

WEBINAR Foundation Models are Going Multimodal

James Le Developer Experience, Twelve Labs

July 27, 2023 | 9:00 AM PT



GPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGP

TchatGPTchatGPTc PTchatGPTchatGPT GPTchatGPTchatGPI GPTchatGPTchatGI tGPTchatGPTchatC atGPTchatGPTchat hatGPTchatGPTchat GPT-4

chatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTchatGPTch

GitHub Copilot



GPTchatGF

tGPTchatG

natGPTchat

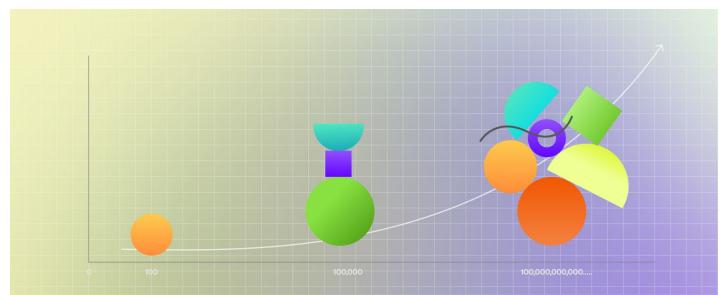
hatGPTcha

chatGPTc

TchatGPT

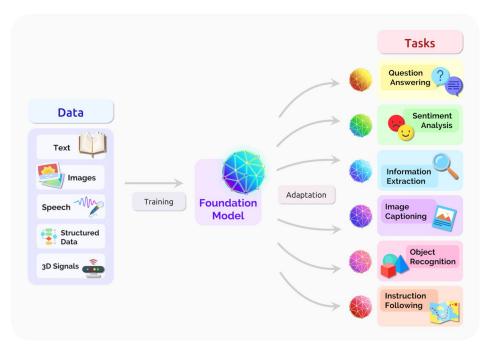
Agenda

- 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



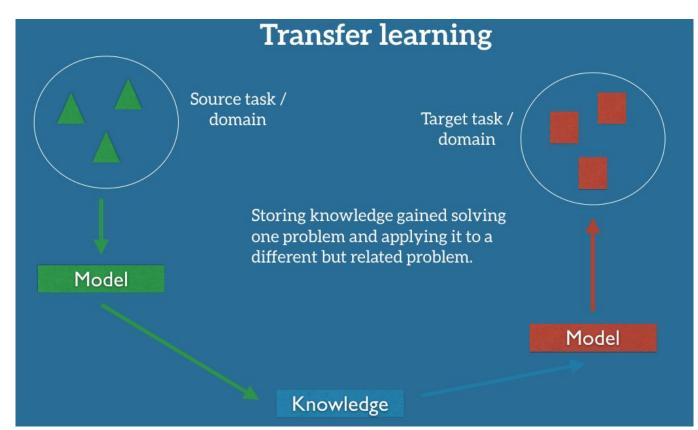
A Gentle Introduction to Foundation Models

What Is A Foundation Model?



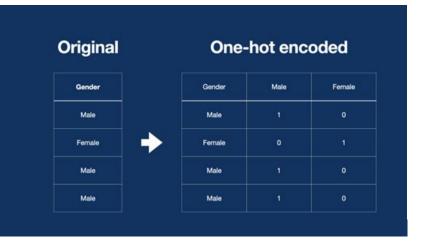
- Learns from a wide range of data using self-supervision at scale
- Leverages deep neural networks and self-supervised learning
- Useful for various downstream tasks

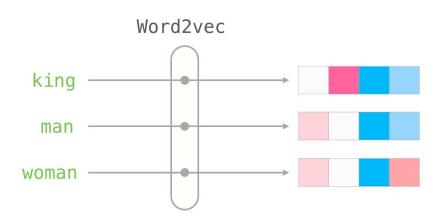
Transfer Learning



800

Word Embeddings





- One-hot encoding encodes words as vectors (embeddings)
- Word2vec maximizes cosine similarity between word embeddings
- Models like <u>ELMo</u>, <u>ULMFiT</u>, and <u>GPT</u> 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

