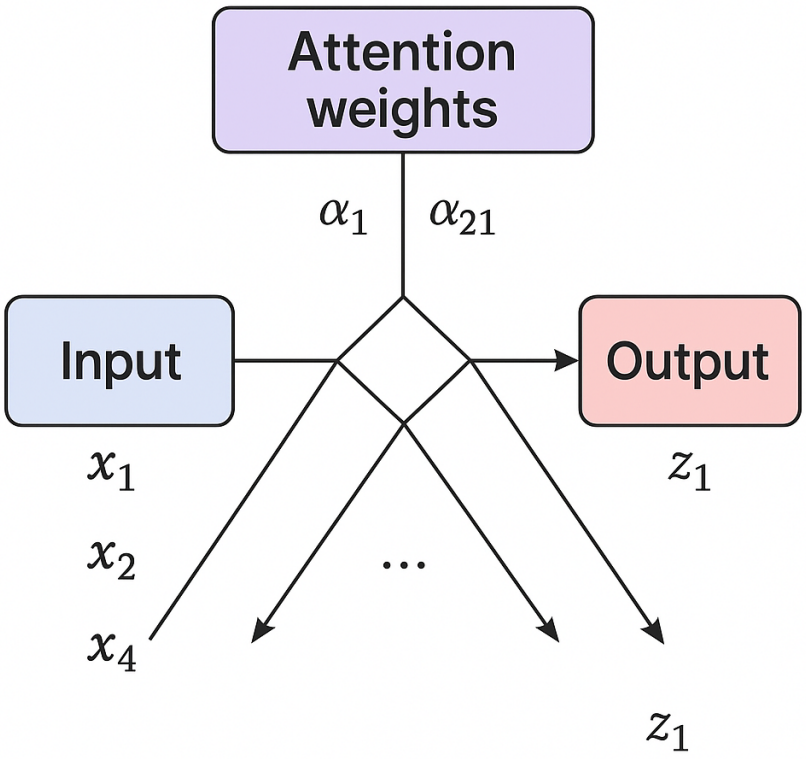

Figure 4: Example of a multimodal transformer (e.g., CLIP) that aligns image and text inputs in a shared semantic embedding space using contrastive learning.

.

Applications

Multimodal transformers have advanced the state of the art across several applications:

- **Image Captioning**: Generating descriptive textual captions for images by jointly modeling visual features and language semantics.
- **Visual Question Answering (VQA)**: Answering natural language questions based on visual inputs, requiring reasoning over both modalities.
- **Text-to-Image/Video Generation**: Creating realistic images or videos from natural language prompts, as demonstrated by DALL·E, Imagen, and others.
- **Cross-modal Retrieval and Alignment**: Matching or retrieving items across modalities — e.g., finding relevant images for a query sentence, or aligning subtitles with video frames.

Challenges

Despite rapid progress, several challenges persist in building robust and generalizable multimodal transformers:

- **Data Imbalance Across Modalities**: Text datasets are abundant, while high-quality labeled image-audio-video data is scarcer and more expensive to collect. This leads to unbalanced learning dynamics during training.
- **Fusion Techniques**: Deciding when and how to fuse information from different modalities remains an open problem. Early fusion (input-level) and late fusion (output-level) each have trade-offs in terms of flexibility and performance.
- **Alignment of Embeddings**: Ensuring that representations from different modalities lie in a shared, semantically meaningful embedding space is challenging, especially when modalities differ in structure and granularity.
- **Evaluation Metrics**: Measuring multimodal performance is difficult due to subjective or task-specific goals. For example, in text-to-image generation, assessing visual quality and semantic relevance requires both automatic metrics and human judgment.

6 Unified and General-Purpose Architectures

As transformer architectures mature and scale, the goal of building general-purpose models capable of performing diverse tasks across multiple modalities is becoming increasingly tangible. Several architectures have emerged that aim to unify learning across different

input types (text, image, audio, video, etc.) and task formats (classification, generation, reasoning), pushing the boundaries of artificial general intelligence (AGI).

Perceiver and Perceiver IO

The **Perceiver** architecture, introduced by DeepMind, proposes a modality-agnostic framework capable of handling high-dimensional and heterogeneous inputs such as images, audio, and point clouds. Instead of applying attention directly to the full input — which is computationally expensive — Perceiver uses a latent bottleneck to attend to inputs in a scalable manner.

Perceiver IO, an extension of the original model, adds flexibility by allowing the model to output arbitrary structured data, making it suitable for a wider range of tasks. Its input-output design allows it to generalize across modalities and tasks without modifying the core architecture. This makes it particularly powerful for use cases like video classification, audio analysis, and multi-sensor fusion.

