zilliz

WEBINAR

Foundation Models are Going Multimodal

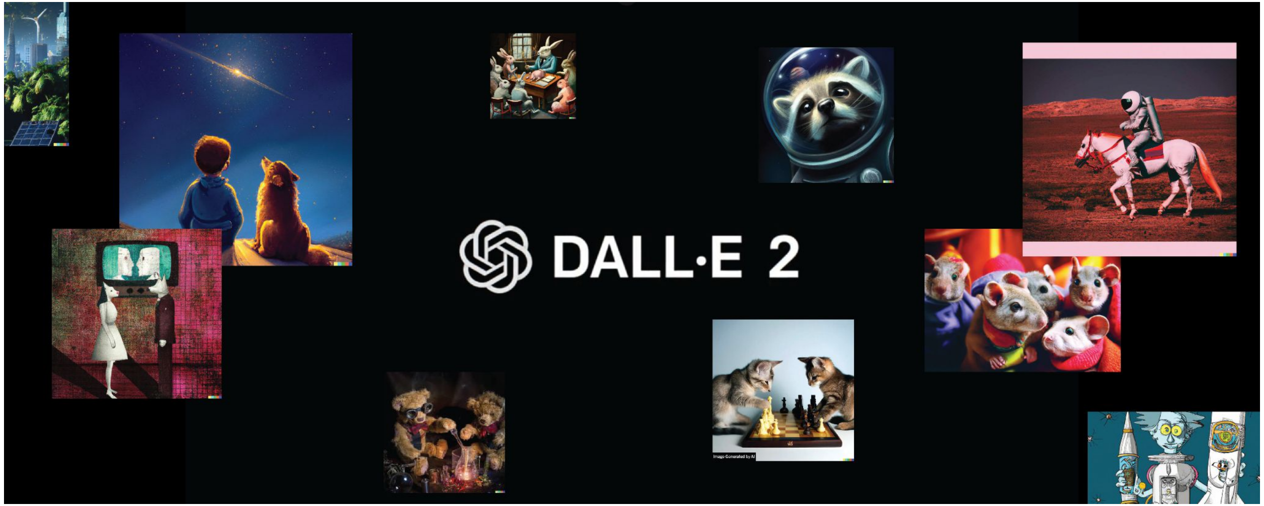
James Le

Developer Experience, Twelve Labs

July 27, 2023 | 9:00 AM PT

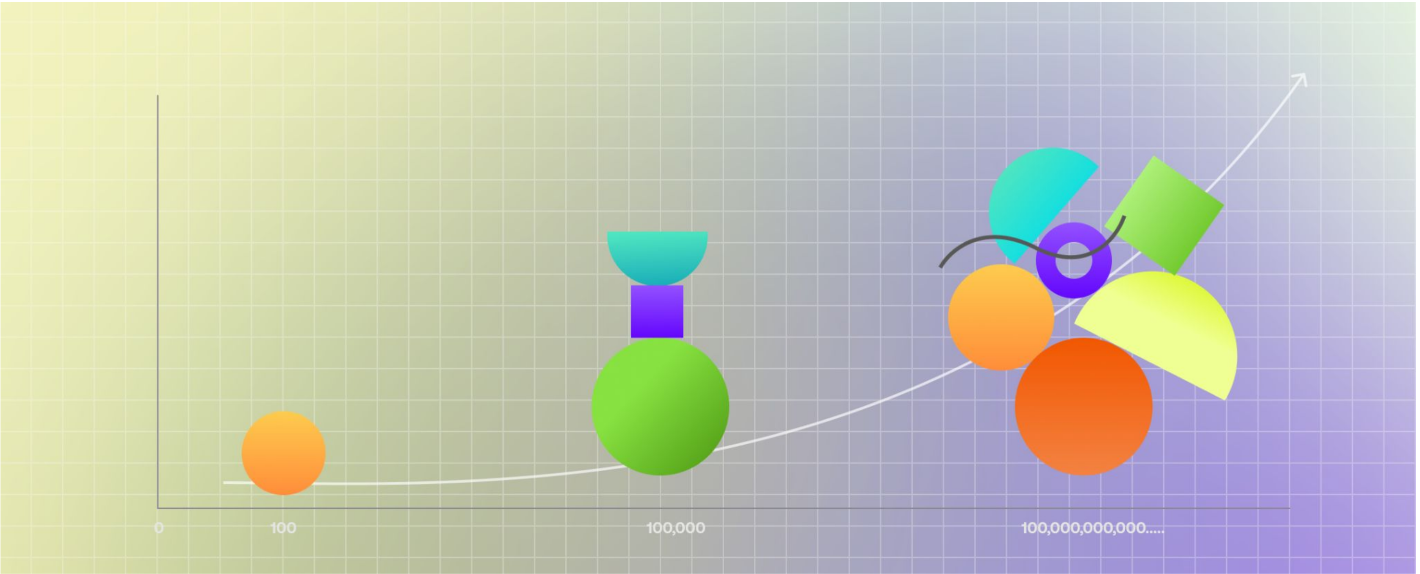

GPT-4


GitHub Copilot


DALL·E 2

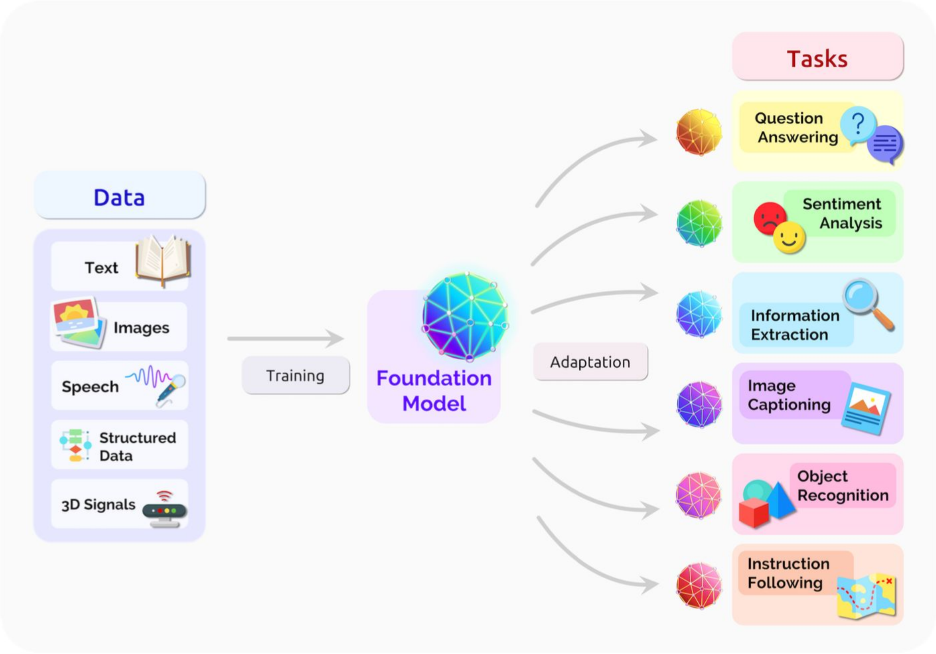
Agenda

1. A Gentle Introduction to Foundation Models
2. The Birth of Large Language Models
3. The Rise of Large Vision-Language Models
4. The New Paradigm of Video Foundation Models



