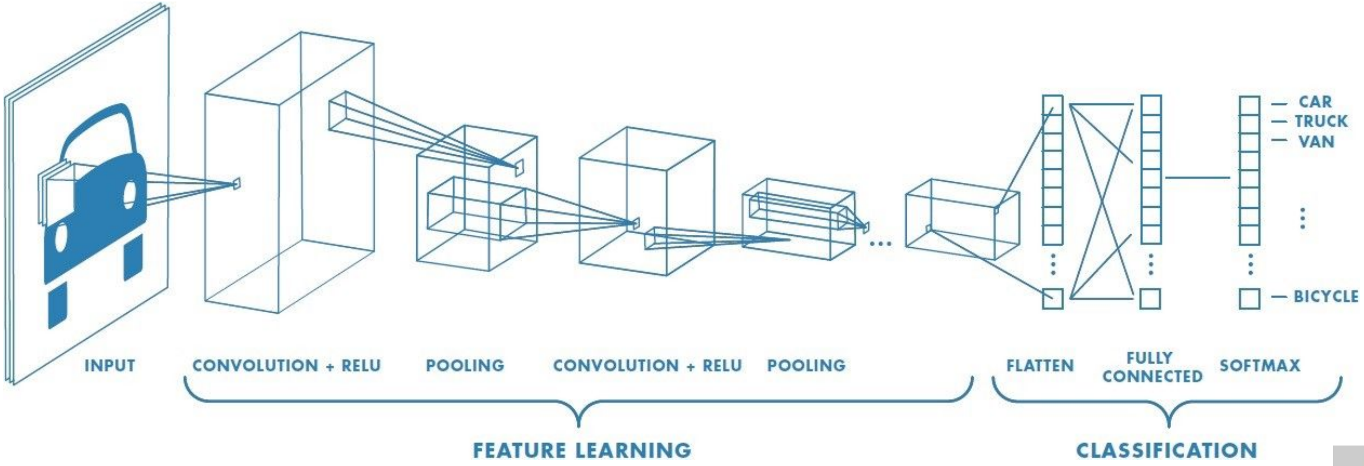
Analyzing Comprehension of Spatial Relations in Joint Text-Image Models

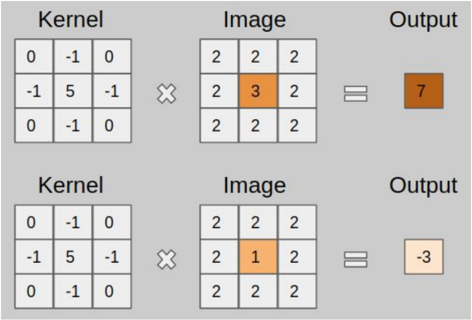
Andrew Briand Chris Haberland David Krug

Outline

- Background
  - Text-Image Models and Tasks
  - Vision Transformers
- CLIP
  - Introduction
  - Using CLIP in Energy Based Models
- DALL-E
  - Introduction
  - Spatial Understanding
- Our experiments

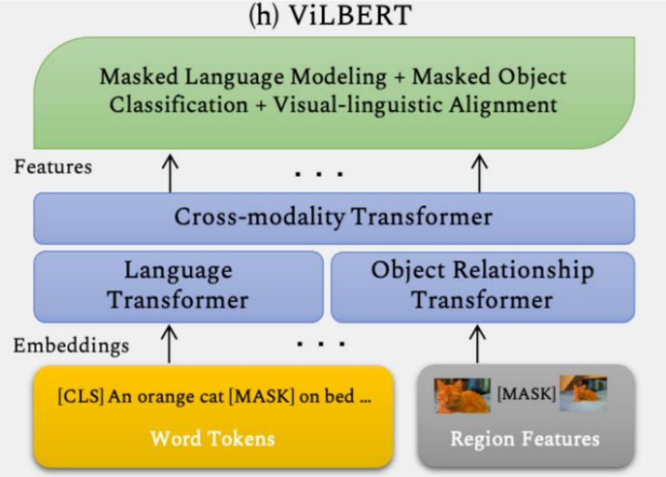
Convolutional Neural Networks





Coupled Models - Multi-stream

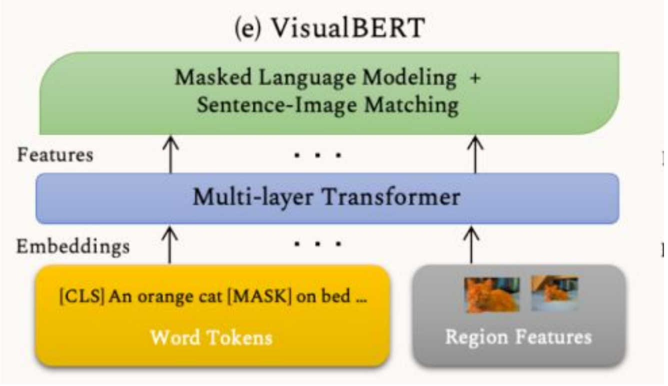
- ViLBERT: Earliest example of applying BERT to multi-model data
  - Language input consists of embedded tokens
  - Image inputs are bounding boxes and their extracted features from object detection model
    - 5-d spatial encoding including top-left, bottom-right, and proportion of image covered.
  - Cross-modality transformer takes queries from one stream and applies them to keys and values from the other
  - Pre-training tasks:
    - Masked inputs (both image and text)
    - Alignment prediction (binary: does the caption describe the image?)



Source: [Transformers in Vision: A Survey]()

Coupled Models - Single Stream

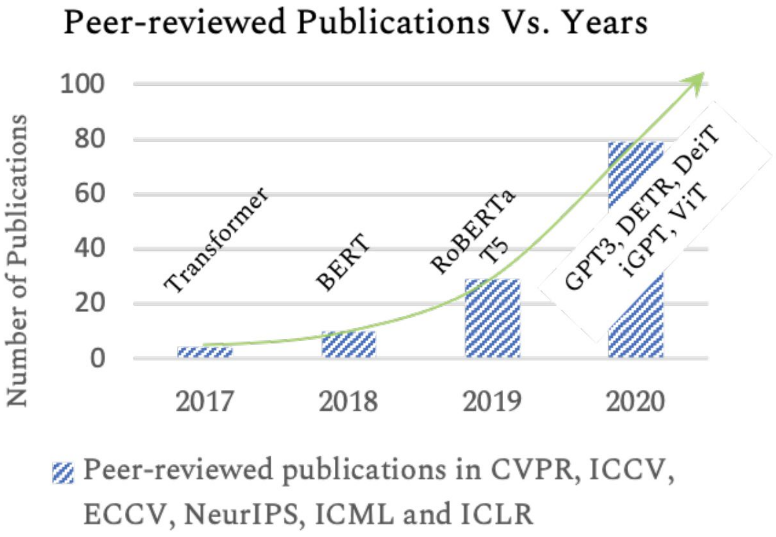
- VisualBERT
  - Word tokens input to language transformer
  - Regions extracted from object detector
    - Sum of embedding from CNN, segment embedding indicating it is an image, and positional embeddings from corresponding words when that info is available
  - All fed into single, multi-layer transformer
  - Pretraining
    - MLM with image
    - Given two captions, determine if both are good or if one good and one bad



