Multimodality in Vision and Language Models

Monica Riedler

Master seminar “The GPTs and their Ilk”

Prof. Hinrich Schütze, Haotian Ye, M.Sc.

23.06.2023

Overview

1. Introduction
2. Applications of Multimodality
3. Multimodal Models (Vision-Language)
   - CLIP
   - Stable Diffusion
   - PaLI
4. Conclusion

Introduction

Current Landscape of AI models:

- Large Language Models (LLMs) like GPT-3 and ChatGPT have become increasingly powerful, revolutionizing how computers understand and generate language.
- Vision models, such as GANs, VAEs, and diffusion models, have made significant advancements, enhancing the computer's ability to process and create visual content.

Combining Modalities - Text, Speech, and Image:

- By combining text, speech, and image capabilities, we can create AI models that are even more powerful and versatile.
- This combination allows AI systems to understand and work with different types of information, leading to more comprehensive and adaptable solutions.

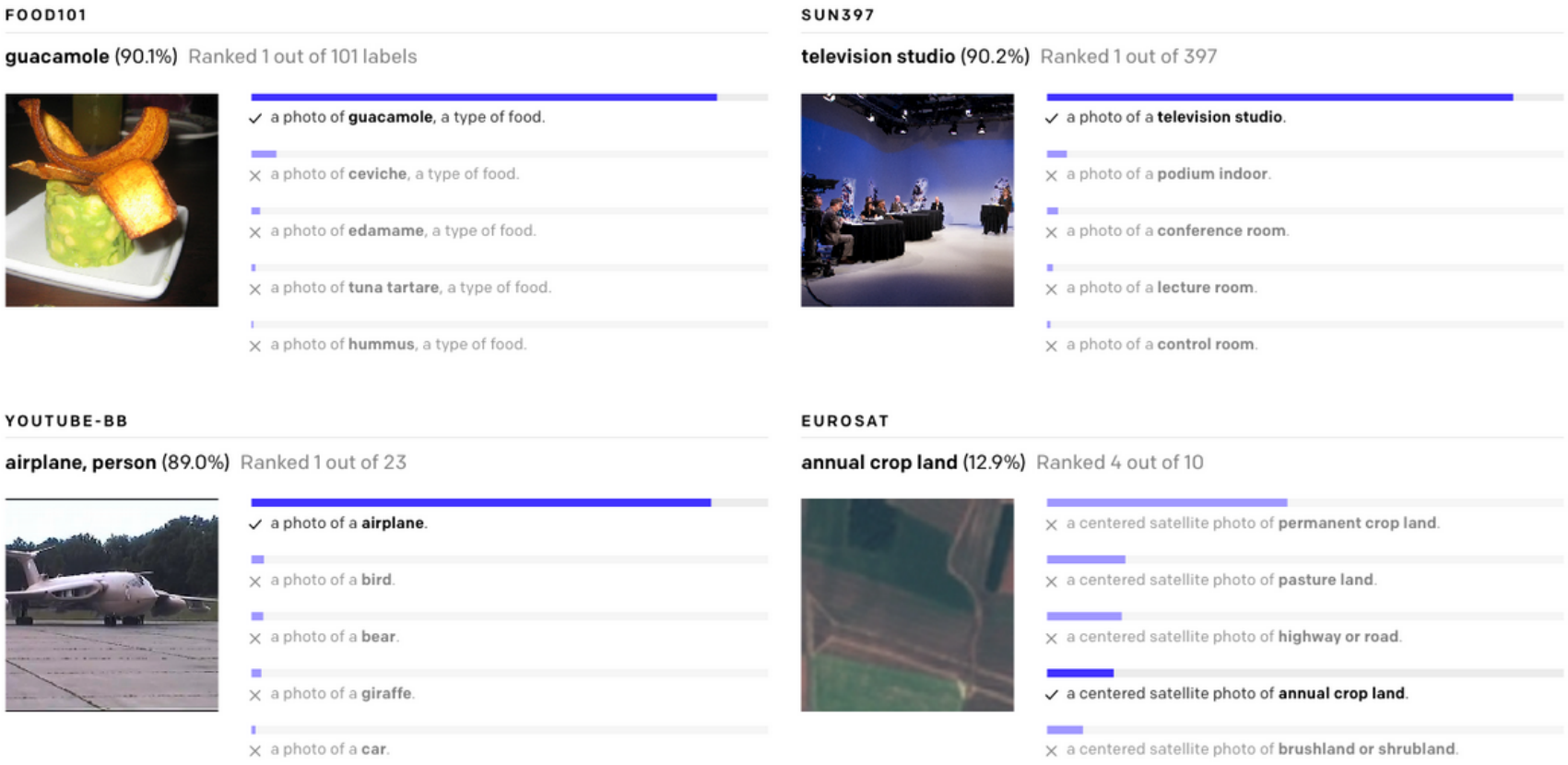
Multimodality:

- Multimodal learning occurs when a machine learning model processes different input types, known as modalities, such as image, text, and audio, and learns to represent them in a shared feature space.
- This enables multimodal models to capture relationships and correlations between different modalities for improved understanding and decision-making.

Applications of Multimodality

- Vision-Language Applications:
  - Zero-Shot Image Classification (**CLIP**)
  - Text-To-Image Retrieval (ALIGN)
  - Text-To-Image Generation (**Stable Diffusion**, DALL-E)
  - VQA: Visual Question Answering (**PaLI**)
- Other modalities:
  - Text-To-Speech or vice-versa (SLAM, mSlam)
  - Emotion Recognition (COGMEN)

Zero-Shot Image Classification (CLIP)



Images source: https://openai.com/research/clip

Text-To-Image Generation (Stable Diffusion)

Prompt:
Beautiful waterfall in a lush jungle, with sunlight shining through the trees

Prompt:
Glimpses of a herd of wild elephants crossing a savanna

Images source: https://stability.ai/stablediffusion

Visual Question Answering (PaLI)

"Answer in EN: How fast could you travel on this?"

Model Output: 500 mph

"Answer in EN: What did this organism evolve from?"

Model Output: Dinosaur

Images source: https://sites.research.google/pali/

CLIP: Contrastive Language-Image Pretraining

What does CLIP do?

- CLIP is a multimodal vision and language model developed by OpenAI.
- It learns to associate images and their textual descriptions.
- Its task formulation is general enough to derivate other tasks from it like image classification.