A Gentle Introduction to Foundation Models

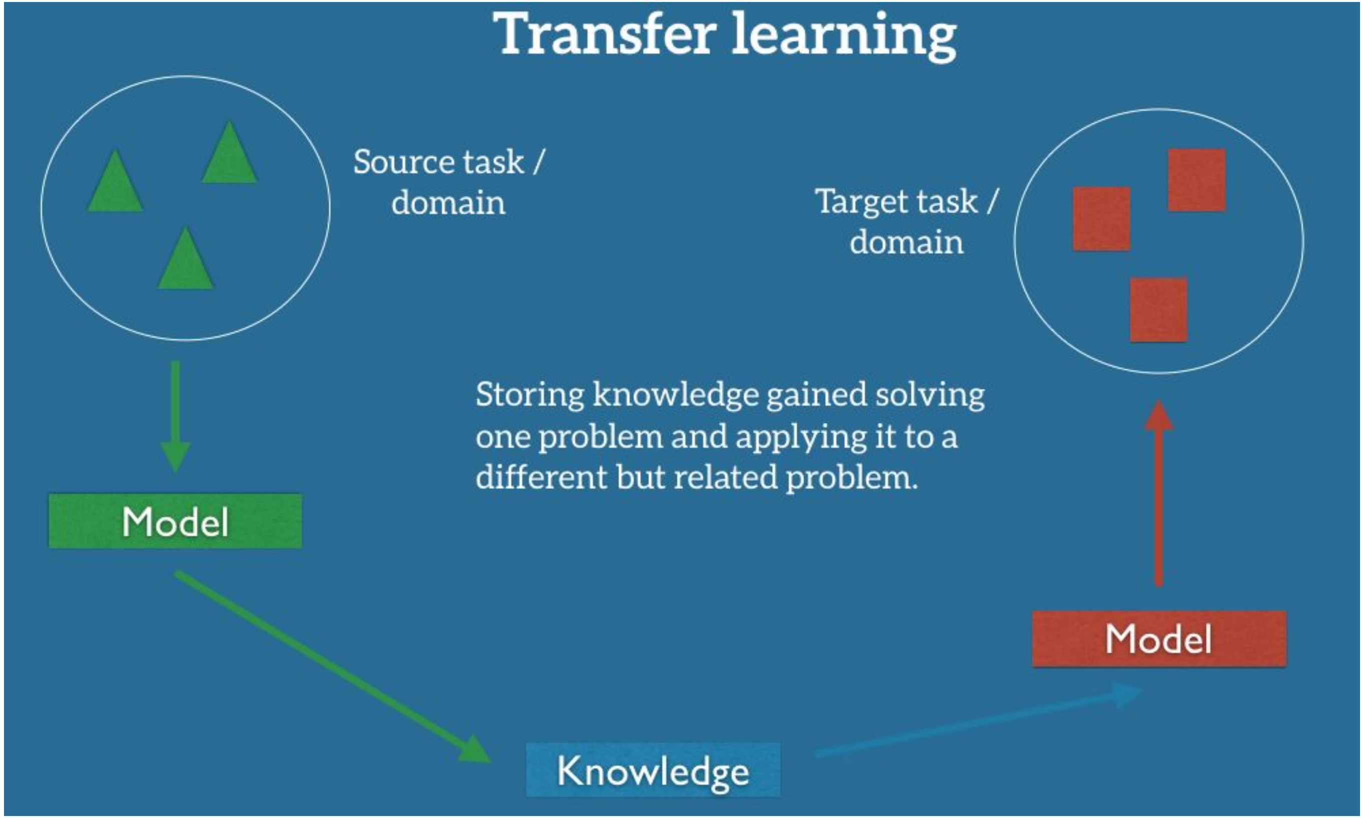
What Is A Foundation Model?



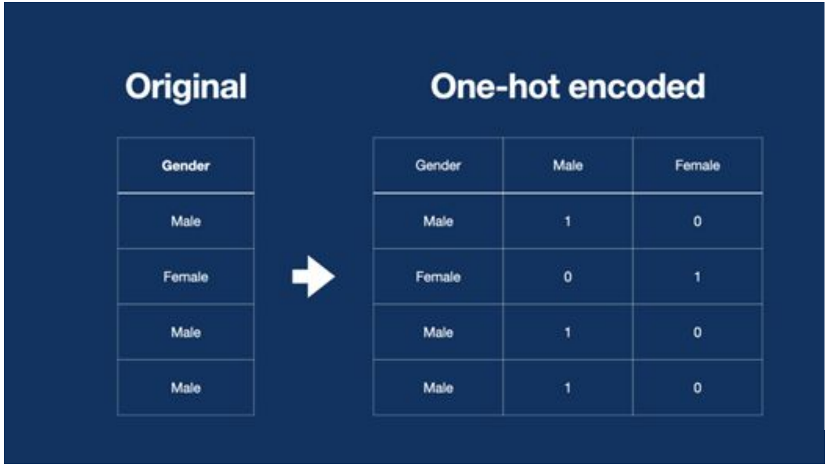
- Learns from a wide range of data using self-supervision at scale
- Leverages **deep neural networks** and **self-supervised learning**
- Useful for various downstream tasks

Source: On the Opportunities and Risks of Foundation Models (Stanford HAI, 2017)

Transfer Learning



Word Embeddings

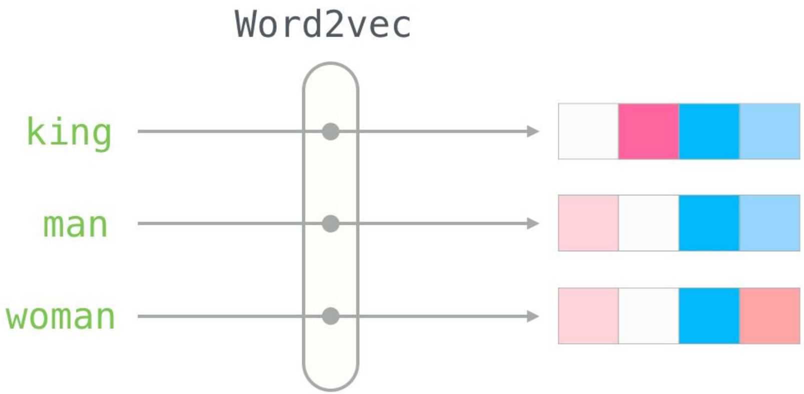


Original

Gender
Male
Female
Male
Male

One-hot encoded

Gender	Male	Female
Male	1	0
Female	0	1
Male	1	0
Male	1	0



- **One-hot encoding** encodes words as vectors (embeddings)
- **Word2vec** maximizes cosine similarity between word embeddings
- Models like ELMo, ULMFiT, and GPT employed pre-trained language models to achieve SOTA results on downstream NLP tasks

Transformers

- Analyze tokens simultaneously
- **Attention mechanism** to support parallelization
- More computationally efficient than Recurrent Neural Nets