Gato: A Generalist Agent

Gato, also from DeepMind, represents one of the first attempts at creating a single transformer-based agent that can handle a diverse array of tasks — from image captioning to robotic control — using the same set of model weights. Gato is trained using a unified sequence-to-sequence framework where all tasks are framed as token prediction problems, regardless of modality or task type.

Despite its simplicity, Gato performs competitively across a wide range of benchmarks and device types, illustrating that a single model can acquire the capacity to generalize in multi-task and multi-modal settings with proper data formatting and training.

Gemini: A Multimodal Frontier

Google's latest **Gemini** family of models integrates the strengths of large language models (LLMs) like PaLM with state-of-the-art vision and audio processing capabilities. Designed as truly multimodal from the ground up, Gemini can accept and reason over text, images, videos, and audio streams.

Unlike models that treat multimodality as an extension of a text-only foundation, Gemini is optimized for joint modality alignment, memory, and interaction, making it better suited for real-world applications like intelligent assistants, creative tools, and robotics. Gemini also incorporates reinforcement learning and memory-based planning, positioning it as a step toward agents capable of decision-making and interaction in complex environments.

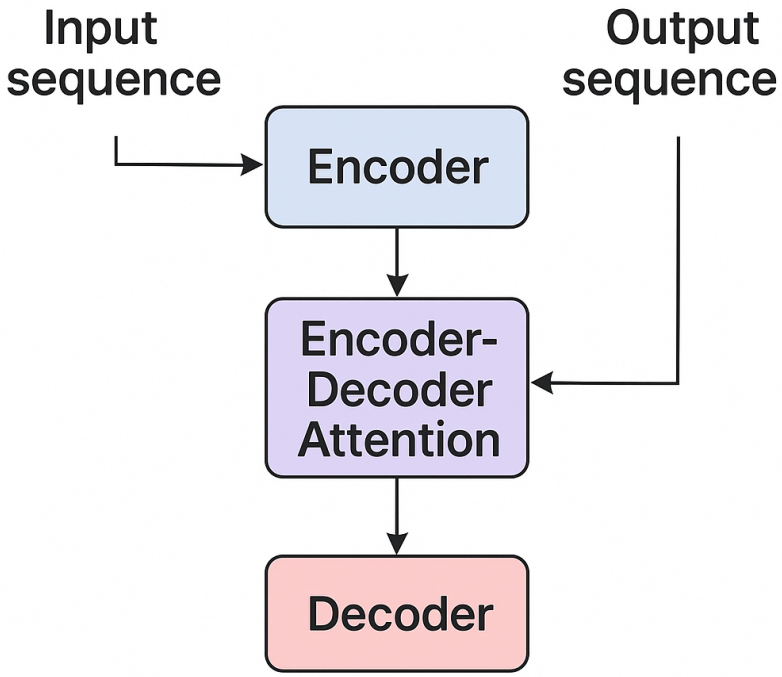

Figure 5: Unified transformer architecture (e.g., Perceiver IO or Gato) capable of handling multiple modalities and tasks under a shared model pipeline.

.

Implications for General Intelligence and Scalability

These unified architectures signal a paradigm shift from specialized, task-specific models toward more general-purpose AI systems. The implications are profound:

- **Scalability**: Modular architectures like Perceiver scale linearly with input size, enabling practical deployment in domains with large or high-dimensional inputs (e.g., genomics, 3D vision).
- **Cross-task Transfer**: Models like Gato and Gemini demonstrate that shared representations across modalities and tasks can enable transfer learning and multitask generalization, reducing the need for task-specific fine-tuning.
- **Toward AGI**: By unifying perception, language, and action under a single architecture, these models represent meaningful progress toward the long-term goal of artificial general intelligence (AGI).

- **Real-World Readiness**: With their ability to handle multimodal inputs and produce structured outputs, these architectures are well-positioned for real-world deployment in domains such as robotics, healthcare, education, and digital assistants.

7 Real-World Applications

The rapid advancement of transformer-based models beyond NLP has opened up a diverse range of real-world applications across industries. The ability of these models to integrate and reason over multiple modalities — such as text, images, audio, and sensor data — has made them foundational components in intelligent systems. Below are several domains where cross-modal and general-purpose transformers are driving innovation.

Google Gemini: Multimodal AI Assistant

Gemini, developed by Google DeepMind, exemplifies the integration of multimodal understanding in next-generation AI assistants. It is designed to process text, images, video, and audio in a unified way, enabling capabilities such as:

- Analyzing diagrams or screenshots while conversing about them in natural language.
- Understanding multimodal queries like "What is happening in this video clip?".
- Summarizing PDFs, reading handwriting, or answering questions about graphs and plots.