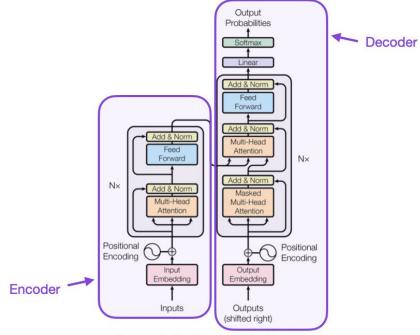
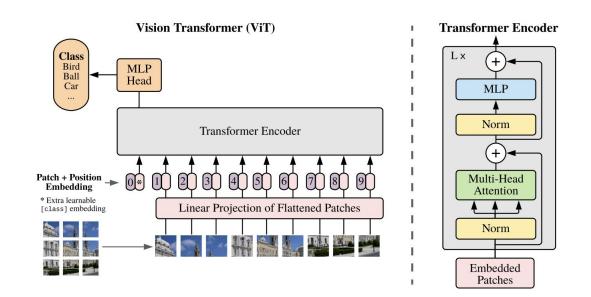


Figure 1: The Transformer - model architecture.

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

Vision Transformers

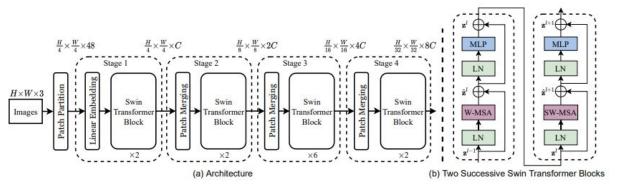
- Split an image into patches
- Treat image patches as token inputs
- Exceed SOTA results on image classification tasks
- Require a lot of compute power
- Not useful for tasks that involve visual elements of varying size



Transformer Variants

Swin Transformers

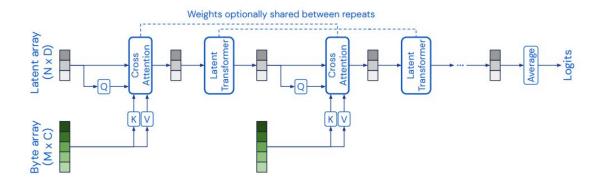
- Hierarchical feature maps
- Shifted window attention



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

Perceiver

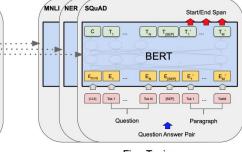
- Latent units to form an attention bottleneck
- Associate position- and modality-specific features with each input



The Birth of Large Language Models

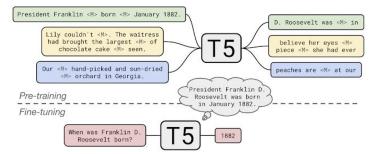
Large Language Models







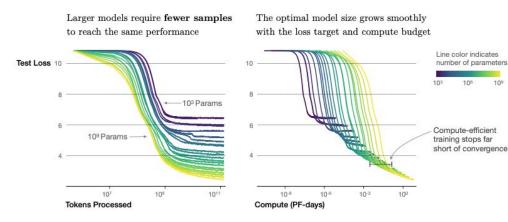
Source: Language Models are Unsupervised Multitask Learners (Radford et. al, 2018)



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

Source: <u>BERT: Pre-training of Deep Bidirectional Transformers for</u> Language Understanding (Devlin et. al, 2018)

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



Source: <u>Scaling Laws for Neural Language Models</u> (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

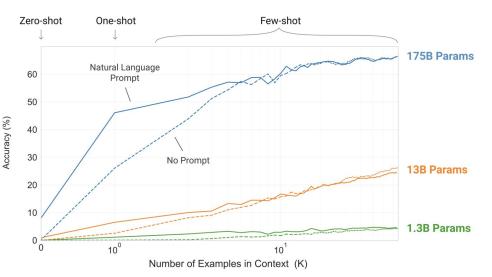


Table 1: List of emergent abilities of large language models and the scale (both training FLOPs and number of model parameters) at which the abilities emerge.