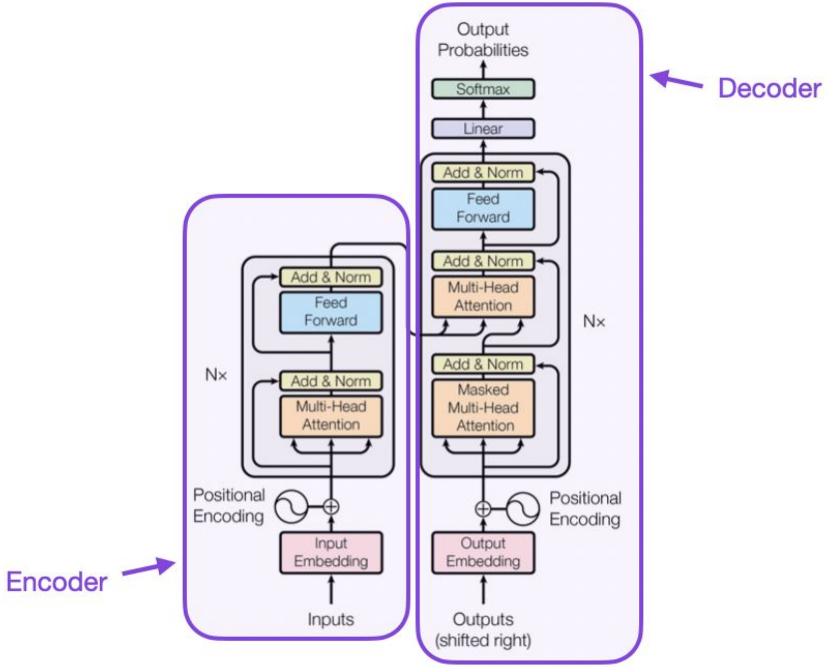
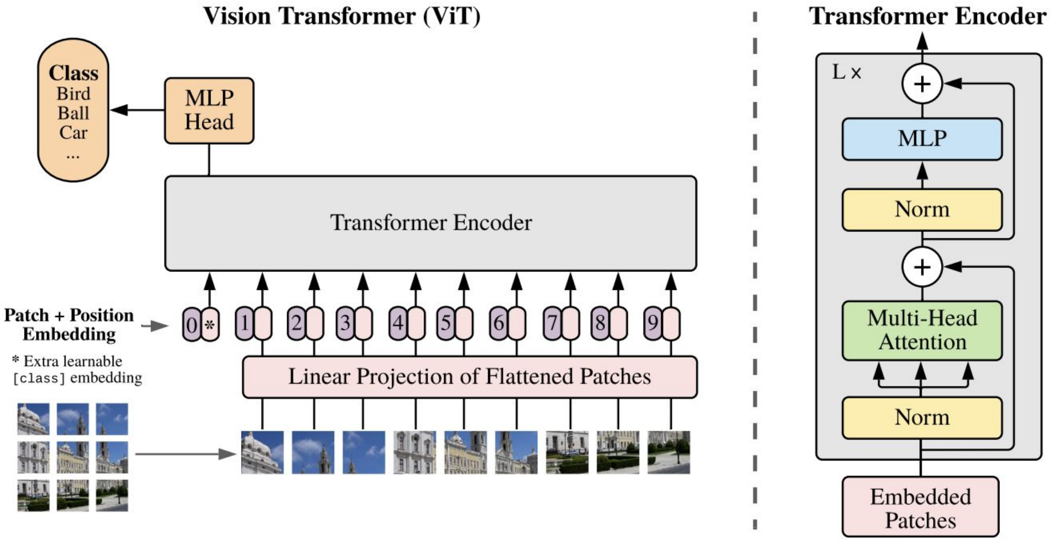


Figure 1: The Transformer - model architecture.

Source: Attention Is All You Need (Vaswani et. al, 2017)

Vision Transformers

- Split an image into patches
- **Treat image patches as token inputs**
- Exceed SOTA results on image classification tasks
- **Require a lot of compute power**
- Not useful for tasks that involve visual elements of varying size

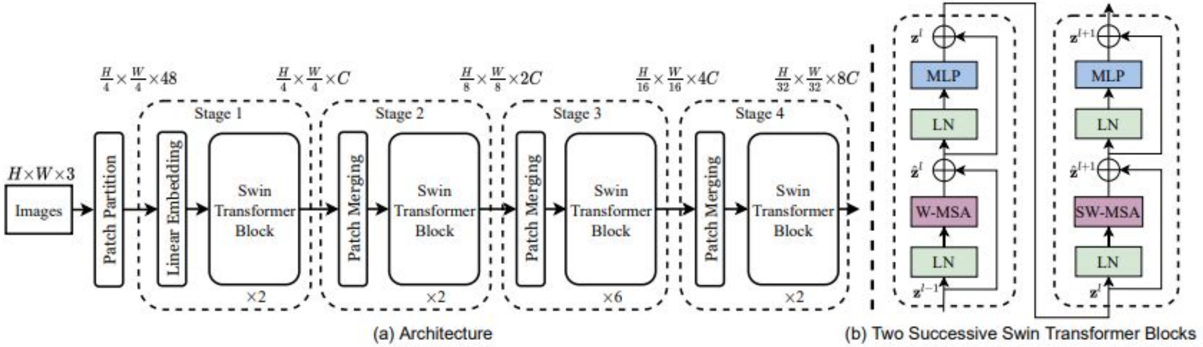


Source: [An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale](#) (Dosovitskiy et. al, 2021)

Transformer Variants

Swin Transformers

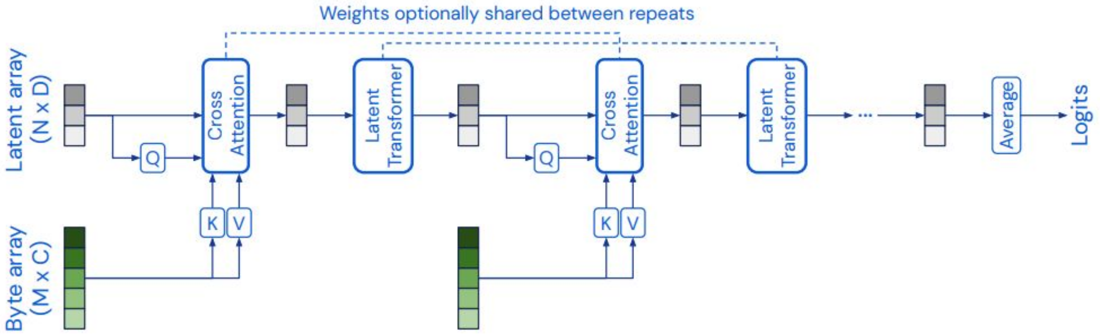
- Hierarchical feature maps
- Shifted window attention



Source: [Swin Transformer: Hierarchical Vision Transformer using Shifted Windows]() (Liu et. al, 2021)

Perceiver

- Latent units to form an attention bottleneck
- Associate position- and modality-specific features with each input

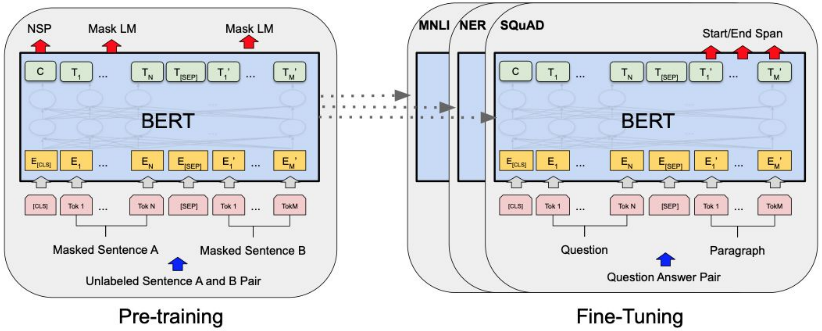


Source: [Perceiver: General Perception with Iterative Attention]() (Jaegle et. al, 2021)

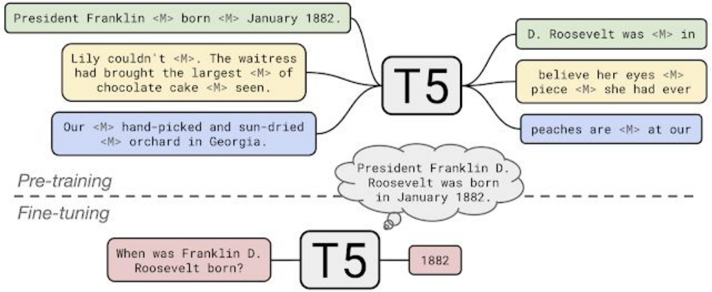
The Birth of Large Language Models

Large Language Models

Source: [Language Models are Unsupervised Multitask Learners](#) (Radford et. al, 2018)