This positions Gemini not just as a chatbot, but as a powerful productivity and reasoning tool with potential applications in education, research, software engineering, and business workflows.

Autonomous Vehicles and Sensor Fusion

In self-driving vehicles, understanding the environment requires fusing information from heterogeneous sources such as:

- Visual cameras (RGB images).
- LiDAR sensors (3D point clouds).
- GPS and IMU data.
- Semantic maps and natural language instructions.

Multimodal transformer architectures enable the unified modeling of sensor data, resulting in improved perception, prediction, and planning. For instance, Perceiver IO and BEVFormer architectures have been explored for bird's-eye-view object detection using multi-view fusion.

Medical Imaging and Diagnostics

Transformers like **ViT** and **Swin Transformer** are being adopted in medical imaging tasks due to their ability to model global context and long-range dependencies. Applications include:

- Tumor classification in histopathology slides.
- Brain lesion segmentation in MRI.
- Retinal disease detection from fundus images.

Transformer models are also being used to fuse imaging with textual data (e.g., radiology reports), improving diagnostic accuracy and interpretability in clinical decision support systems.

AR/VR and Embodied AI Assistants

Multimodal transformers are powering intelligent agents in augmented and virtual reality (AR/VR) environments. These systems can:

- Understand voice commands, gestures, and eye gaze.
- Generate contextual responses or visual augmentations.
- Navigate and interact with 3D environments in real time.

Such capabilities are enabled by architectures that can process audio, vision, and motion data in a synchronized fashion, critical for virtual tutors, industrial training tools, and immersive gaming assistants.

8 Challenges and Open Problems

Despite the impressive progress of transformer architectures across modalities, several fundamental challenges remain unresolved. These limitations span computational efficiency, interpretability, ethical concerns, and model robustness, and they present key areas for future research.

Long Sequence Inefficiency in Vision and Speech

Transformers scale quadratically with input sequence length due to the self-attention mechanism, making them computationally expensive for high-resolution images or long audio clips. In computer vision, processing large images as sequences of small patches (e.g., ViT) can result in excessive memory use. In speech, long utterances or streaming input add latency and complexity. Emerging approaches like **Sparse Attention**, **Performer**, and **Longformer** seek to address these inefficiencies, but trade-offs remain between performance and efficiency.

Model Interpretability and Explainability

As transformers become larger and more complex, understanding their decision-making processes becomes increasingly difficult. This opacity is problematic in high-stakes domains like healthcare, law, and autonomous systems. Unlike CNNs or decision trees, attention maps do not always provide reliable explanations. There is an ongoing need for techniques that can offer human-interpretable rationales for predictions across modalities.

Dataset Bias and Fairness in Multimodal Models

Multimodal transformers are often trained on large-scale web data (e.g., CLIP's 400M image-text pairs), which may reflect societal biases, stereotypes, or culturally skewed distributions. These biases can propagate or even amplify in downstream applications. Moreover, the imbalanced availability of data across modalities (e.g., more English text than non-English, more images than audio) introduces challenges in equitable performance and generalization.

Compute and Energy Efficiency

Training large-scale multimodal transformers like Gemini or Flamingo requires vast compute resources, carbon emissions, and financial cost, raising concerns about sustainability and accessibility. Furthermore, deploying such models on mobile or edge devices remains challenging due to memory and latency constraints. Research into model compression, quantization, knowledge distillation, and efficient transformer variants is critical for enabling broader use.

Training Instability and Catastrophic Forgetting

Multimodal and multitask training settings often lead to instability, where the model struggles to converge or overfits to one dominant modality. Additionally, when fine-tuned on new modalities or tasks, transformers can exhibit **catastrophic forgetting** — losing performance on previously learned capabilities. Strategies such as continual learning, rehearsal methods, and parameter-efficient adaptation (e.g., LoRA, adapters) are active areas of investigation.

9 Future Directions

As transformer models continue to push the boundaries of AI across modalities, several promising research directions are emerging. These aim to improve efficiency, adaptability, generalization, and real-world usability. Below, we highlight key areas that are expected to shape the next generation of cross-modal transformer architectures.

Efficient Transformers

To address the computational limitations of standard self-attention, many researchers are developing efficient transformer variants. Models such as:

- **Linformer** approximates full attention using low-rank projections.
- **Performer** employs kernel-based approximations to reduce attention complexity to linear time.
- **Longformer** introduces sparse attention patterns suited for long documents or sequences.