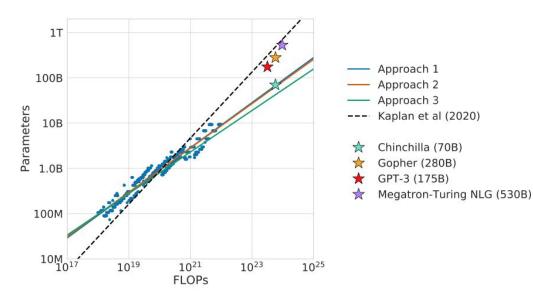
	Emergent scale			
	Train. FLOPs	Params.	Model	Reference
Few-shot prompting abilities				
Addition/subtraction (3 digit)	2.3E + 22	13B	GPT-3	Brown et al. (2020)
• Addition/subtraction (4-5 digit)	3.1E + 23	175B		
 MMLU Benchmark (57 topic avg.) 	3.1E + 23	175B	GPT-3	Hendrycks et al. (2021a
 Toxicity classification (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: Language Models Are Few-Shot Learners (Brown et. al, 2020)

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.

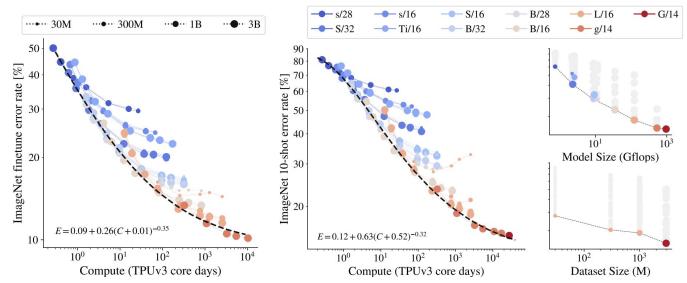
"Chinchilla" Scaling Laws (by DeepMind)



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

- More accurate than the OpenAl'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



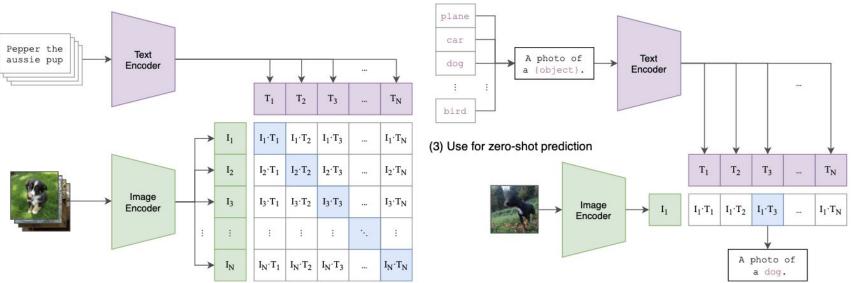
- Experiments with Vision Transformers on a range of parameters (5M-2B), a range of datasets (1M-3B), and a range of compute budgets (1-10,000 TPUs)
- Simultaneously scaling total compute and model size is effective
- Larger models perform better in few-shot learning

Source: <u>Scaling Vision Transformers</u> (Zhai et. al, 2022)