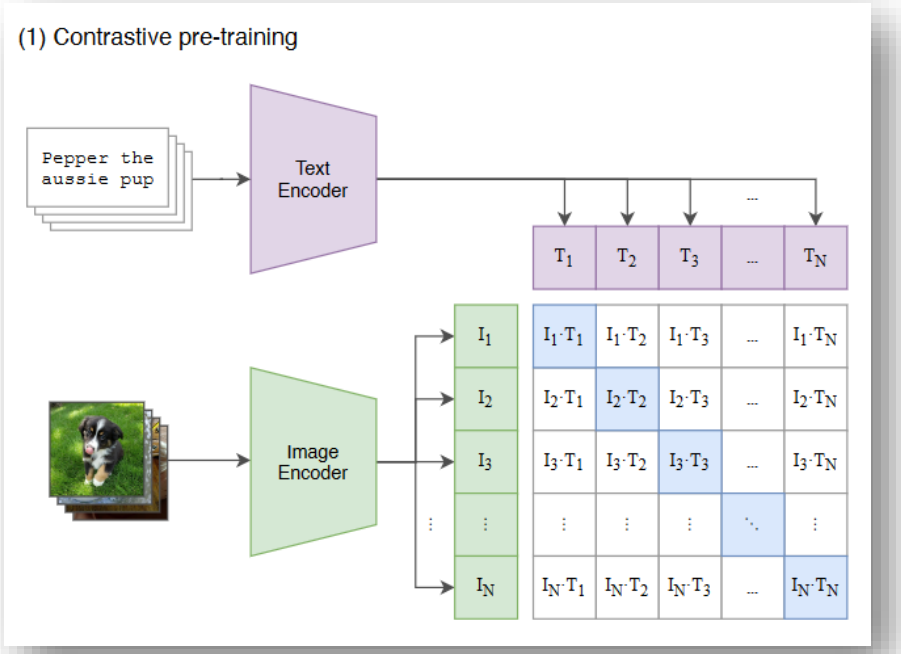
Advantages of CLIP:

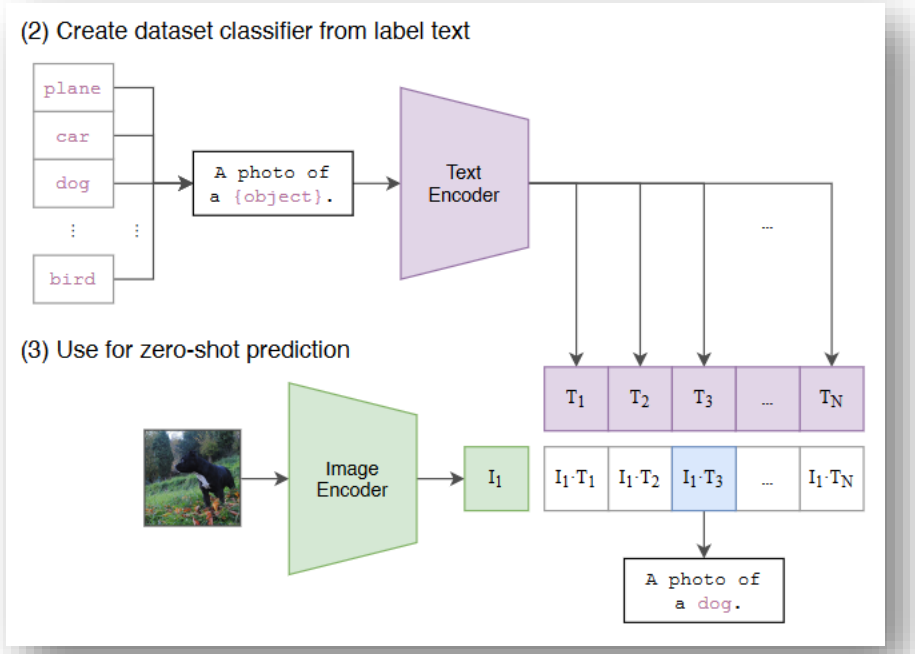
- Unlike existing approaches, CLIP is not limited to the dataset classes due to its training procedure.
- It possesses extensive knowledge of English words, allowing it to formulate English classes that encompass more language than just the dataset classes.
- This capability enables CLIP to generalize to new classes, exhibiting impressive performance on previously unseen datasets and classes, including datasets for OCR, facial emotion recognition, texture detection and action recognition in videos.
- Impressive zero-shot performance achieved by a model trained solely to predict the similarity between a text prompt and an image.

CLIP: Contrastive Language-Image Pretraining

How does CLIP work?

- CLIP is trained on a large dataset of image-text pairs from the internet.
- It leverages a **contrastive learning** approach, which maximizes the similarity between an image and its corresponding description while minimizing the similarity between the image and an unrelated description.
- It uses a transformer architecture, which enables it to capture complex relationships between images and text.





Images source: Learning transferable visual models from natural language supervision (Radford et al., PMLR 2021)

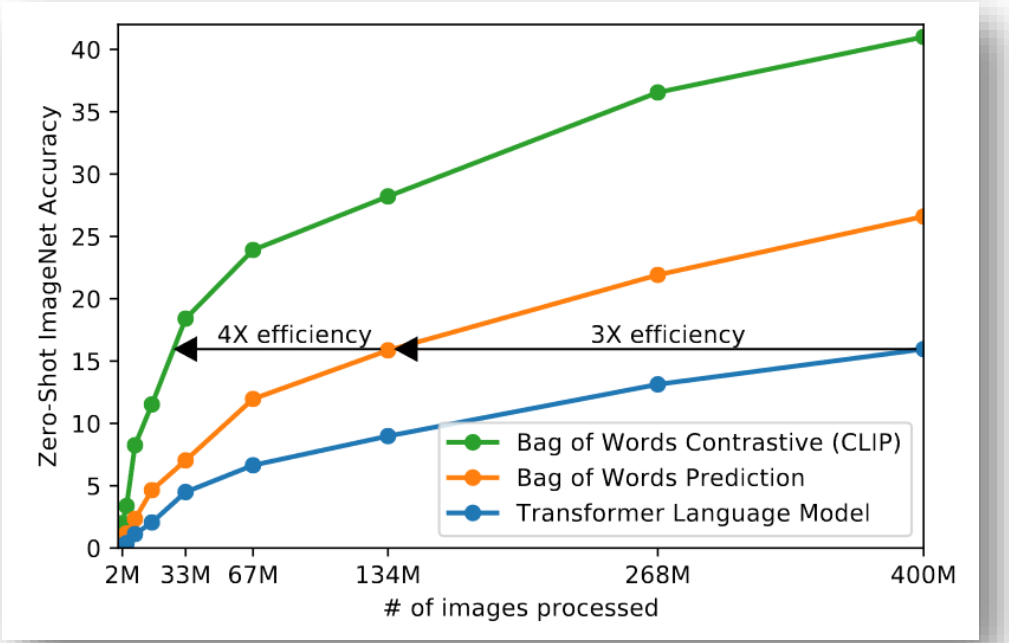
CLIP: Contrastive Language-Image Pretraining

Why is CLIP so successful?