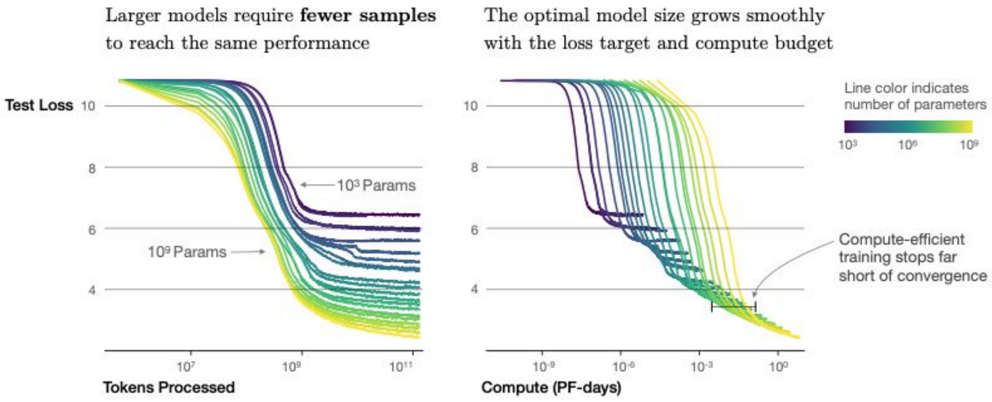


Source: [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#) (Devlin et. al, 2018)



Source: [Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer](#) (Raffel et. al, 2020)

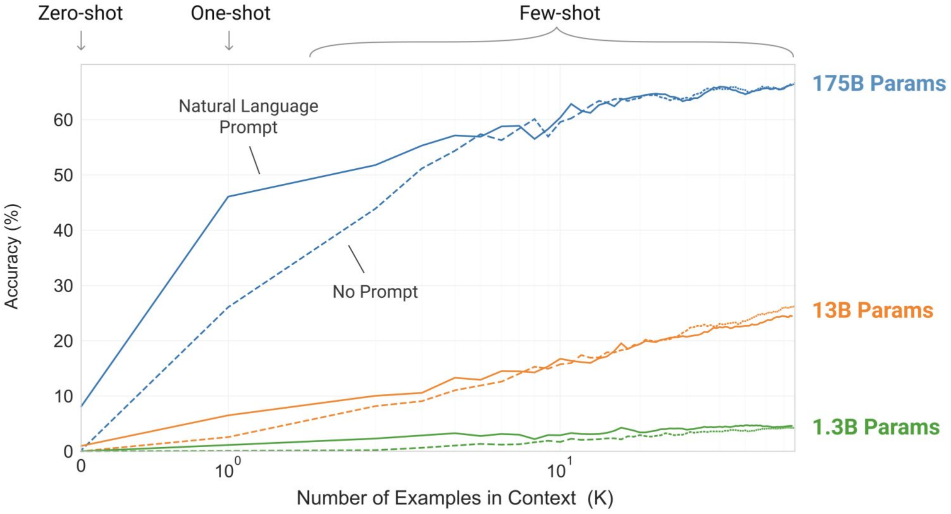
Scaling Laws (by OpenAI): Performance = Data Size x Parameter Size x Compute Size



Source: [Scaling Laws for Neural Language Models](#) (Kaplan et. al, 2020)

- Test loss follows a power law w.r.t **model size**, **dataset size**, and **compute used for training**
- Other architectural details have minimal effects
- Larger models are significantly more sample-efficient

Emergent Abilities of LLMs



Source: Language Models Are Few-Shot Learners (Brown et. al, 2020)

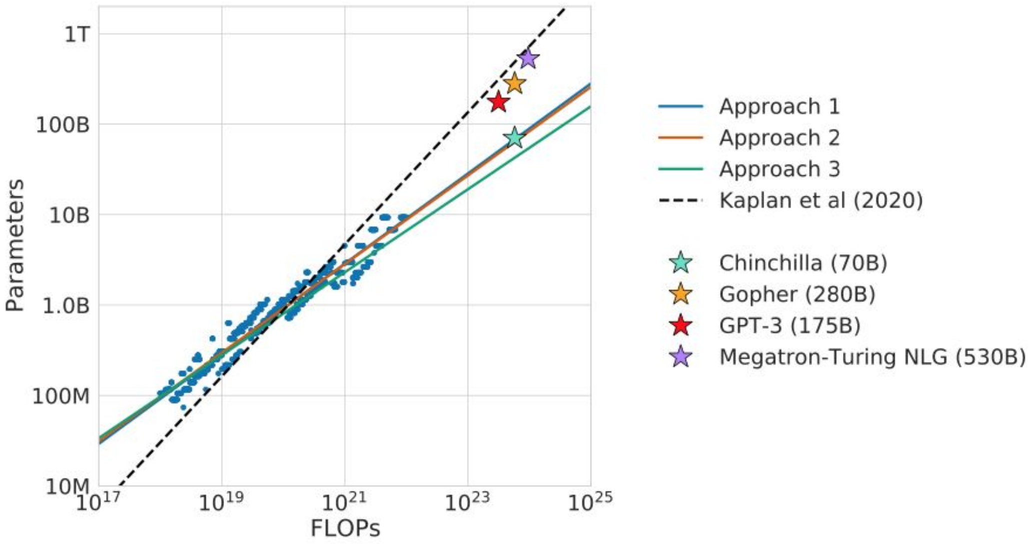
Table 1: List of emergent abilities of large language models and the scale (both training FLOPs and number of model parameters) at which the abilities emerge.

	Emergent scale			
	Train. FLOPs	Params.	Model	Reference
Few-shot prompting abilities				
• Addition/subtraction (3 digit)	2.3E+22	13B	GPT-3	Brown et al. (2020)
• Addition/subtraction (4-5 digit)	3.1E+23	175B		
• MMLU Benchmark (57 topic avg.)	3.1E+23	175B	GPT-3	Hendrycks et al. (2021a)
• Toxicity classification (CivilComments)	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Truthfulness (Truthful QA)	5.0E+23	280B		
• MMLU Benchmark (26 topics)	5.0E+23	280B		
• Grounded conceptual mappings	3.1E+23	175B	GPT-3	Patel & Pavlick (2022)
• MMLU Benchmark (30 topics)	5.0E+23	70B	Chinchilla	Hoffmann et al. (2022)
• Word in Context (WiC) benchmark	2.5E+24	540B	PaLM	Chowdhery et al. (2022)
• Many BIG-Bench tasks (see Appendix E)	Many	Many	Many	BIG-Bench (2022)
Augmented prompting abilities				
• Instruction following (finetuning)	1.3E+23	68B	FLAN	Wei et al. (2022a)
• Scratchpad: 8-digit addition (finetuning)	8.9E+19	40M	LaMDA	Nye et al. (2021)
• Using open-book knowledge for fact checking	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Chain-of-thought: Math word problems	1.3E+23	68B	LaMDA	Wei et al. (2022b)
• Chain-of-thought: StrategyQA	2.9E+23	62B	PaLM	Chowdhery et al. (2022)
• Differentiable search index	3.3E+22	11B	T5	Tay et al. (2022b)
• Self-consistency decoding	1.3E+23	68B	LaMDA	Wang et al. (2022b)
• Leveraging explanations in prompting	5.0E+23	280B	Gopher	Lampinen et al. (2022)
• Least-to-most prompting	3.1E+23	175B	GPT-3	Zhou et al. (2022)
• Zero-shot chain-of-thought reasoning	3.1E+23	175B	GPT-3	Kojima et al. (2022)
• Calibration via P(True)	2.6E+23	52B	Anthropic	Kadavath et al. (2022)
• Multilingual chain-of-thought reasoning	2.9E+23	62B	PaLM	Shi et al. (2022)
• Ask me anything prompting	1.4E+22	6B	EleutherAI	Arora et al. (2022)

Source: Emergent Abilities of Large Language Models (Wei et. al, 2022)

- The more examples you give the model, the better its performance will be. And the larger the model, the better its performance gets.
- The model behavior surges unpredictably from random performance to above random at a specific scale threshold.