Source: [Transformers in Vision: A Survey]()

Background: Transformers

- Transformers begin to be applied for vision - 2020



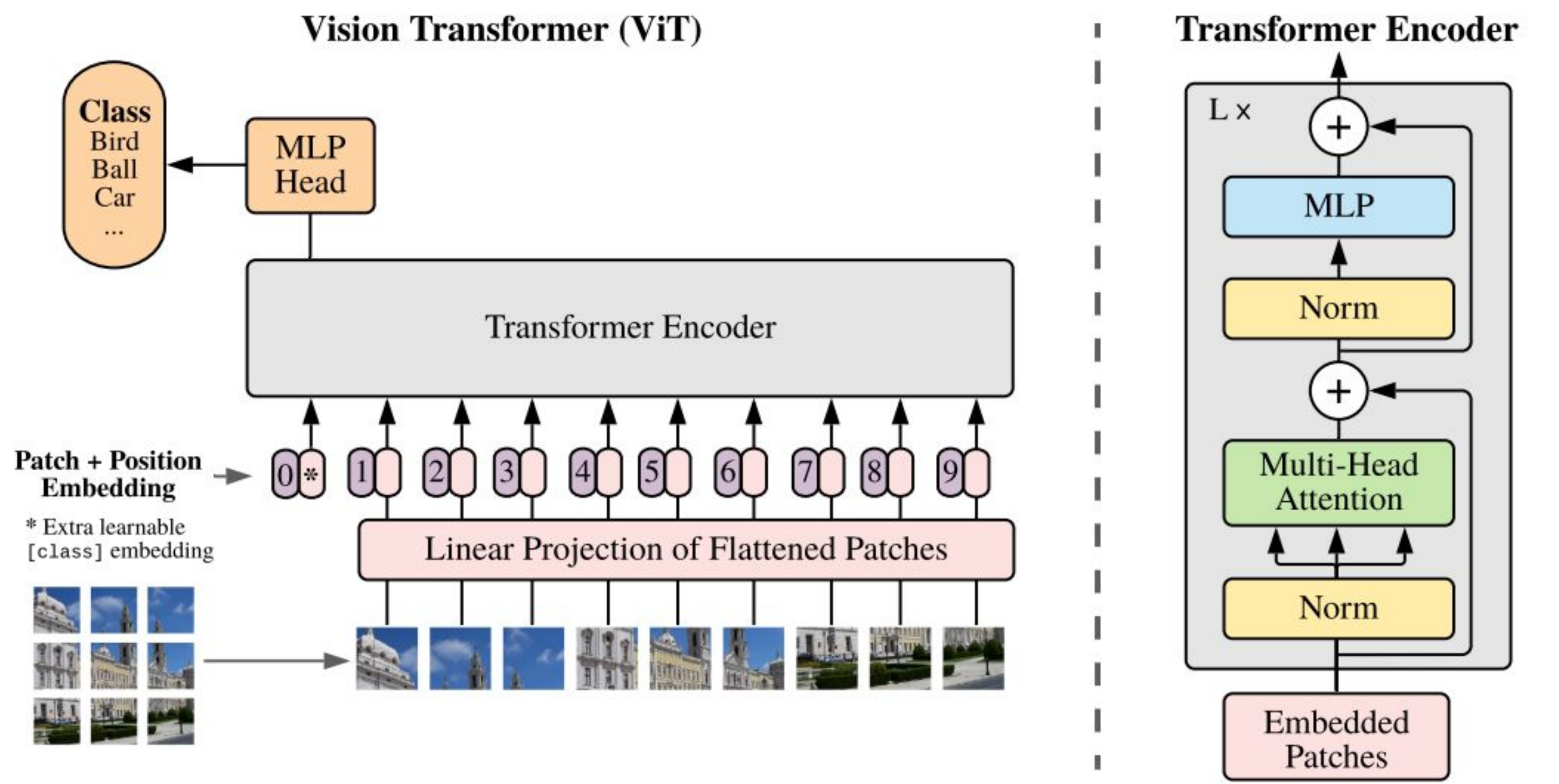
Source: [Transformers in Vision: A Survey]

Vision Transformer Input

- ViT
  - Input is a sequence of projections (learnable) of flattened image patches with positional embeddings
    - Embeddings are learned 1D embeddings, 2D positionally aware embeddings offered no performance improvement
  - Supervised pre-training on large classification datasets like JFT
- Can also be done with image features taken from a CNN
  - Seems from the paper that smaller transformers did better in such a hybrid setup but larger ones did not

Source: An Image is Worth 16x16 Words

Image Encodings: Vision Transformer



Vision Transformer Pretraining

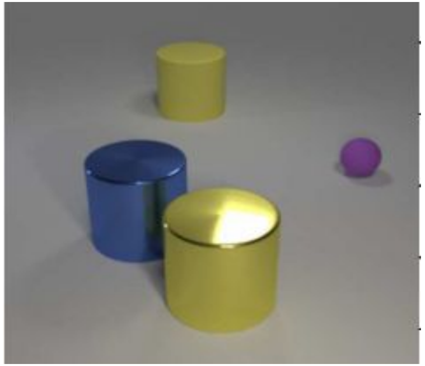
- Can be analogous to MLM:
  - Predict parts of image
  - Color parts of image
  - Predict rotation of image
- Contrastive training

Image-Text Tasks

- Object recognition
- Facial emotion recognition
- Hateful meme recognition
- Geolocation
- Distance to objects
- [ImageNet]

OCR: 158

Counting: Four



Action classification: Line dancing

Object classification: Motorcycle

Contrastive Language-Image for Pretraining (CLIP)

Learning Transferable Visual Models From Natural Language Supervision

---

Alec Radford [*1] **Jong Wook Kim** [*1] **Chris Hallacy** [1] **Aditya Ramesh** [1] **Gabriel Goh** [1] **Sandhini Agarwal** [1]
Girish Sastry [1] **Amanda Askell** [1] **Pamela Mishkin** [1] **Jack Clark** [1] **Gretchen Krueger** [1] **Ilya Sutskever** [1]

Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task

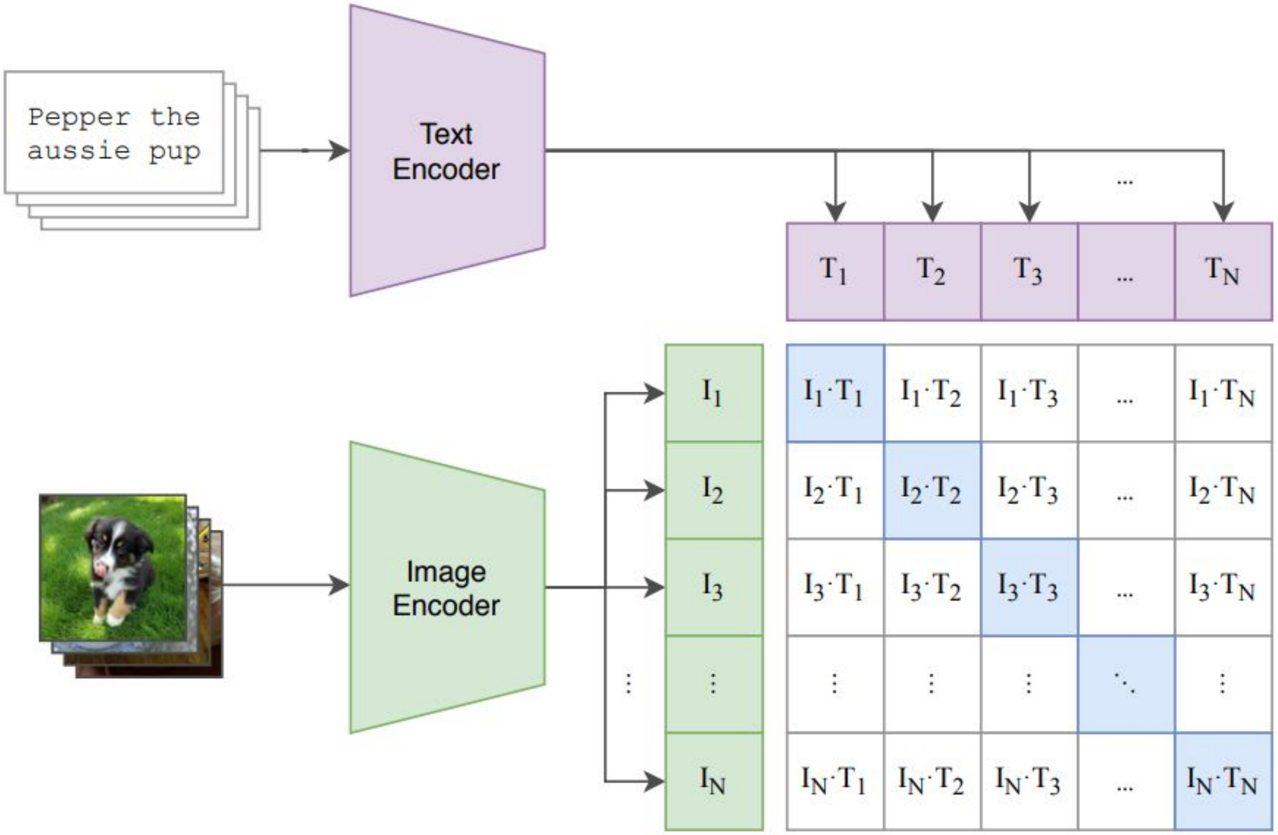
Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of "text-to-text" as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets removing the need for specialized output heads or dataset specific customization. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset

Source

CLIP: Overview

- From OpenAI in early 2021
- Task: pair images with captions
  - Uses natural language for zero-shot transfer to downstream tasks
  - Scalable and efficient task for learning image-text embeddings
- Two-stream architecture: image encoder separate from text encoder
- Data: 400 million (image, text) pairs collected from the internet
- Results: performs surprisingly well given no specific training data

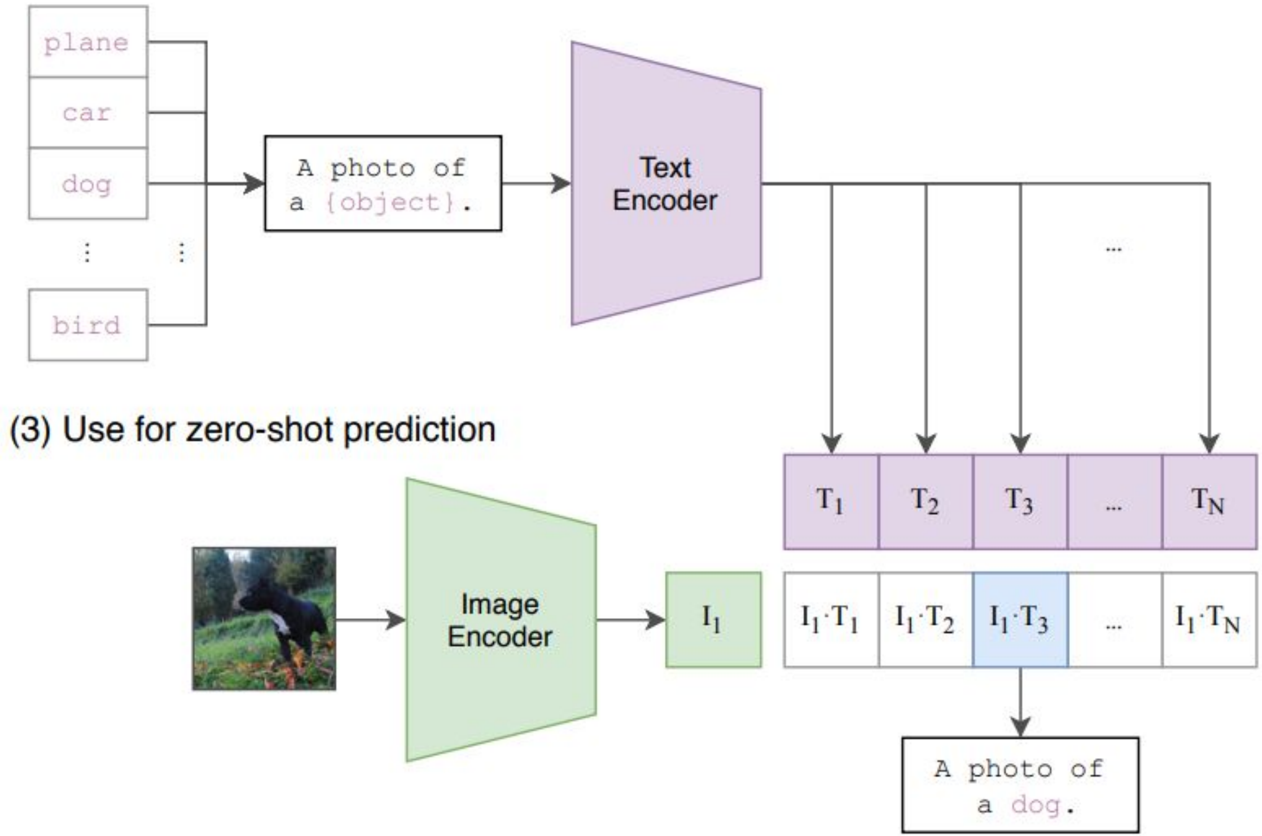
CLIP: Architecture - Training



CLIP: Text Encoder Details

- transformer
- 63M parameters
- 12 layers deep
- 512-dimensional representations
- 8 attention heads
- lower-cased byte pair encoding with 50k vocabulary size

CLIP: Architecture - Zero Shot Inference



CLIP: Zero Shot Inference - Prompt Engineering

Use natural language rather than one word

- “cat” becomes “A photo of a cat.”