- **Amount of data:**
  Impressive zero-shot performance was possible thanks to the very large and diverse amount of training data (400 million image text-pairs, 100 times more data than ImageNet).
- **Shared vision and language representations:**
  Embeddings are forced into a shared feature space making them ready for a variety of image recognition tasks on a diverse range of datasets (OCR, textures, action recognition).
- **Computational efficiency:**
  - Transformer architecture: Both encoders in CLIP utilize transformers, enabling efficient parallelization compared to architectures such as ResNet.
  - Contrastive learning approach: The use of a contrastive objective proved to be 4x more efficient compared to a predictive objective.

CLIP: Contrastive Language-Image Pretraining

Performance Evaluation:



Other models:

- Image-to-caption language models with Bag of Words Prediction and Transformers.
- Given an image, predict the exact caption of the image.

CLIP:

- Easier proxy task of predicting only which text as a whole is paired with an image -> higher efficiency.
- Switching from a predictive to a contrastive objective -> higher accuracy.

Source: Learning transferable visual models from natural language supervision (Radford et al., PMLR 2021)

CLIP:
Contrastive Language-Image Pretraining

Limitations:

- CLIP's performance heavily depends on the quality and diversity of the training data, and it may struggle with specific domains or fine-grained tasks that were underrepresented in the training data (e.g., differentiating between species of flowers).
- CLIP shows difficulties in generalizing to out-of-distribution images.
- It requires substantial amounts of data for training.

Stable Diffusion

What does Stable Diffusion do?

- Stable diffusion is a text-to-image model which, given a text prompt, returns an image that matches the text.
- Belongs to a class of generative models called Latent Diffusion Models.

Advantages of Stable Diffusion:

- Efficiency: SD achieves a much higher degree of efficiency compared to standard Diffusion Models as the operations are applied to a reduced latent space instead of using the pixel space directly. This leads to faster training times and more computationally efficient generation of images.
- Realism: Stable diffusion models produce diverse, highly detailed images which preserve the semantic structure of the data.
- Open-source: Stable Diffusion offers an open-source alternative to OpenAI DALLE-2, making both the code and model weights freely available to the community, ensuring transparency and accessibility. (https://github.com/CompVis/stable-diffusion)

Stable Diffusion

How does Stable Diffusion work?

Diffusion models work in two steps:

- Forward Diffusion Process: Adds noise to a training image gradually turning it into an image containing only noise.
- Reverse Diffusion Process: Starting from a noisy image, recover the original image.

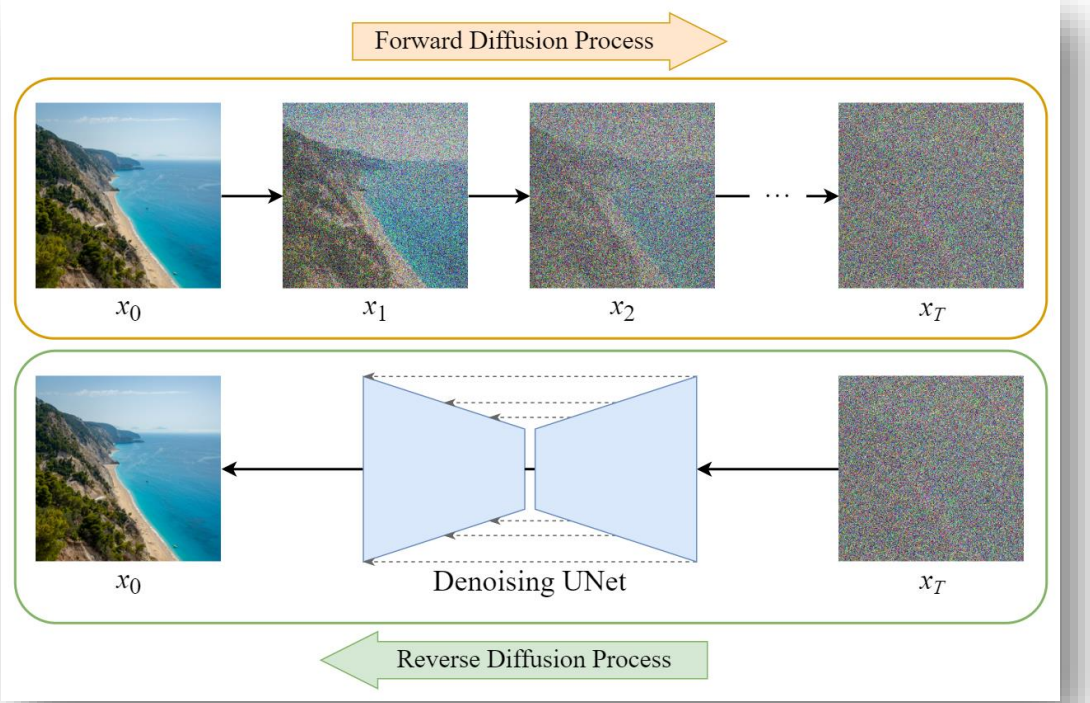


Image source: https://medium.com/@steinsfu/diffusion-model-clearly-explained-cd331bd41166

Stable Diffusion

- **Forward Diffusion Process:**
  Create training examples by adding noise to images.