The Rise of Large Vision-Language Models

OpenAl's CLIP

(1) Contrastive pre-training

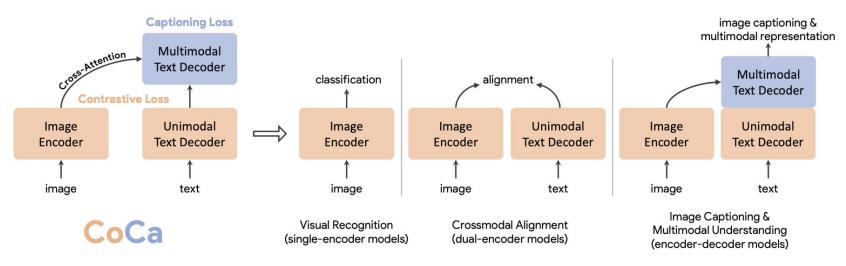


(2) Create dataset classifier from label text

- 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) 1

Google's CoCa



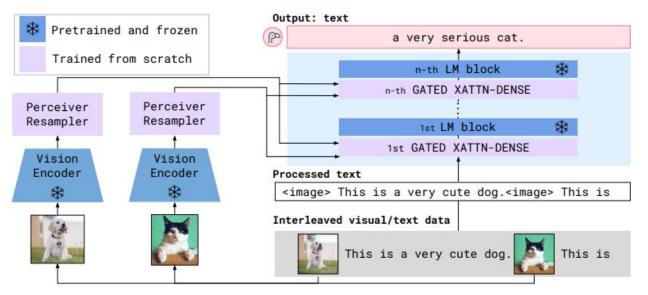
Pretraining

Zero-shot, frozen-feature or finetuning

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

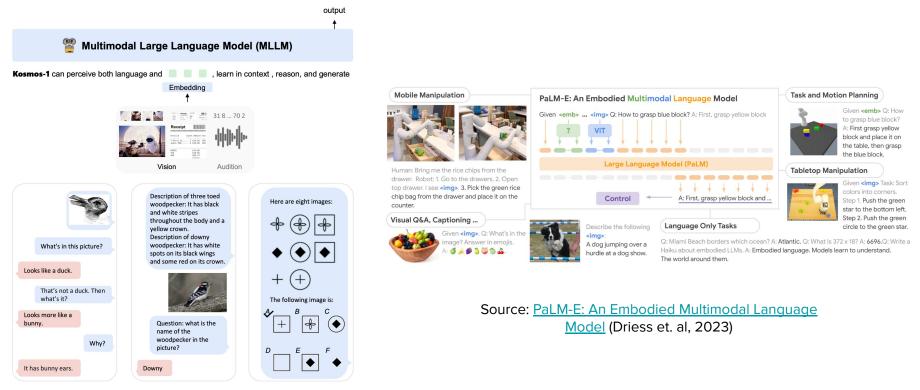
Source: <u>CoCa: Contrastive Captioners are Image-Text Foundation Models</u> (Yu et. al, 2022)

DeepMind's Flamingo



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

Latest Vision-Language Models



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

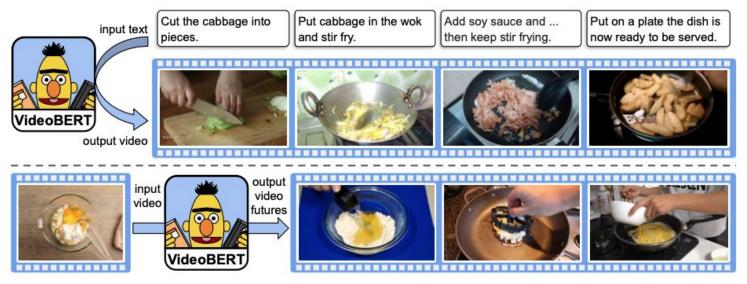
• A unique challenge to temporal modeling

- Videos contain a temporal dimension
- Requires specialized techniques and models that are not commonly used in other modalities