Specify general category

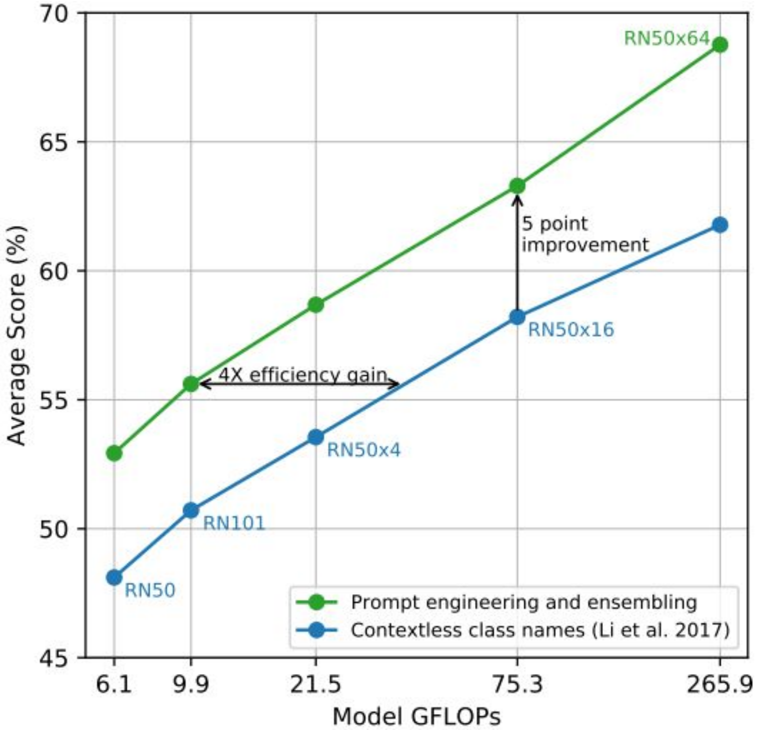
- Food101 : A photo of a {label}, a type of food

Specify unusual photo angles

- Satellite imagery: A satellite image of {label}

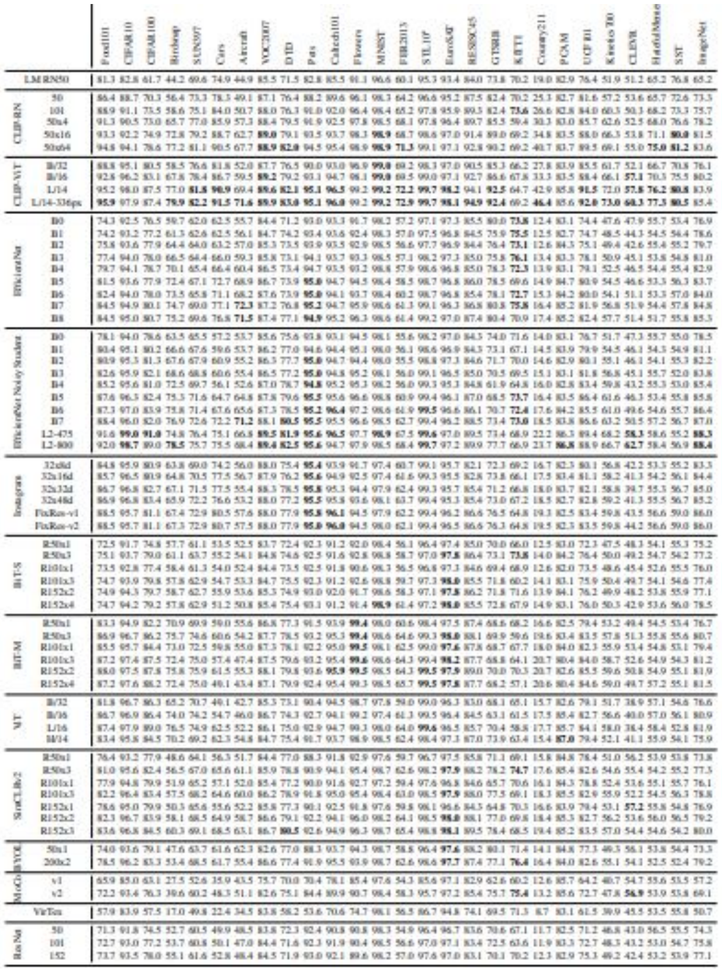
Ensemble with different contexts

- A photo of a big {label}
- A photo of a small {label}



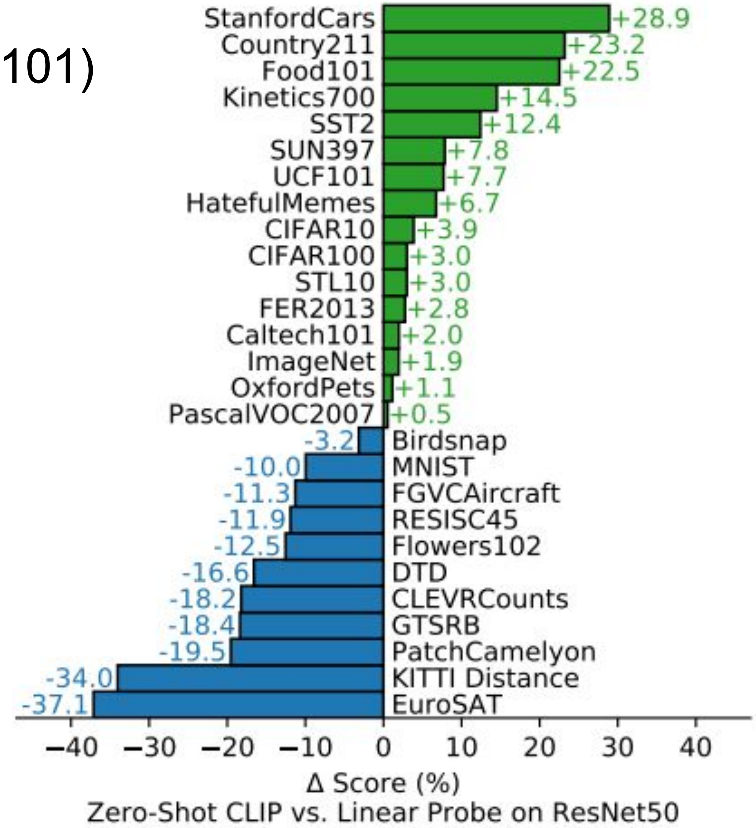
CLIP: Results

- For 16 out of 27 dataset evaluation suites, CLIP's zero-shot outperforms a supervised linear classifier fit on ResNet-50 features