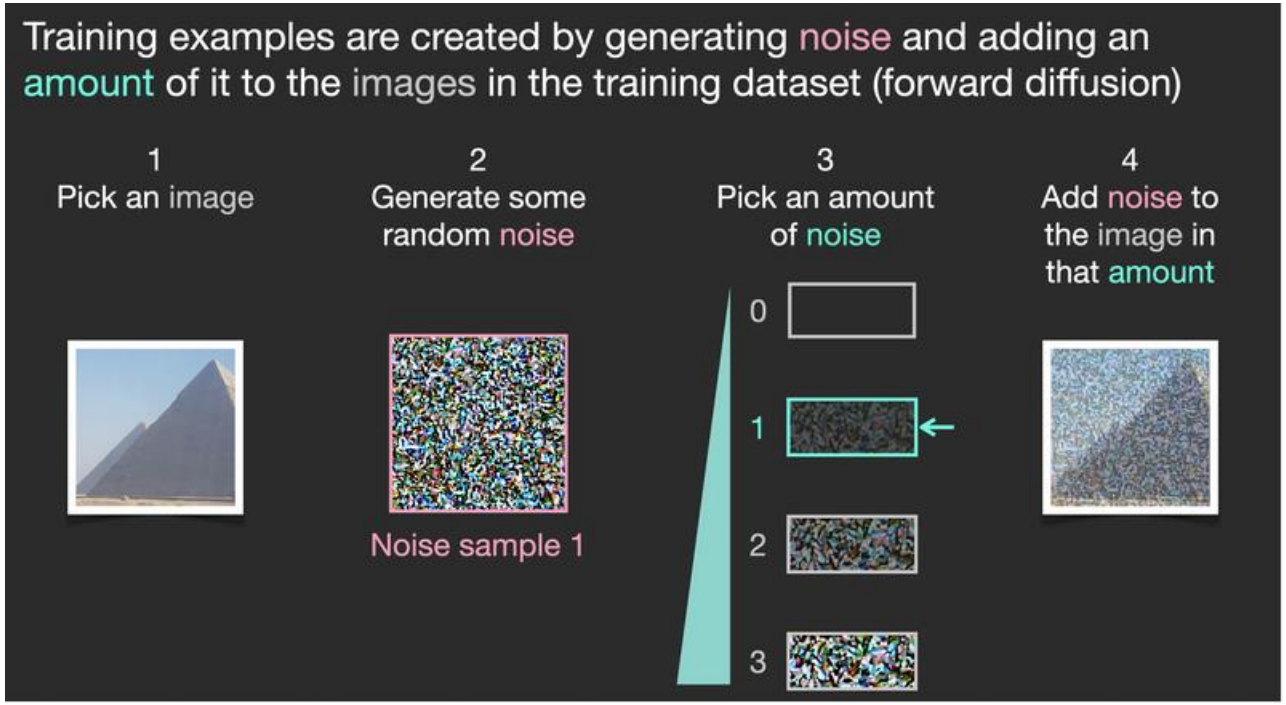


Image source: https://jalammar.github.io/illustrated-stable-diffusion/

Stable Diffusion

- **Forward Diffusion Process:** We can create many such training examples having a noisy image and an amount of noise as input, and a noise sample as output.

Image source: https://jalammar.github.io/illustrated-stable-diffusion/

Stable Diffusion

- **Forward Diffusion Process:**
  With this dataset we can train a noise predictor, which can help us generate images.

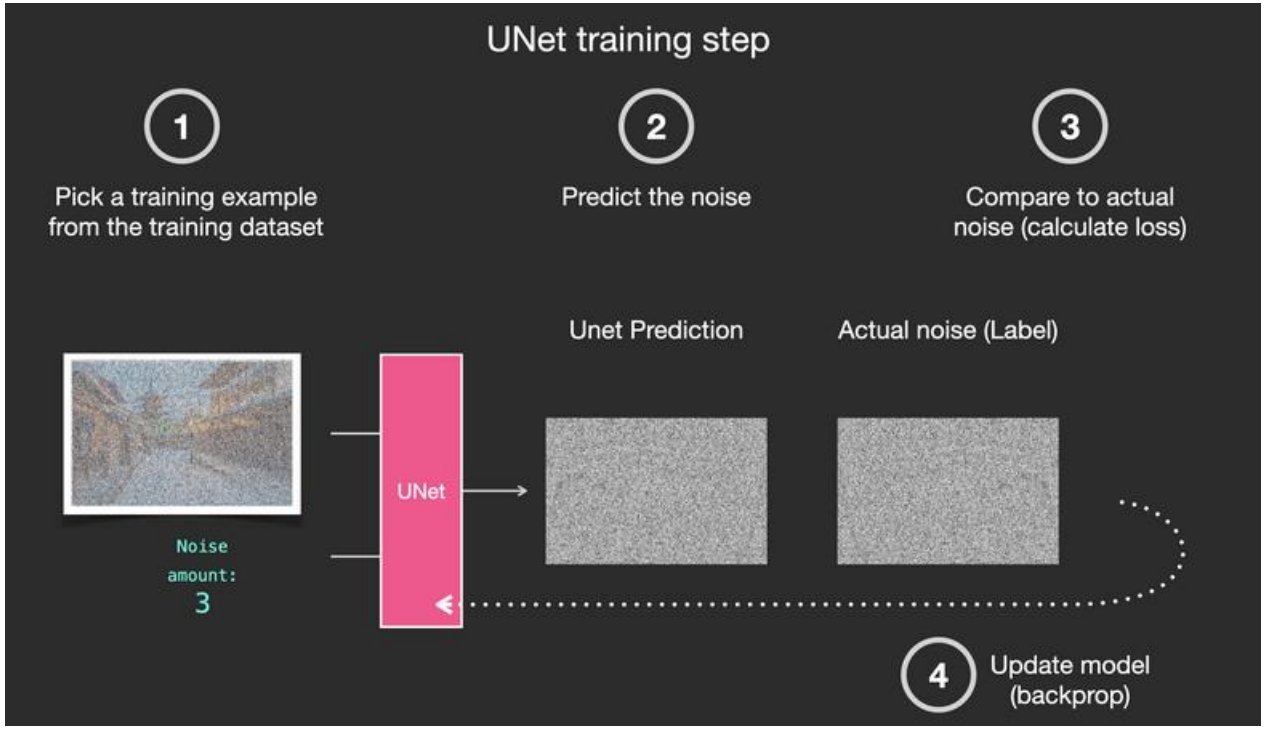


Image source: https://jalammar.github.io/illustrated-stable-diffusion/

Stable Diffusion

Reverse Diffusion Process:

- The noise predictor takes a noisy image and a noise amount as inputs and predicts a noise sample.
- By subtracting the predicted noise sample from the noisy image, we obtain an image that is closer to the original image.
- This step is iteratively repeated to refine the image reconstruction at each step, progressively moving it closer to the original image.