• Synchronized audio cues require additional processing

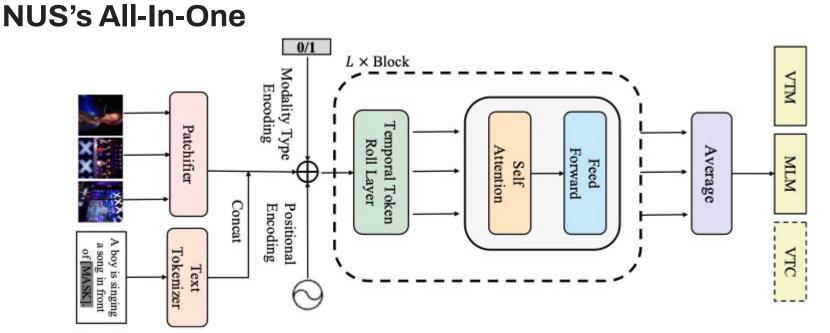
- Sounds or conversations happening within the video
- Need the same level of attention as visual analysis

Google's VideoBERT



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

24

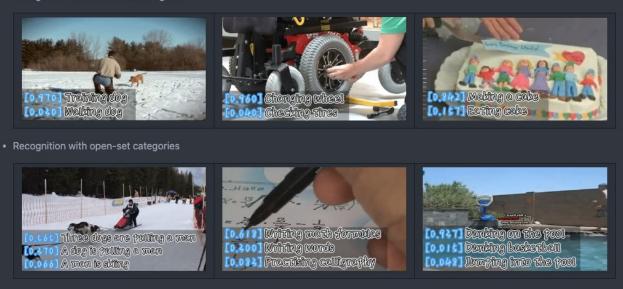


- Captures video-language representations from raw visual and textual signals in a unified architecture
- Uses a temporal rolling operations to capture temporal representations of sparsely sampled frames
- Performs well on video QA, text-to-video retrieval, multiple-choice QA, and visual reasoning

Source: <u>All in One: Exploring Unified Video-Language Pre-training</u> (Wang et. al, 2022)

Microsoft's X-Clip

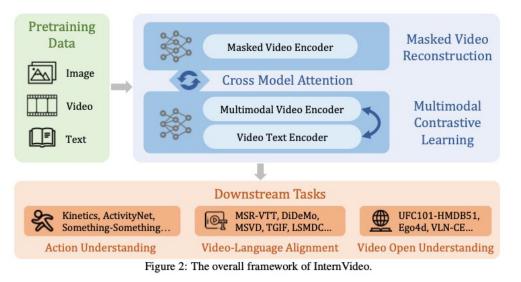
Recognition with closed-set categories



- 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: <u>Expanding Language-Image Pretrained Models for General Video Recognition</u> (Ni et. al, 2022)

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) 27

MERLOT Reserve and VideoCoCa



Figure 1: **WERLOT** 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: <u>MERLOT Reserve: Multimodal Neural Script Knowledge</u> <u>through Vision and Language and Sound</u> (Zellers et. al, 2022)

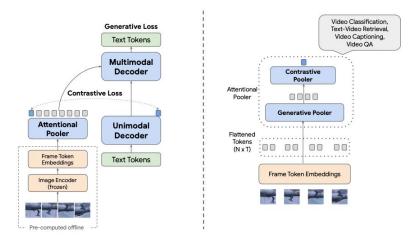


Figure 1. Left: Overview of VideoCoCa. All weights of the pretrained CoCa model are reused, without the need of learning new modules. We compute frame token embeddings offline from the frozen CoCa image encoder. These tokens are then processed by a generative pooler and a contrastive pooler on all flattened 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

Input video frames x



Input transcribed speech $3.02s \rightarrow 4.99s$: Please stay calm! $42.87s \rightarrow 45.97s$: Hey my friend!