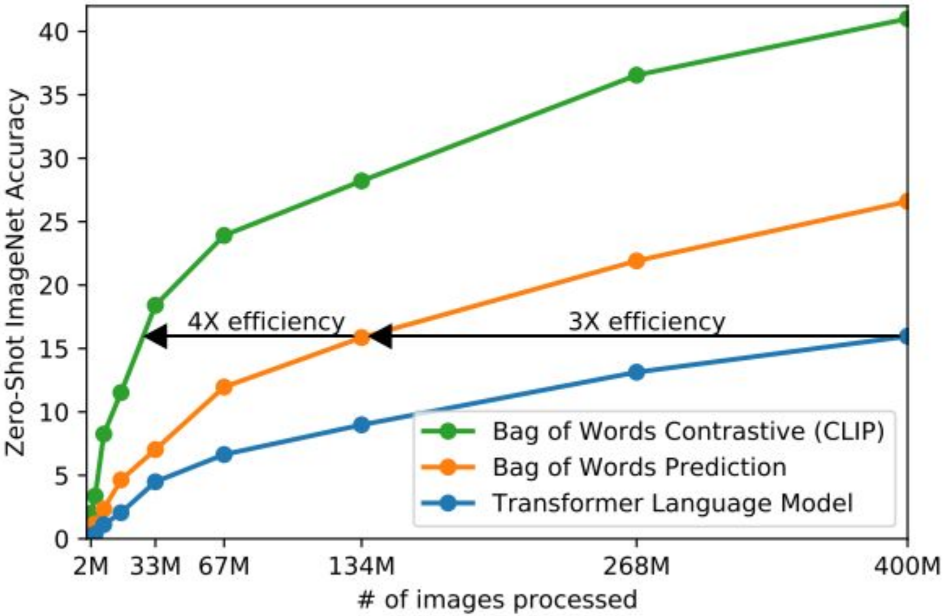


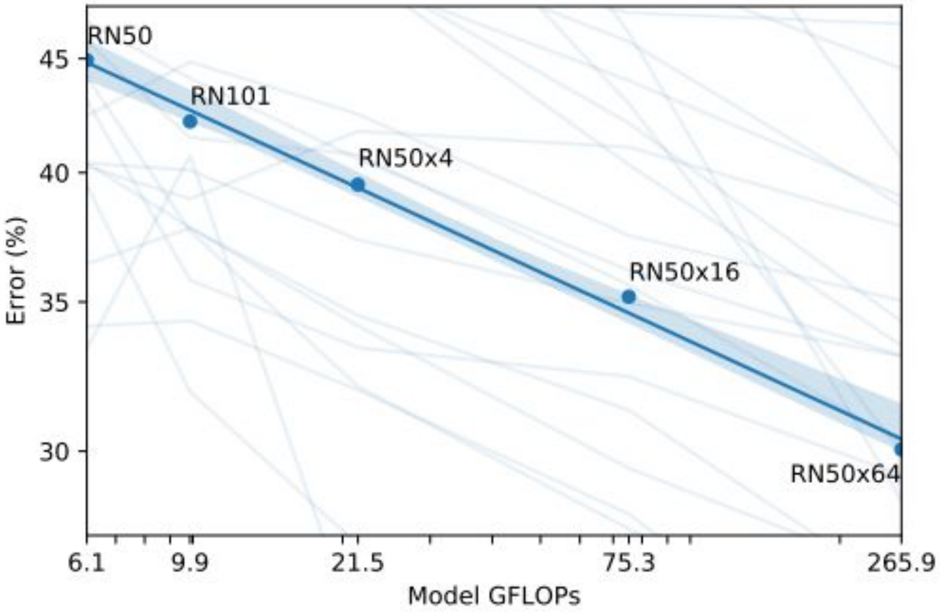
CLIP: Results cont.

- Good at action recognition (Kinetics700, UFC101)
    - possibly because NL offers more verbs in training data
- Bad at complex/abstract tasks:
    - classifying satellite imagery, german traffic signs, lymph node tumors
    - counting
    - measuring distance (to nearest car)
- Should we expect good zero-shot performance at detecting tumors?