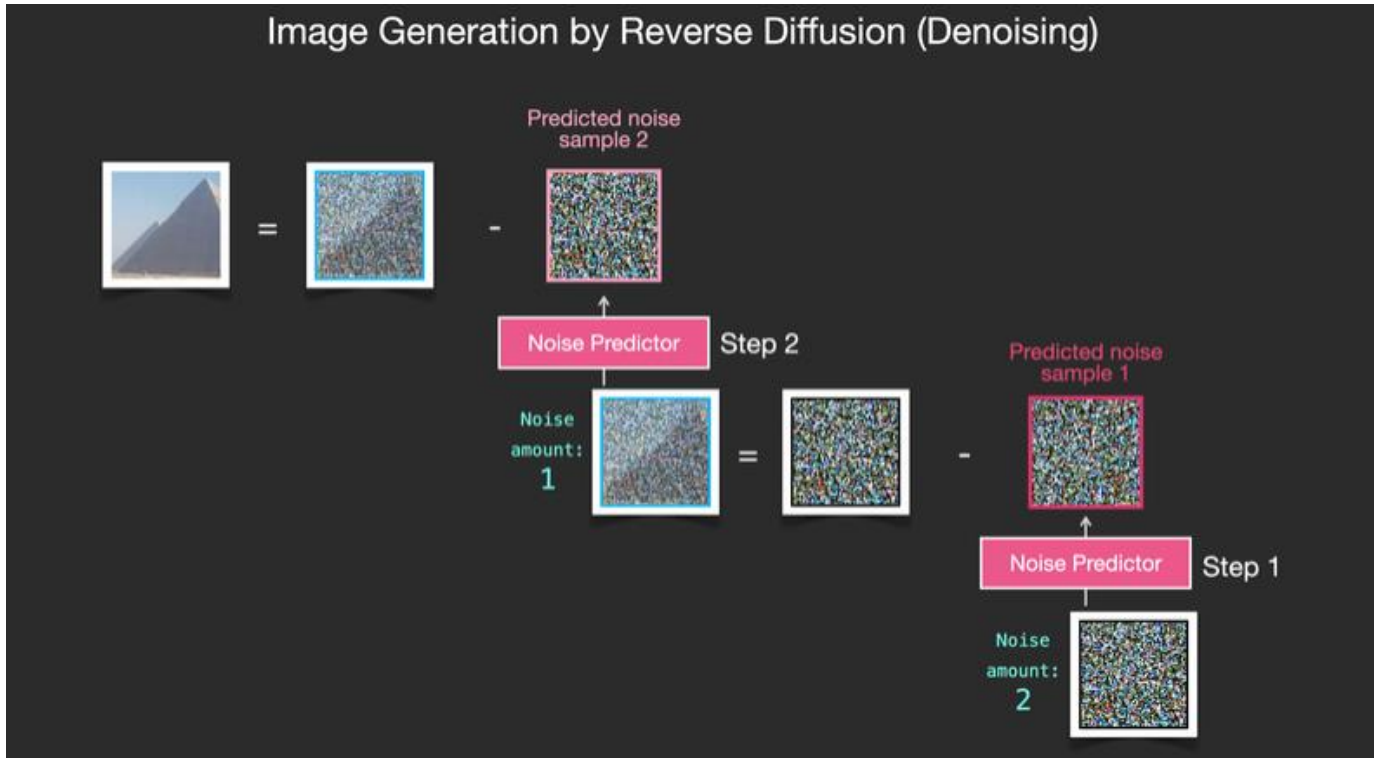


Image source: https://jalammar.github.io/illustrated-stable-diffusion/

Stable Diffusion

Latent Diffusion Model:

- The architecture described on the previous slides operates on the pixel space directly. This is inefficient and computationally very expensive!
- In Latent Diffusion Models the images are compressed to a reduced latent space using an autoencoder.
- The autoencoder compresses the image into latent space using an encoder and reconstructs it from the compressed information using a decoder.

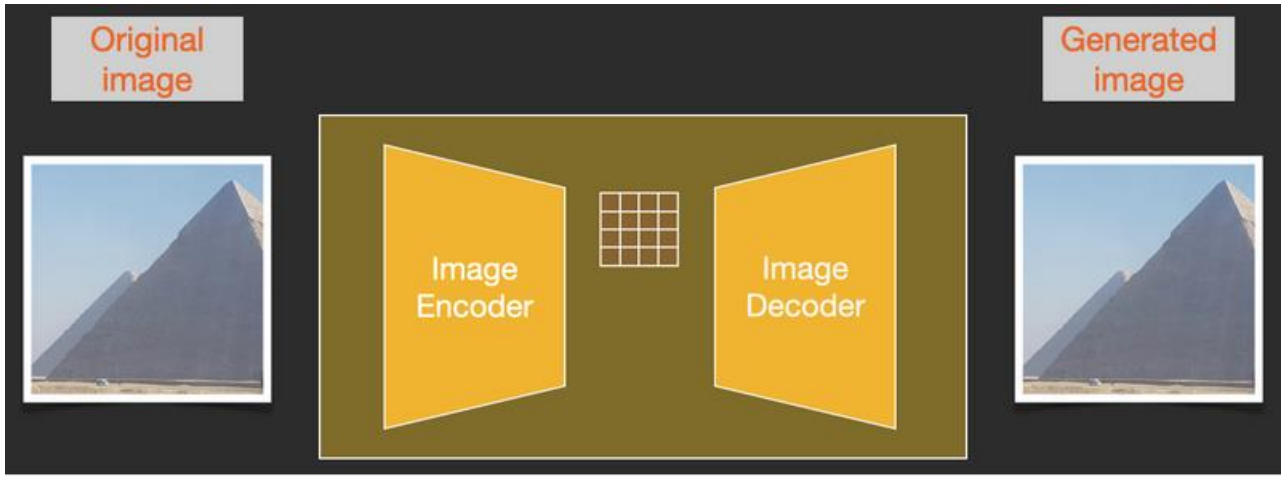


Image source: https://jalammar.github.io/illustrated-stable-diffusion/

Stable Diffusion

Text Conditioning:

- To generate images out of prompts, a language understanding component is needed, that takes a prompt as input and generates token embeddings.
- This component is needed by the model to understand the prompt and include it in the image generation process.
- The text component in the stable diffusion model is a CLIP model.