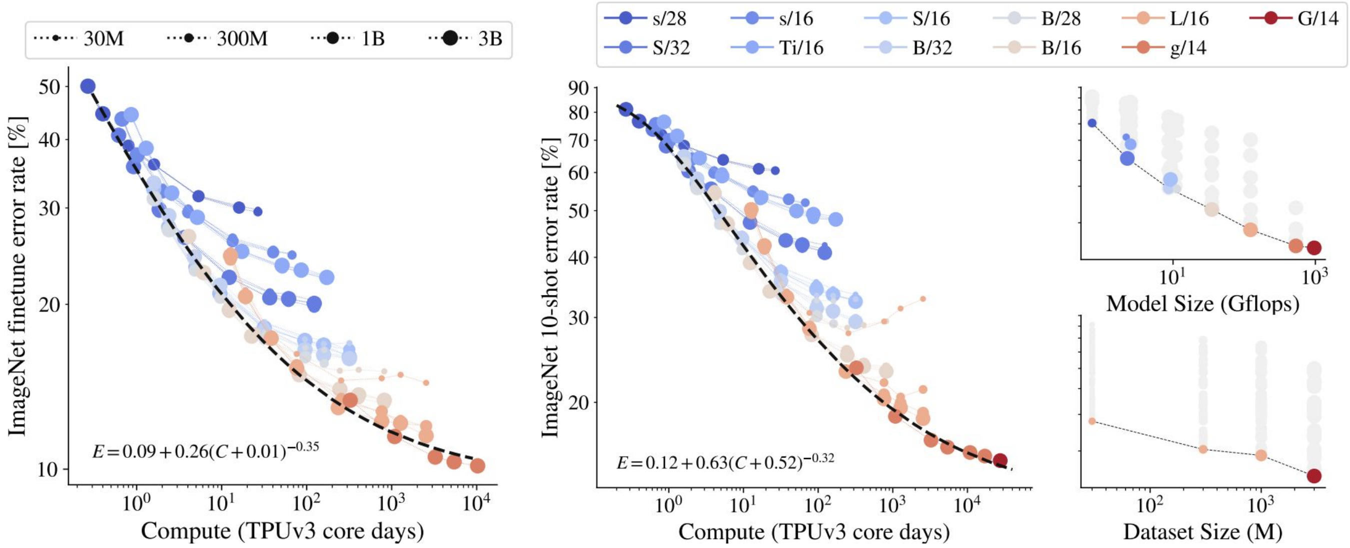
"Chinchilla" Scaling Laws (by DeepMind)



Source: Training Compute-Optimal Large Language Models (Hoffman et. al, 2022)

- More accurate than the OpenAI's original one
- Trainever 400 LLMs with a range of parameters (70M-16B) on a range of tokens (5B-500B)
- Most LLMs are **under-trained** - they haven't seen enough data
- Chinchilla exceeded Gopher
- More LLMs showed up by scaling model size and training on larger datasets from diverse sources

Scaling Laws for Vision

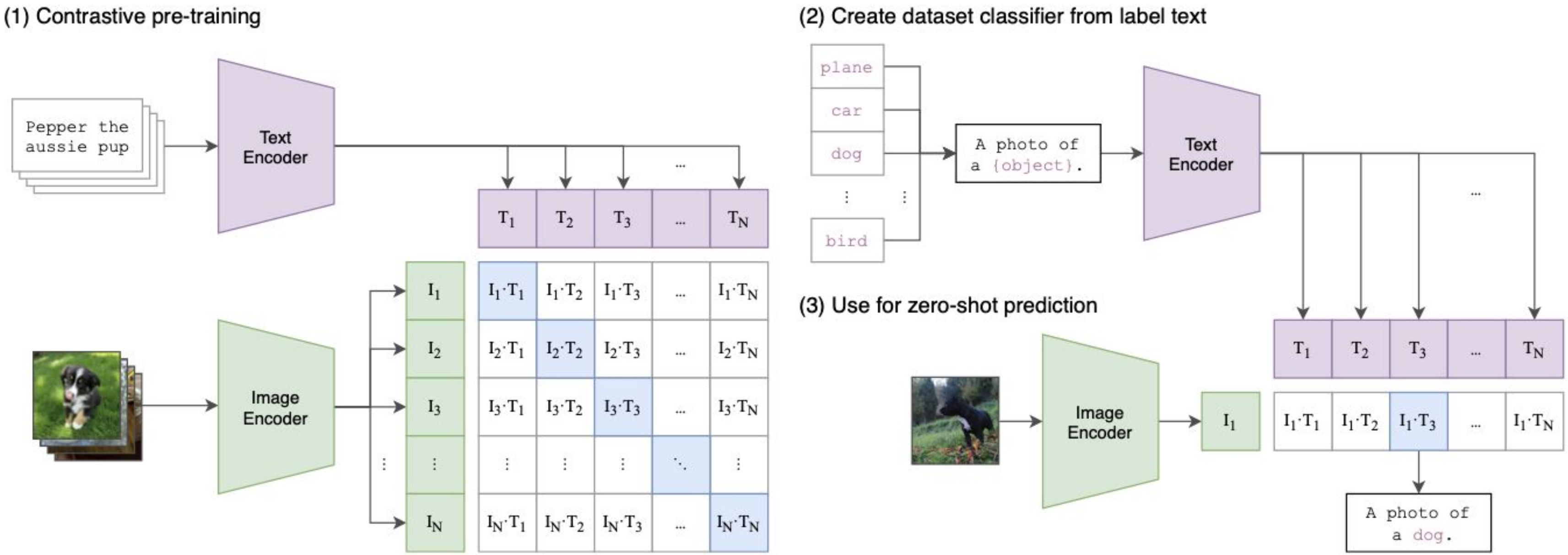


- Experiments with Vision Transformers on a range of parameters (5M-2B), a range of datasets (1M-3B), and a range of compute budgets (1-10,000 TPUs)
- Simultaneously scaling total compute and model size is effective
- Larger models perform better in few-shot learning

Source: Scaling Vision Transformers (Zhai et. al, 2022)

The Rise of Large Vision-Language Models

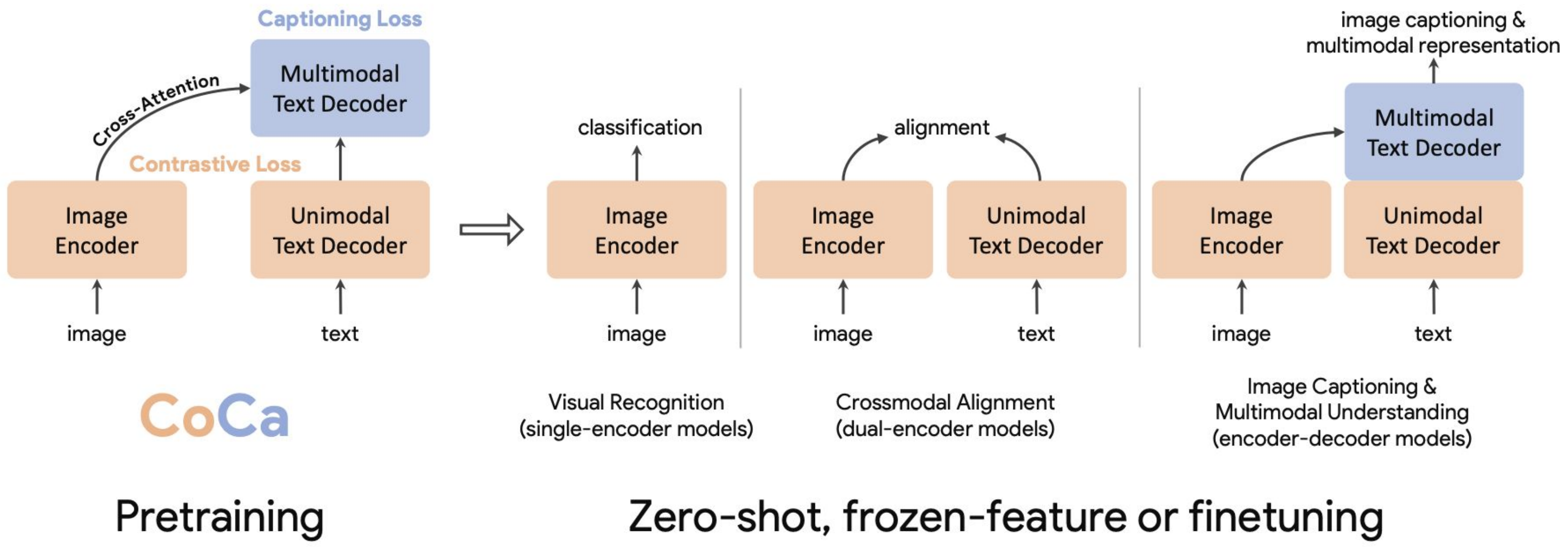
OpenAI’s CLIP



- Contrastive Learning matches correct image and text pairs.
- Map images and text using embeddings: (1) linear probe or (2) “zero-shot” learning