Source: <u>Track Anything: Segment Anything Meets Videos</u> (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. OpenAl's GPTs, Google's T5 and BERT
- b. Scaling Laws -> Emergent Abilities of LLMs
- 3. The Rise of Large Vision-Language Models
 - a. Open Al'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

mesLe 🕴 🎧 💠

Join Our Discord



Multimodal Minds ~	mannouncement Follow For community announcement	7월4 🜲 🖈 🚨 Search	۵ 🖬 🖉
GOAL: LVL 1 0/2 Boosts >			
	JamesLe 03/02/2023 4:00 PM Hey @everyone, again welcome to Multimodal Minds - whether you are old folks who have been here for a while or new folks who just joined our community.	04**065	/ # 4 0
	This is the list of channels and their descriptions in our Server in order to help you get the most out of our community:		
	Tart Here:		
# welcome-party	#welcome-party: This is where new members get welcomed. #rules: This is the code of conduct for our community.		
I Tules	## Transversement. This is a place for all major community-wide announcements.		
📹 🎮 announcement 🗠 o			
	🚅 Lounge:		
-ILOUNGE +	 # \$introduce-yourself: This is where new members introduce themselves. We recommend you share details about who you are, how you are connected with Multimodal. # and accussions: This is for organic general discussions between community members. 	AI, and what you are interested in.	
# Wintroduce-vourself	 # a discussions: This is for organic general discussions between community members. # / shameless-promotion: This is where you can promote your product/services. 		
Hi I am Calvin our compa			
	🗢 Community:		
# discussions	• # multimodal-ai: This is where we share latest updates in Multimodal AI research and applications.		
#≟ ∕ [®] shameless-promotion	# gehare-resources: This is where we share other relevant information on Generative AI, Large-Language Models, AI Ethics, AI Infrastructure, etc.		
COMMUNITY +	 # gevents-and-meetups: This is where we keep track of events and meetups (mostly local to the San Francisco Bay Area). # @ event-n-meetups-discussions: This is where we share highlights and pictures of events and meetups we attend. 		
# 💼 multimodal-ai	 # event this metupe sources and the state is a state ingraging and pictures of vertices we attempt we attempt the state is a state of the state of		
# Share-resources	 # Each are your work: This is where you share your project, research, articles, videos (anything you created). 		
₩ ∰events-and-meetups	# video-engineering: This is for discussions around specific video engineering topics such as video encoding, audio encoding, video broadcasting, video publishing, etc.		
	More below		
	4 <i>8</i> 2		
揖 🗣 event-n-meetups-dis	JamesLe 05/08/2023 4:39 PM		
建 ? ask-for-help	Cantinuing from above @everyone		
# share-your-work	I Twelve Labs Product:		
GitHub - groovybits/gaib	 #module Labs Product: #module Labs Product: #module Labs Product. 		
Live Interviews from CV	##Approduct-feedback: This is for giving feedback as you use the Twelve Labs product.		
	# "Nbug-reports: This is for reporting bugs that you encounter while using the Twelve Labs product.		
	• # 🌢 feature-requests: This is for requesting new features that you might want to see on the Twelve Labs product.		
	• ## 20 provide the provided the second		
# 🔤cool-queries	 # M product-support: This is where you can ask any product-centric questions. 		
指 Aproduct-feedback	Twelve Labs Platform:		
# Sbug-reports	Quantization: This is a forum for conversations about Twelve Labs Classification APL		
	Q) Search : This is a forum for conversations about Twelve Labs Search API.		
# offeature-requests	 Qit playground: This is a forum for conversations about Twelve Labs Playground. 		
描 😨 🛄 applications-and-t	😨 Al Essentials Library:		
# Mproduct-support	 # ai-100-feet-overview: This is a list of resources for people who are new to AI. 		
THI I TWELVE LABS PLATFORM +	## ai-4-non-technical: This is a list of resources for non-technical folks to learn Al.		
Ch Classification	 # al-4-developers: This is a list of resources for developers to learn Al. 		
Ch Search	 #rai-4-advanced-practitioners: This is a list of resources for advanced practitioners to level up their Al game. 		
	Qforum-discussions: This is a forum to discuss how you have utilized these resources (and whether you want to add more).		
Q₁ 🛝 playground			*
	Message #imannouncement		# 🖾 🖺 💿