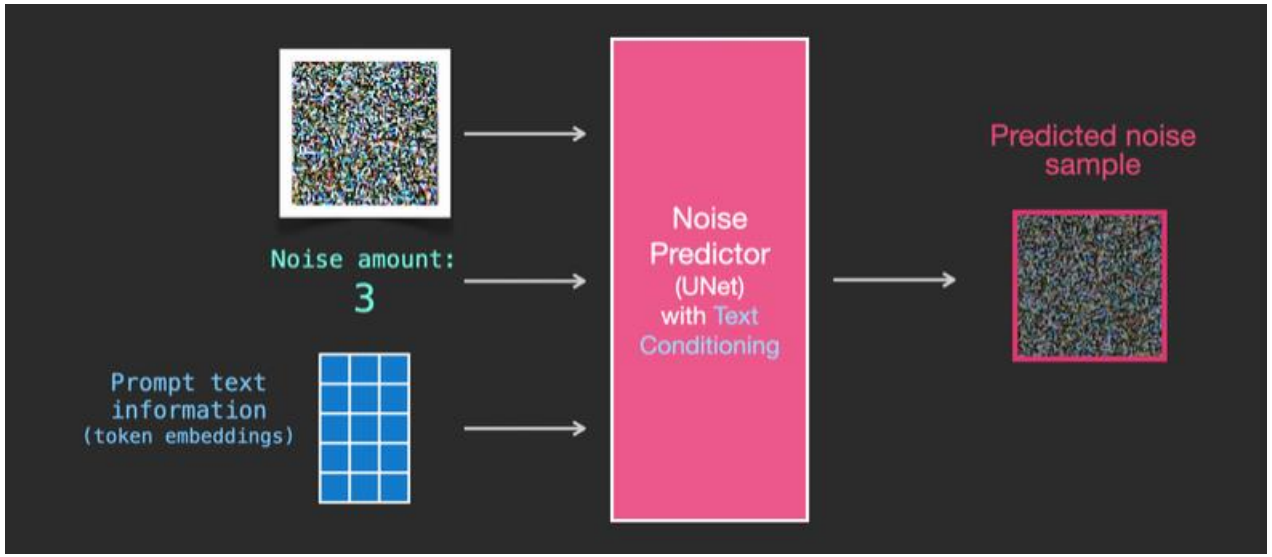


Image source: https://jalammar.github.io/illustrated-stable-diffusion/

Stable Diffusion

Attention Layers:

- The noise predictor is a UNet consisting of different ResNet layers
- An attention layer is added between the ResNet layers to merge the text representations into the ResNets
- Now the ResNets can incorporate the textual information in the processing of the noisy images to generate images that reflect the content of the text prompts.

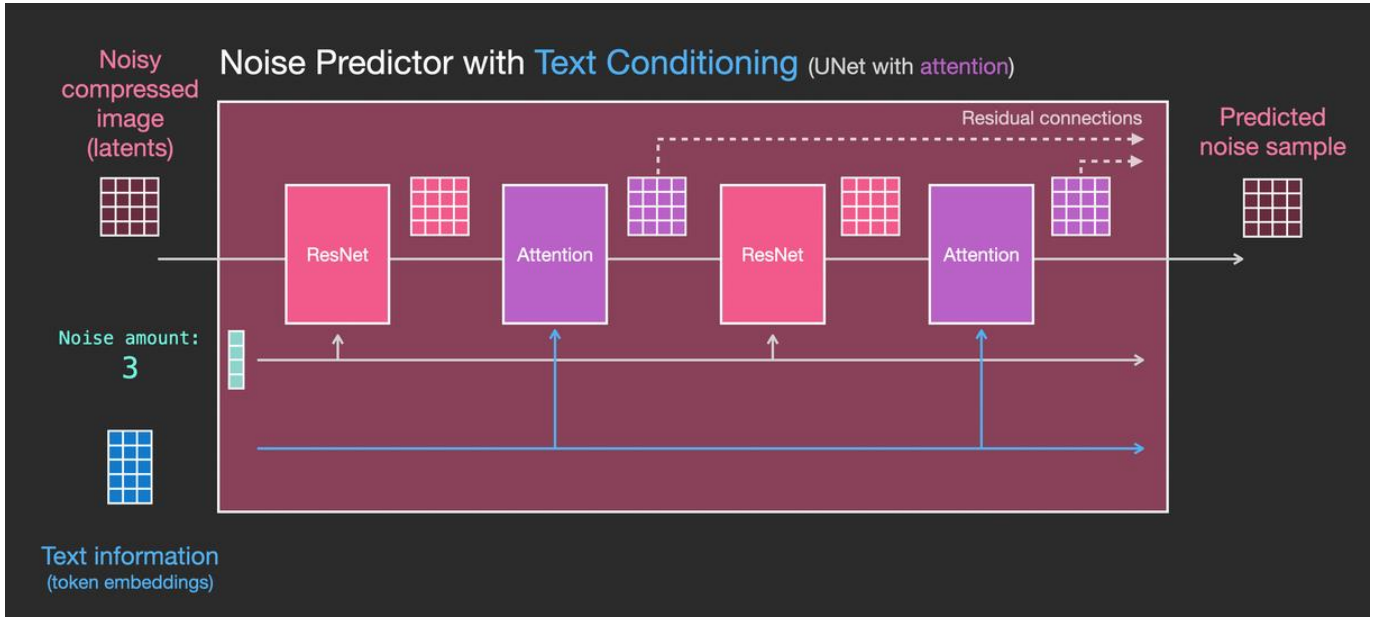


Image source: https://jalammar.github.io/illustrated-stable-diffusion/

Stable Diffusion

Putting it all together:

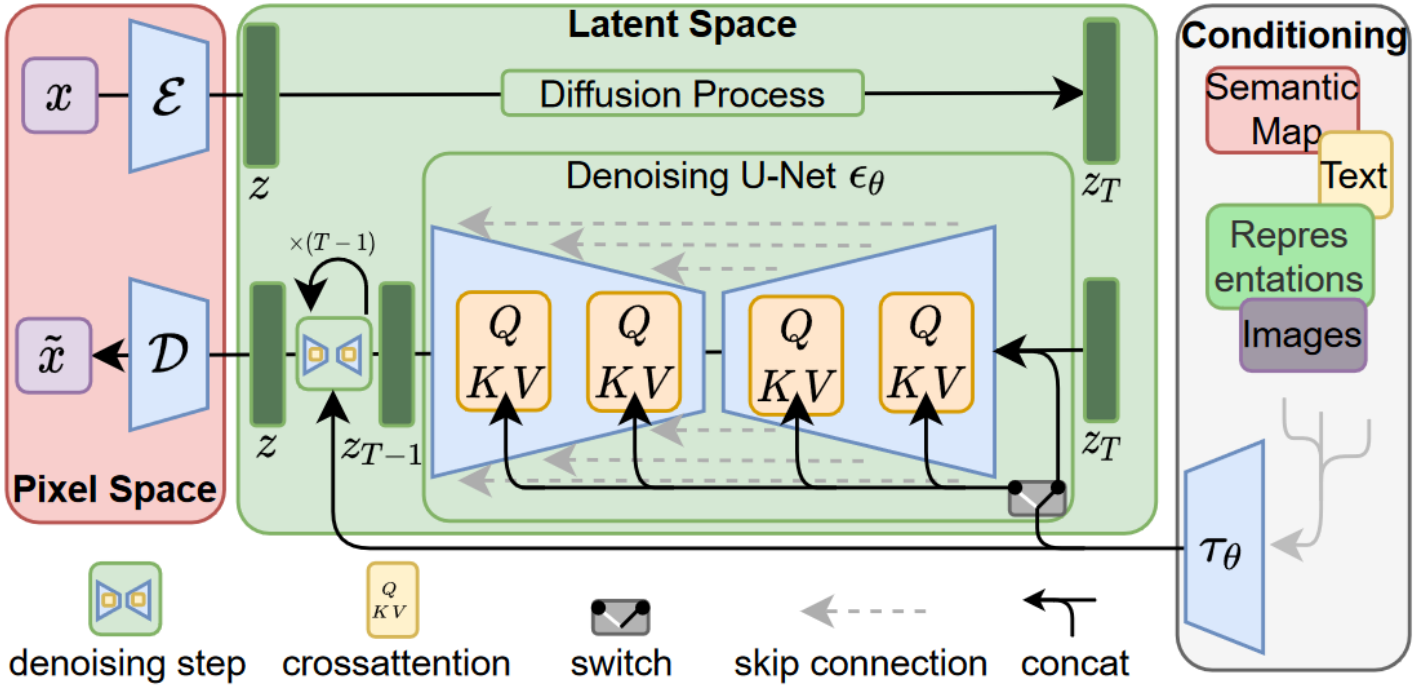


Image source: "High-Resolution Image Synthesis with Latent Diffusion Models." (Rombach et al., CVPR 2022)

Stable Diffusion

Evaluation Results:

- Comparison if different LDM configurations with recent state-of-the-art methods for image generation on the ImageNet dataset.

Method	FID↓	IS↑	Precision↑	Recall↑	N_{params}
SR3 [72]	11.30	-	-	-	625M
ImageBART [21]	21.19	-	-	-	3.5B
ImageBART [21]	7.44	-	-	-	3.5B
VQGAN+T [23]	17.04	70.6±1.8	-	-	1.3B
VQGAN+T [23]	5.88	**304.8**±3.6	-	-	1.3B
BigGan-deep [3]	6.95	203.6±2.6	**0.87**	0.28	340M
ADM [15]	10.94	100.98	0.69	**0.63**	554M
ADM-G [15]	4.59	186.7	0.82	0.52	608M
ADM-G,ADM-U [15]	3.85	221.72	0.84	0.53	n/a
CDM [31]	4.88	158.71±2.26	-	-	n/a
LDM-8 (ours)	17.41	72.92±2.6	0.65	0.62	395M
LDM-8-G (ours)	8.11	190.43±2.60	0.83	0.36	506M
LDM-8 (ours)	15.51	79.03±1.03	0.65	**0.63**	395M
LDM-8-G (ours)	7.76	209.52±4.24	0.84	0.35	506M
LDM-4 (ours)	10.56	103.49±1.24	0.71	0.62	400M
LDM-4-G (ours)	3.95	178.22±2.43	0.81	0.55	400M
LDM-4-G (ours)	**3.60**	247.67±5.59	**0.87**	0.48	400M

2 metrics for evaluating the quality of a generated image:

- **FID:** Fréchet Inception Distance
- **IS:** Inception Score

Source: "High-Resolution Image Synthesis with Latent Diffusion Models." (Rombach et al., CVPR 2022)

Stable Diffusion

Limitations:

- LDMs (Latent Diffusion Models) have reduced computational requirements compared to pixel-based approaches, but their sequential sampling process is slower than GANs.
- LDMs may not be suitable for tasks requiring high precision, as their reconstruction capability can become a bottleneck for tasks that require fine-grained accuracy in pixel space, such as image super-resolution.

PaLI:
Pathways Language and Image model

What does PaLI do?

- PaLI is a multilingual language-image model.
- It generates text based on visual and textual inputs and can perform vision, language and multimodal tasks in many languages (visual question-answering, image captioning, scene-text understanding, image classification).

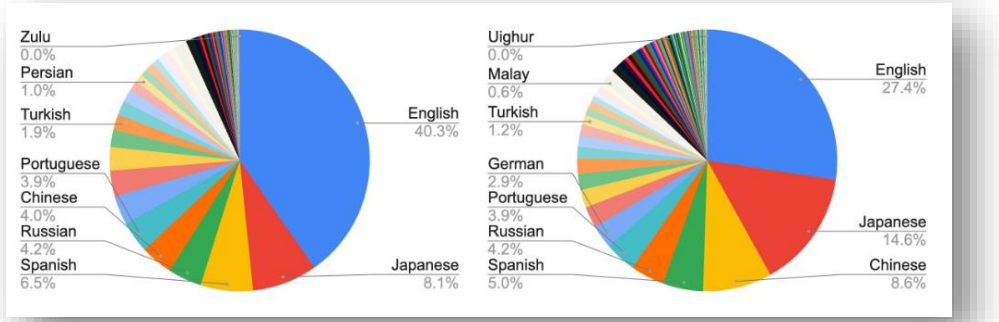
Advantages of PaLI:

- PaLI achieves SOTA in multiple vision and language tasks, while maintaining a simple, modular and scalable design.
- It is a multilingual and multimodal model that scales both across tasks and across languages.

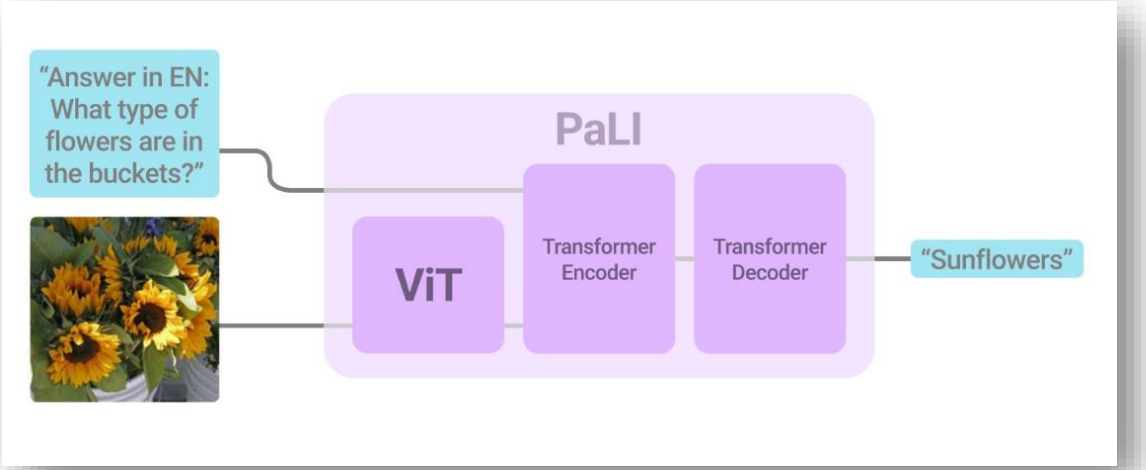
PaLI: Pathways Language and Image model

How does PaLI work?