Source: [Learning Transferable Visual Models From Natural Language Supervision]() (Radford et. al, 2021)

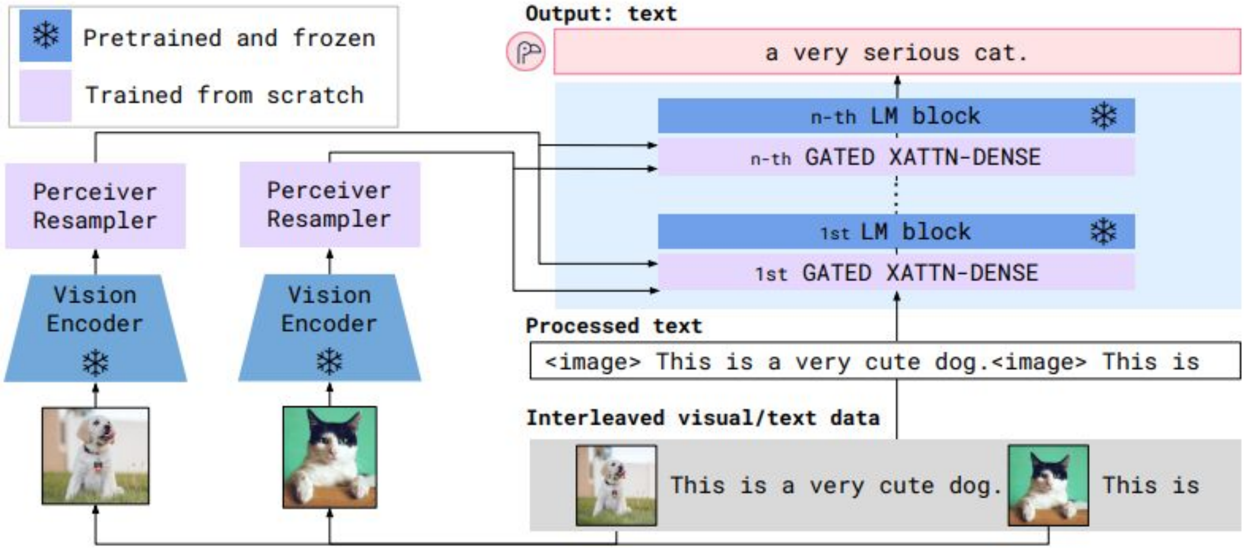
Google's CoCa



- Combines contrastive learning and generative learning
- Learns global representations from unimodal image and text embeddings
- Learns fine-grained region-level features from multimodal embeddings

Source: [CoCa: Contrastive Captioners are Image-Text Foundation Models] (Yu et. al, 2022)

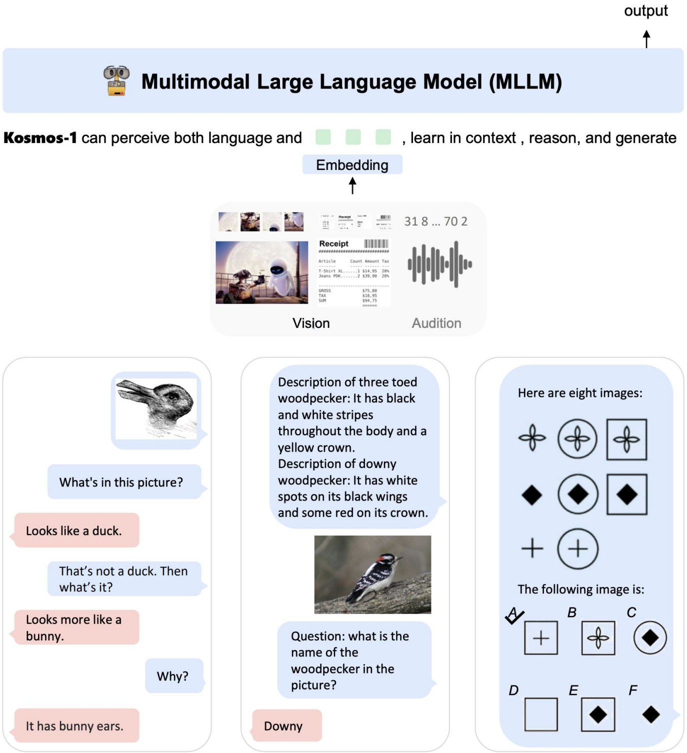
DeepMind’s Flamingo

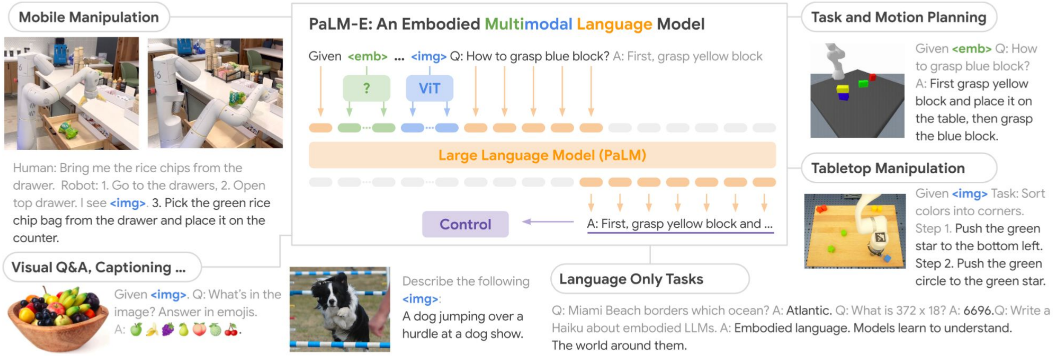


- A vision model that understands visual scenes
- A language model that helps with reasoning

Source: [Flamingo: a Visual Language Model for Few-Shot Learning] (Alayrac et. al, 2022)

Latest Vision-Language Models





Source: [PaLM-E: An Embodied Multimodal Language Model](#) (Driess et. al, 2023)