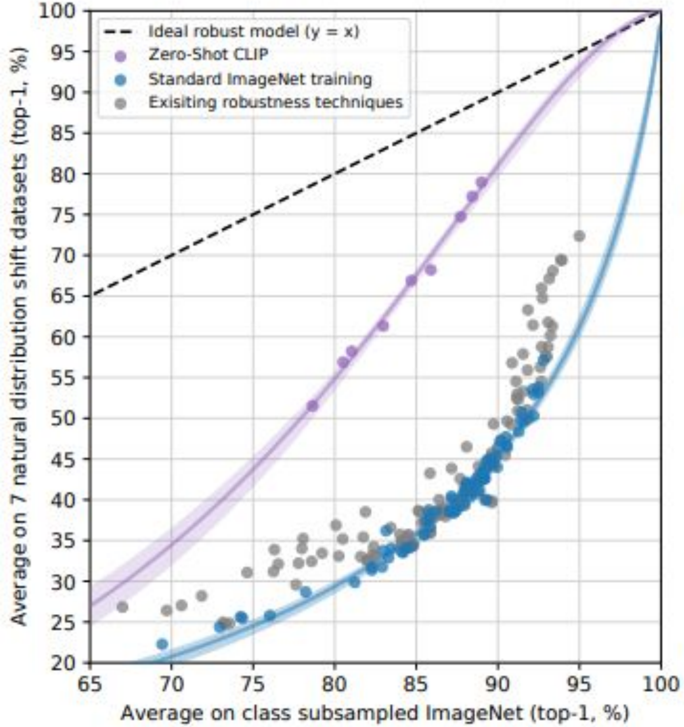


CLIP: Results - Scalability





CLIP: Robustness to Distributional Shift

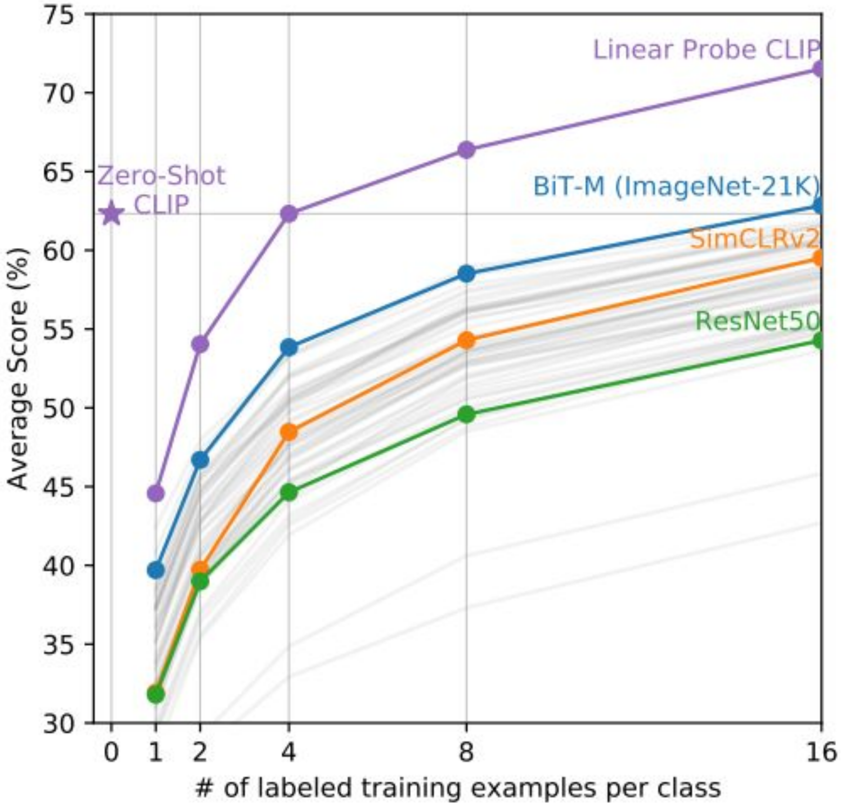


CLIP does not learn like a human

- 5 humans each classified 3600 images of dogs and cats
  - 37 dog/cat breed plus “I don’t know”
- Human performance increased a lot with one training example per class
- Most performance increase happened where human confidence was low
- Humans “know what the don’t know” and “update their priors”
- CLIP’s few-shot performance doesn’t “make effective use of prior knowledge”

CLIP: Linear Probing

- Logistic regression classifier on top of CLIP requires 4 labeled training examples match CLIP's performance



CLIP: Issues and Limitations

- Polysemy: crane (construction vs. bird), boxer (dog vs. athlete)
- Severe distributional shifts (handwritten OCR)
- Subject to biases found in training data
  - Given an image of a store, can CLIP identify potential thieves?
    - Black, young, and male people are misclassified most
  - Men are more likely to be labeled with a high status job

CLIP: Summary

- Main idea: CLIP uses natural language, which allow for scalability
- Using natural language provides robustness to distributional shifts
    - Still somewhat vulnerable to severe shifts
- More work need to be done combining prior (zero-shot) knowledge with new (one-shot) knowledge

Using CLIP in Energy Based Models

Learning to Compose Visual Relations

Nan Liu *
University of Michigan
liunan@umich.edu

Shuang Li *
MIT CSAIL
lishuang@mit.edu

Yilun Du *
MIT CSAIL
yilundu@mit.edu

Joshua B. Tenenbaum
MIT CSAIL, BCS, CBMM
jbt@mit.edu

Antonio Torralba
MIT CSAIL
torralba@mit.edu

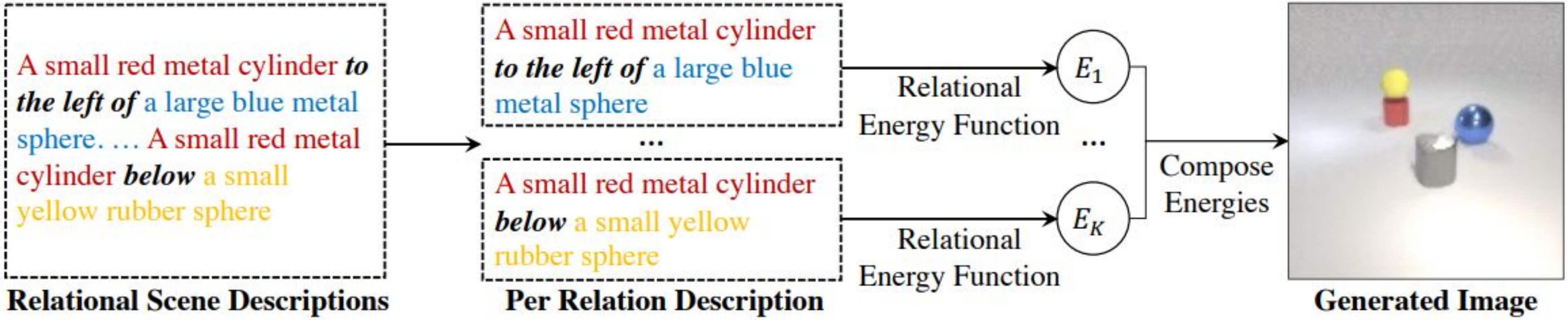
Source

Abstract

The visual world around us can be described as a structured set of objects and their associated relations. An image of a room may be conjured given only the description of the underlying objects and their associated relations. While there has been significant work on designing deep neural networks which may compose individual

Energy Based Models (EBMs)

- Using existing multi-modal models like CLIP and DALL-E is "naive"
  - Due to a "lack of *compositionality* in the language encoder"
- Their approach is to:
  - factorize the scene description wrt. each individual relation
  - separate EMBs encode each relation
  - the encodings are composed to produce the scene encoding



EBMs: How do they work?