These architectures enable transformers to scale to longer sequences in speech and vision while reducing memory and energy costs — a prerequisite for real-time or embedded systems.

Continual and Lifelong Learning Transformers

Traditional transformer training assumes fixed tasks and static datasets, but real-world environments are dynamic. Future models must:

- Learn incrementally from new data.
- Avoid catastrophic forgetting.
- Adapt to new domains or modalities without full retraining.

Continual learning strategies, such as replay-based learning, modular architectures, and regularization-based methods, are being actively explored to make transformers more lifelong learners.

Federated and On-Device Transformer Deployment

Deploying large multimodal models on-device (e.g., smartphones, wearables, autonomous robots) enables privacy, personalization, and reduced latency. However, this requires:

- Model compression (quantization, pruning).
- Federated learning techniques that update models locally.
- Lightweight attention mechanisms and reduced parameterization.

On-device transformers are particularly relevant for medical, consumer, and IoT applications, where data sensitivity and response time are critical.

Multilingual and Multimodal Models

The future of AI lies in building models that generalize across both languages and modalities. Existing large language models often struggle with low-resource languages or multilingual understanding. When combined with vision or speech inputs, this complexity increases further.

Progress in this direction includes models such as:

- **mBERT** and **XLM-R** for multilingual text.
- **PaLI** (Pathways Language and Image) for multilingual vision-language tasks.

Unified multilingual-multimodal models would empower global applications, including education, accessibility, and cross-cultural communication.

Combining Transformers with Reinforcement Learning

Transformers are being increasingly combined with reinforcement learning (RL) for complex reasoning and decision-making. Examples include:

- **Decision Transformer**: Treats RL as a sequence modeling problem using trajectory data.
- **Gato**: Combines supervised pretraining with policy-like behavior across environments.

Integrating transformers into RL pipelines opens doors for interactive, memory-aware agents capable of acting in open-ended, multimodal environments — a step toward general-purpose intelligent systems.

10 Conclusion

Since their introduction in 2017, transformer architectures have significantly transformed the landscape of artificial intelligence, beginning with their success in natural language processing. Over the past few years, the core design principles of transformers—self-attention, parallelism, and scalability—have been effectively extended beyond text to tackle challenges in computer vision, speech processing, and multimodal integration.

This survey explored the growing ecosystem of transformer-based models across diverse domains. From ViT revolutionizing image classification, to wav2vec 2.0 advancing self-supervised learning in speech, to CLIP and Gemini setting new standards in multimodal reasoning, transformers have emerged as a unifying architecture for perception and cognition.

A key trend highlighted throughout the survey is the convergence of modalities into a single, shared modeling space. Unified architectures like Perceiver IO, Gato, and Gemini exemplify

this shift toward general-purpose AI systems capable of handling a broad array of tasks and input formats. These advances not only simplify model design and deployment but also bring us closer to systems that exhibit properties of artificial general intelligence (AGI) — such as flexible problem-solving, real-world grounding, and cross-domain transfer.

Nevertheless, substantial challenges remain. Computational efficiency, interpretability, fairness, and training stability are active areas of research that must be addressed to unlock the full potential of transformers across domains and devices.

Looking forward, continued progress in efficient modeling, continual learning, federated deployment, and reinforcement learning integration will play a crucial role in shaping the next wave of general-purpose AI. As the line between individual modalities continues to blur, transformers are well-positioned to serve as the foundation for the next generation of intelligent systems — adaptable, scalable, and universal.

11 References

1. Vaswani, Ashish et al. “Attention Is All You Need.” Advances in Neural Information Processing Systems, 2017.

2. Devlin, Jacob et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, 2018.

3. Radford, Alec et al. “Improving Language Understanding by Generative Pre-Training.” OpenAI, 2018.

4. Raffel, Colin et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” Journal of Machine Learning Research, 2020.

5. Dosovitskiy, Alexey et al. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv, 2020.

6. Baevski, Alexei et al. “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.” Advances in Neural Information Processing Systems, 2020.

7. Radford, Alec et al. “Learning Transferable Visual Models From Natural Language Supervision.” OpenAI, 2021.

8. Alayrac, Jean-Baptiste et al. “Flamingo: A Visual Language Model for Few-Shot Learning.” DeepMind, 2022.

9. DeepLearning.AI. Machine Learning and Deep Learning Specializations by Andrew Ng.

10. Stanford University. CS230: Deep Learning, by Andrew Ng and Kian Katanforoosh.

11. Wikipedia Contributors. Articles on Transformers, Self-Attention, and Multimodal AI.

12. Medium and Towards Data Science Authors. Articles on Deep Learning, Optimization, and Transformer Applications.