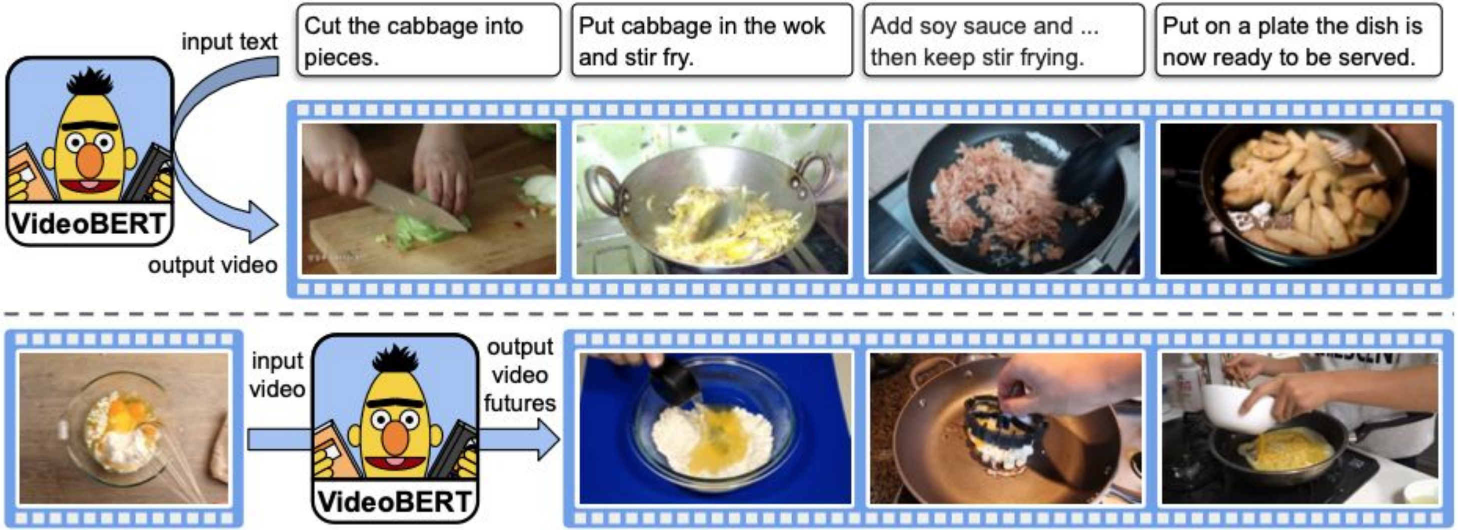
Source: [Language Is Not All You Need: Aligning Perception with Language Models](#) (Huang et. al, 2023)

The New Paradigm of Video Foundation Models

The Challenges of Video Modeling

- **The high computing burden**
  - Videos are much larger in size than text or images
  - Transformer architecture has quadratic complexity with respect to token length
- **A unique challenge to temporal modeling**
  - Videos contain a temporal dimension
  - Requires specialized techniques and models that are not commonly used in other modalities
- **Synchronized audio cues require additional processing**
  - Sounds or conversations happening within the video
  - Need the same level of attention as visual analysis

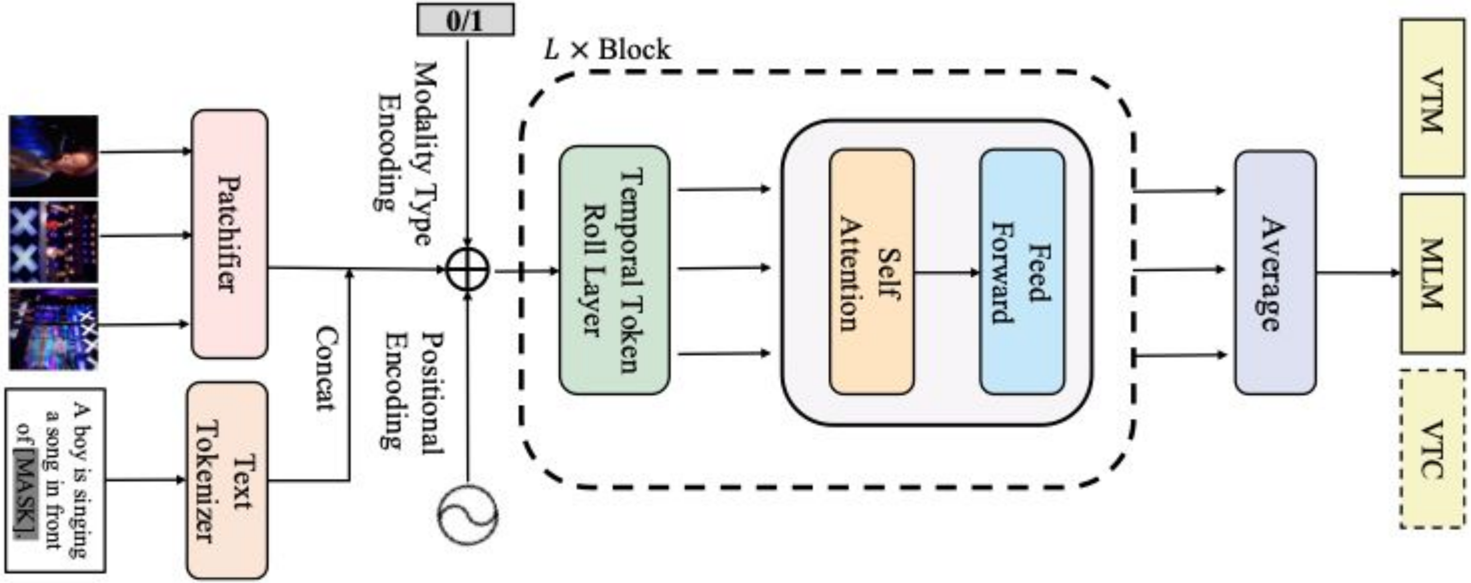
Google's VideoBERT



- Automatic speech recognition + Vector quantization for spatiotemporal visual features + a BERT model for sequences of tokens
- Outperformed existing video captioning models

Source: [VideoBERT: A Joint Model for Video and Language Representation Learning] (Sun et. al, 2019)

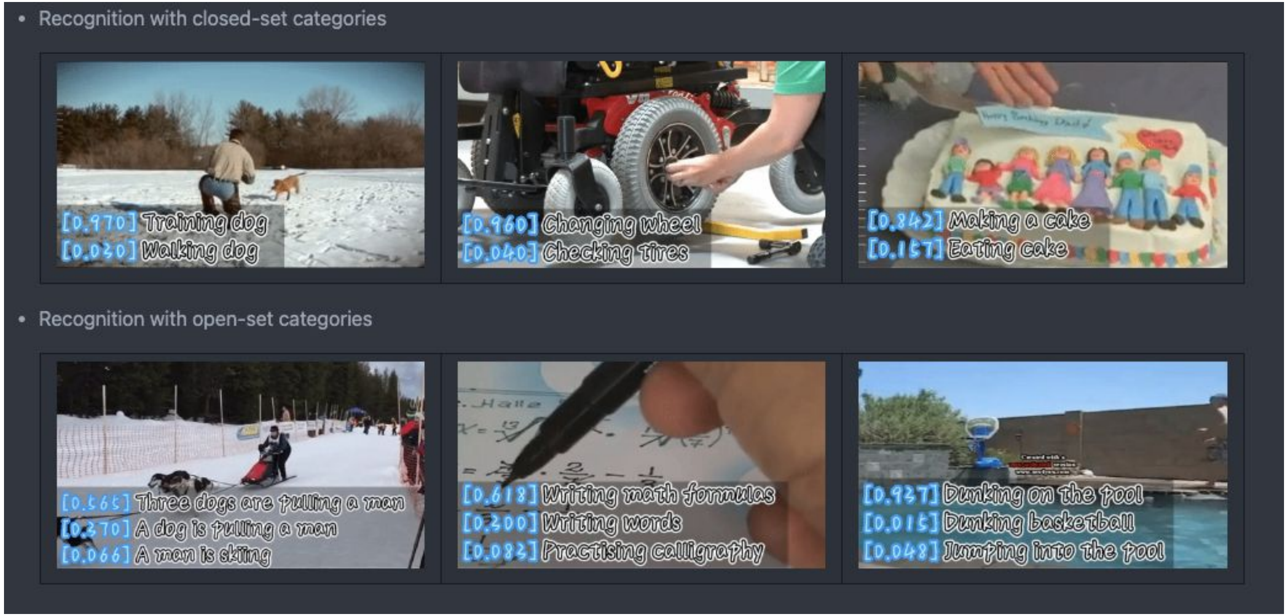
NUS's All-In-One



- Captures video-language representations from raw visual and textual signals in a unified architecture
- Uses a temporal rolling operations to capture temporal representations of sparsely sampled frames
- Performs well on video QA, text-to-video retrieval, multiple-choice QA, and visual reasoning