- EBMs are “a class of unnormalized probability models”
- Parameterize a probability distribution over images $p_\theta(x)$ via a learned energy function E_θ

$$p_\theta(\mathbf{x}) \propto e^{-E_\theta(\mathbf{x})}$$

- Compose models

$$\prod_i p_\theta^i(\boldsymbol{x}) \propto e^{-\sum_i E_\theta^i(\boldsymbol{x})}$$

- The energy function depends on the specific relational model

$$p_\theta(\boldsymbol{x}; r_i) \propto e^{-E_\theta^i(\boldsymbol{x}|\mathrm{Enc}(r_i))}$$

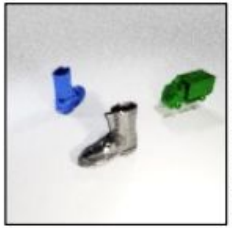
EBMs: Experiment

Query image	CLIP	Fine-tuned CLIP	Ours
	•A maple wood coffee table ***on the right of*** a gray fabric couch ✗ •A gray fabric couch ***on the left of*** a maple wood coffee table ✗ •A maple wood coffee table ***in front of*** a blue fabric stool ✗	•A maple wood coffee table ***on the left*** a gray fabric couch ✓ •A gray fabric couch ***behind*** a blue fabric stool ✗ •A blue fabric stool ***in front of*** a maple wood coffee table ✓	•A maple wood coffee table ***on the left of*** a gray fabric couch ✓ •A gray fabric couch ***on the right of*** a blue fabric stool ✓ •A blue fabric stool ***in front of*** a maple wood coffee table ✓

(a) Top 1 image-text retrieval result on iGibson scenes.

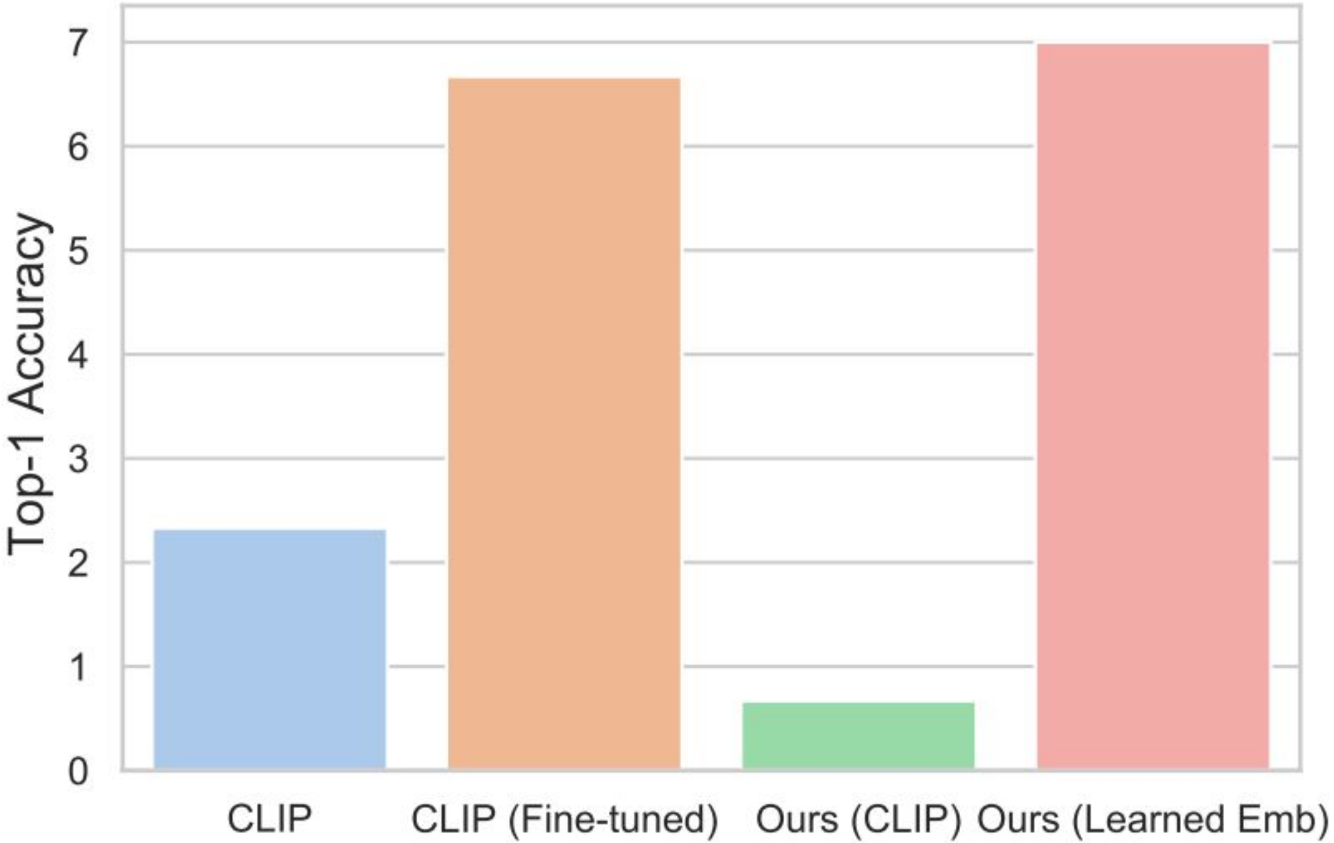
	•A large gray metal sphere ***on the left of*** a small red metal cube ✗ •A small red metal cube ***on the right of*** a large brown metal cube ✗ •A large brown metal cube ***below*** a large green rubber cylinder ✓	•A large gray metal sphere ***above*** a small red metal cube ✓ •A small red metal cube ***behind*** a large brown metal cube ✓ •A large brown metal cube ***below*** a large green rubber cylinder ✓	•A large gray metal sphere ***above*** a small red metal cube ✓ •A small red metal cube ***on the left of*** a large brown metal cube ✓ •A large brown metal cube ***below*** a large green rubber cylinder ✓

(b) Top 1 image-text retrieval result on CLEVR scenes.

	•A blue object ***in front of*** a gray object ✗ •A gray object ***on the left of*** a green object ✓ •A green object ***behind*** a blue object ✗	•A blue object ***in front of*** a gray object ✗ •A gray object ***behind*** a green object ✗ •A green object ***on the left of*** a blue object ✗	•A blue object ***behind*** a gray object ✓ •A gray object ***on the left of*** a green object ✓ •A green object ***on the right of*** a gray object ✓

(c) Top 1 image-text retrieval result on Blender scenes (outside the training distribution).

EBM: Results



Top-1 Accuracy
0
1
2
3
4
5
6
7
CLIP
CLIP (Fine-tuned)
Ours (CLIP)
Ours (Learned Emb)

DALL-E

Zero-Shot Text-to-Image Generation

Aditya Ramesh [1] Mikhail Pavlov [1] Gabriel Goh [1] Scott Gray [1]
Chelsea Voss [1] Alec Radford [1] Mark Chen [1] Ilya Sutskever [1]

Abstract

Text-to-image generation has traditionally focused on finding better modeling assumptions for training on a fixed dataset. These assumptions might involve complex architectures, auxiliary losses, or side information such as object part labels or segmentation masks supplied during training. We describe a simple approach for this task based on a transformer that autoregressively models the text and image tokens as a single stream of data. With sufficient data and scale, our approach is competitive with previous domain-specific models when evaluated in a zero-shot fashion.