- PaLI is trained on the WebLI dataset consisting of 10B pairs of image-text examples from the web, containing text in 109 languages.
- PaLI uses an “image-and-text to text” architecture which is used both during pretraining and fine-tuning and is general enough to be used for different tasks (VQA, Image Captioning, Scene-Text Understanding)
- **Visual Component:** ViT-e, a pretrained vision transformer scaled up to 4B parameters, the largest vanilla ViT to that date.
- **Language Component:** pretrained mT5-XXL from which the PaLI language encoder-decoder is initialized.
- The task to be performed is indicated in the prompt.



Statistics of recognized languages in OCR/alt-text in the WebLI dataset



PaLI model architecture

Images source: Pali: A jointly-scaled multilingual language-image model (Chen et al., 2022)

PaLI: Pathways Language and Image model

Why is PaLI so successful?

- **Amount of data:**
  Trained on 10B image-text pairs (compared to CLIPs 400M pairs).
- **Pretraining Task Mixture:**
  Designed to handle a wide range of tasks in the image-language space (captioning, OCR, English and cross-lingual VQA, English and cross-lingual Visual Question Generation, object detection).
- **Joint scaling of the vision and language components:**
  While the performance gains of scaling vision models tend to saturate in classification tasks, scaling combined vision and language models in PaLI results in significant improvements in vision-language tasks.

PaLI: Pathways Language and Image model

Evaluation Results:

Image Captioning: (CIDEr score)

	COCO	NoCaps		TextCaps		VizWiz-Cap	
Model	Karpathy-test	val	test	val	test	test-dev	test-std
LEMON (0.7B)	139.1	117.3	114.3	-	-	-	-
SimVLM	143.3	112.2	110.3	-	-	-	-
CoCa (2.1B)	143.6	122.4	120.6	-	-	-	-
GIT (0.7B)	144.8	125.5	123.4	143.7	138.2	113.1	114.4
GIT2 (5.1B)	145.0	**126.9**	**124.8**	148.6	145.0	119.4	120.8
OFA (0.9B)	145.3	-	-	-	-	-	-
Flamingo (80B)	138.1	-	-	-	-	-	-
BEiT-3 (1.9B)	147.6	-	-	-	-	-	-
PaLI-3B	145.4	121.1	-	143.6	-	117.2	-
PaLI-15B	146.2	121.2	-	150.1	-	121.7	-
PaLI-17B	**149.1**	**127.0**	**124.4**	**160.0**	**160.4**	**123.0**	**124.7**

Visual Question Answering: (accuracy)

	VQAv2		OKVQA	TextVQA		VizWiz-QA		ST-VQA	
Method	test-dev	test-std	val	val	test	test-dev	test	val	test
SimVLM	80.03	80.34	-	-	-	-	-	-	-
CoCa (2.1B)	82.3	82.3	-	-	-	-	-	-	-
GIT (0.7B)	78.56	78.81	-	59.93	59.75	68.0	67.5	69.1	69.6
GIT2 (5.1B)	81.74	81.92	-	68.38	67.27	70.97	70.1	75.1	75.8
OFA (0.9B)	82.0	82.0	-	-	-	-	-	-	-
Flamingo (80B)	82.0	82.1	57.8*	57.1	54.1	65.7	65.4	-	-
BEiT-3 (1.9B)	84.2	84.0	-	-	-	-	-	-	-
KAT	-	-	54.4	-	-	-	-	-	-
Mia	-	-	-	-	73.67†	-	-	-	-
PaLI-3B	81.4	-	52.4	60.12	-	67.5	-	67.5	69.7
PaLI-15B	82.9	-	56.5	65.49	-	71.1	-	73.2	76.5
PaLI-17B	**84.3**	**84.3**	**64.5**	**71.81**	73.06	**74.4**	**73.3**	**77.1**	**79.9**

Zero-Shot Image Classification: (accuracy)

Model (ImageNet data)	INet	INet-R	INet-A	INet-Sketch	INet-v2	ObjNet
Flamingo-80B (**1-shot**)	71.9	-	-	-	-	-
Flamingo-80B (**5-shot**)	77.3	-	-	-	-	-
PaLI-3B (**0-shot**)	70.06	80.15	37.92	61.11	62.55	38.87
PaLI-15B (**0-shot**)	70.27	81.21	41.16	61.03	62.81	39.51
PaLI-17B (**0-shot**)	**72.11**	**81.97**	**44.70**	**63.83**	**64.46**	**42.62**

Images source: Pali: A jointly-scaled multilingual language-image model (Chen et al., 2022)

PaLI:
Pathways Language and Image model

Limitations:

- Complex Scene Description:
  Limited complex annotations in the source data may result in less thorough descriptions of scenes with multiple objects in VQA.
- Loss of Multilingual Capabilities:
  Fine-tuning on English-only data can lead to a loss of multilingual capabilities -> Need of fine-tuning with a mix of multilingual datasets.

Conclusion

Multimodal AI – a paradigm shift?

- Promising approach for both research and business applications.
- Boundaries between modalities are becoming more fuzzy.
- There is a growing trend of moving beyond one data modality and leverage relationships between different modalities.
- Ad-hoc solutions where a model is designed to solve a single task on a single modality may be gradually replaced by multimodal models in the future.

References

- Letitia Parcalabescu, Nils Trost, and Anette Frank. 2021. What is Multimodality?. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 1–10, Groningen, Netherlands (Online). Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger and Ilya Sutskever. 2021 Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning*.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer. "High-Resolution Image Synthesis with Latent Diffusion Models." *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021): 10674-10685.*
- Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, ... & Soricut, R. (2022). Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database" *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248-255.
- Jay Alammar, The Illustrated Stable Diffusion [Blog Post], https://jalammar.github.io/illustrated-stable-diffusion/
- Han Xiao, The Paradigm Shift Towards Multimodal AI, https://jina.ai/news/paradigm-shift-towards-multimodal-ai/

Questions?