Source: [All in One: Exploring Unified Video-Language Pre-training](#) (Wang et. al, 2022)

Microsoft's X-Clip



- A cross-frame communication Transformer allows frames to exchange information using message tokens
- A multi-frame integration Transformer transfers frame-level representations to video-level
- Performs well in zero-shot and few-shot video recognition tasks

Source: Expanding Language-Image Pretrained Models for General Video Recognition (Ni et. al, 2022)

InternVideo

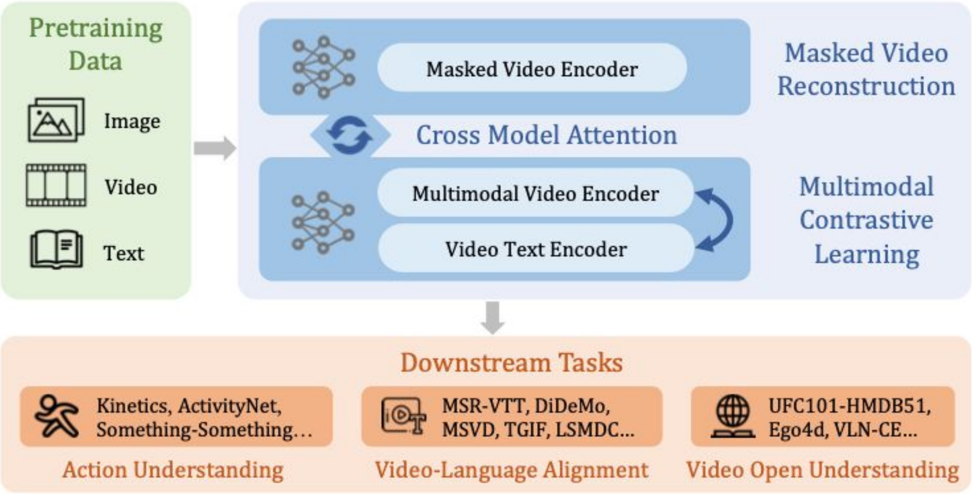


Figure 2: The overall framework of InternVideo.

- Combines masked video modeling and multimodal contrastive learning
- Uses learnable interactions to derive new features from these two Transformers
- Outperformed models in action understanding, video-language alignment, and open-world video applications tasks

Source: InternVideo: General Video Foundation Models via Generative and Discriminative Learning (Wang et. al, 2022)

MERLOT Reserve and VideoCoCa

Figure 1: MERLOT RESERVE learns *multimodal neural script knowledge* representations of video – jointly reasoning over video frames, text, and audio. Our model is pretrained to predict which snippet of text (and audio) might be hidden by the MASK. This task enables it to perform well on a variety of vision-and-language tasks, in both zero-shot and finetuned settings.

Source: [MERLOT Reserve: Multimodal Neural Script Knowledge through Vision and Language and Sound](#) (Zellers et. al, 2022)

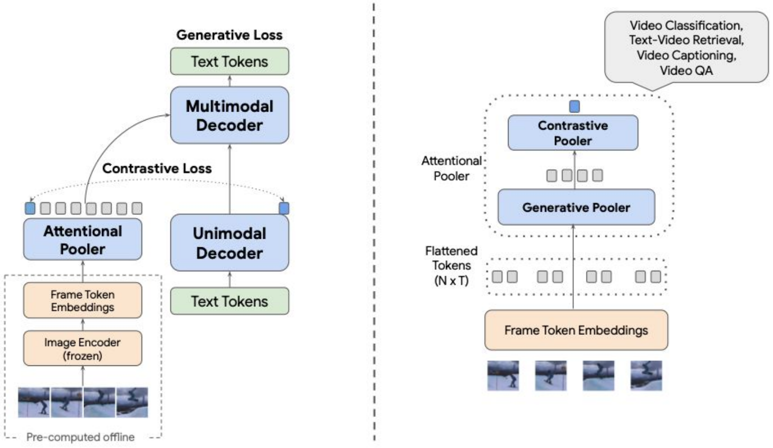
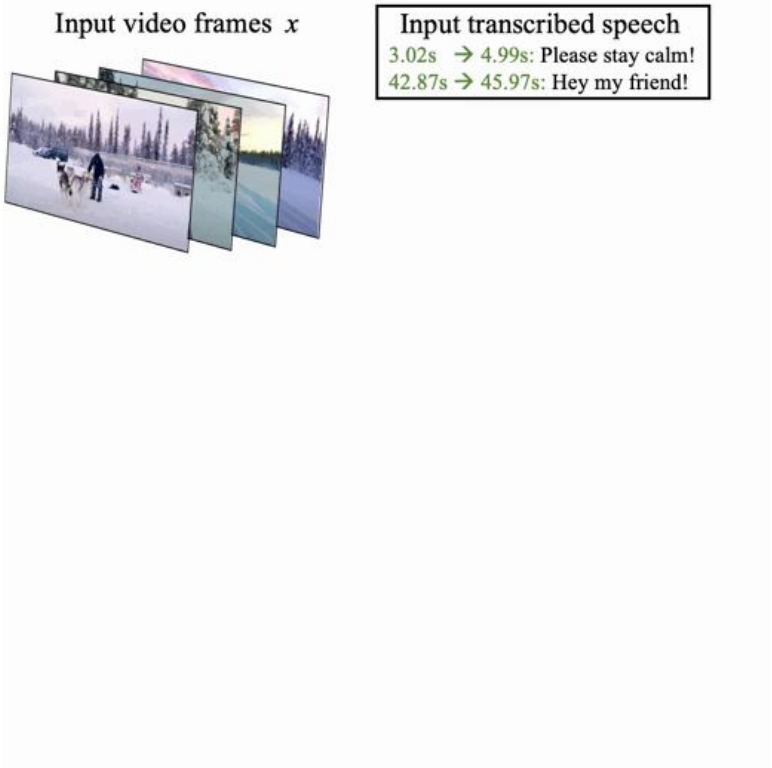


Figure 1. **Left**: Overview of VideoCoCa. All weights of the pretrained CoCa model are reused, without the need of learning new modules. We compute frame token embeddings offline from the frozen CoCa image encoder. These tokens are then processed by a generative pooler and a contrastive pooler on all flattened frame tokens, yielding a strong zero-shot transfer video-text baseline. When continued pretraining on video-text data, the image encoder is frozen, while the attentional poolers and text decoders are jointly optimized with the contrastive loss and captioning loss, thereby saving heavy computation on frame embedding. **Right**: An illustration of the attentional poolers and flattened frame token embeddings. We flatten $N \times T$ token embeddings as a long sequence of frozen video representations.

Source: [VideoCoCa: Video-Text Modeling with Zero-Shot Transfer from Contrastive Captioners](#) (Yan et. al, 2023)

Vid2Seq and Track Anything



Source: Track Anything: Segment Anything Meets Videos (Yang et. al, 2023)

Source: Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning (Yang et. al, 2023)

Conclusion

1. **A Gentle Introduction to Foundation Models**
   a. Transfer Learning and Word Embeddings
   b. Transformers and Their Variants
2. **The Birth of Large Language Models**
   a. OpenAI's GPTs, Google's T5 and BERT
   b. Scaling Laws -> Emergent Abilities of LLMs
3. **The Rise of Large Vision-Language Models**
   a. Open AI's CLIP
   b. Google's CoCa, DeepMind's Flamingo
   c. Microsoft's Kosmos-1, Google's PaLM-E
4. **The New Paradigm of Video Foundation Models**
   a. Challenges of Video Modeling
   b. VideoBERT, All-In-One, X-CLIP, InternVideo
   c. MERLOT Reserve, VideoCoCa, Vid2Seq, Track Anything
   d. Twelve Labs' Marengo!

Twelve Labs

Full Blog Post

https://app.twelvelabs.io/blog/foundation-models-are-going-multimodal

Join Our Discord