[Source]

Generative models

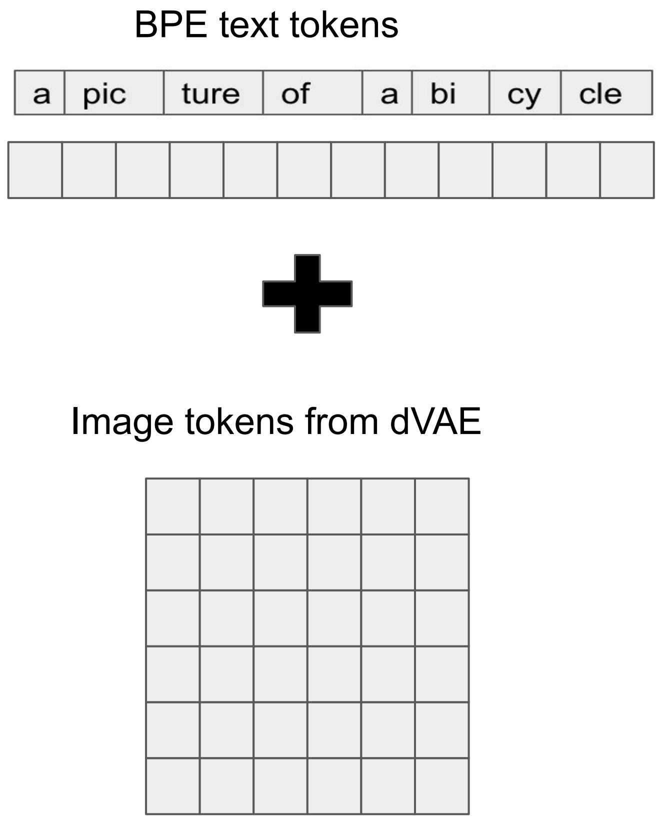
- Can we get an image from text alone based off of dual-modal data?
- End goal is to learn the conditional distribution of images given some string of text;

- Examples:
    - Dall-E
    - GLIDE
    - Dall-E 2 (uses diffusion model as generator)

Dall-E paper

- **Stage 1.** We train a discrete variational autoencoder (dVAE)[1] to compress each 256×256 RGB image into a 32×32 grid of image tokens, each element of which can assume 8192 possible values. This reduces the context size of the transformer by a factor of 192 without a large degradation in visual quality (see Fig-

- **Stage 2.** We concatenate up to 256 BPE-encoded text tokens with the $32 \times 32 = 1024$ image tokens, and train an autoregressive transformer to model the joint distribution over the text and image tokens.

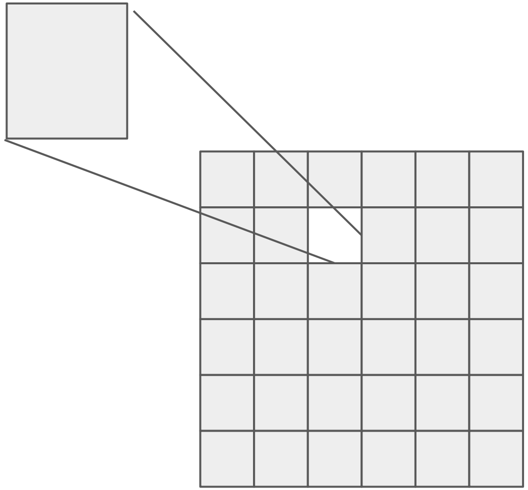


The text and image tokens are concatenated and modeled as a single stream of data.

What’s an image token

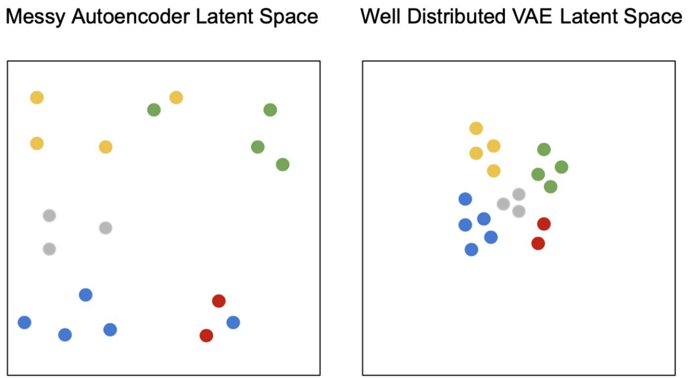
“...encode the image using 32 × 32 = 1024 tokens with vocabulary size 8192. The image tokens are obtained using argmax sampling from the dVAE encoder logits”

The image tokens are the mappings of regions of the image produced by the dVAE model



dVAE z(x) of image

VAE



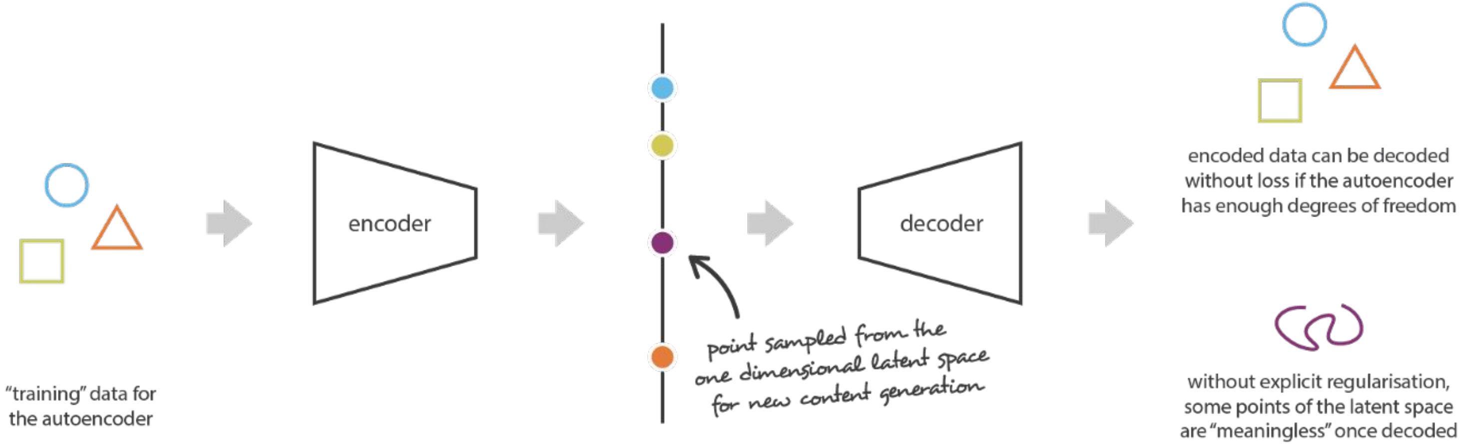
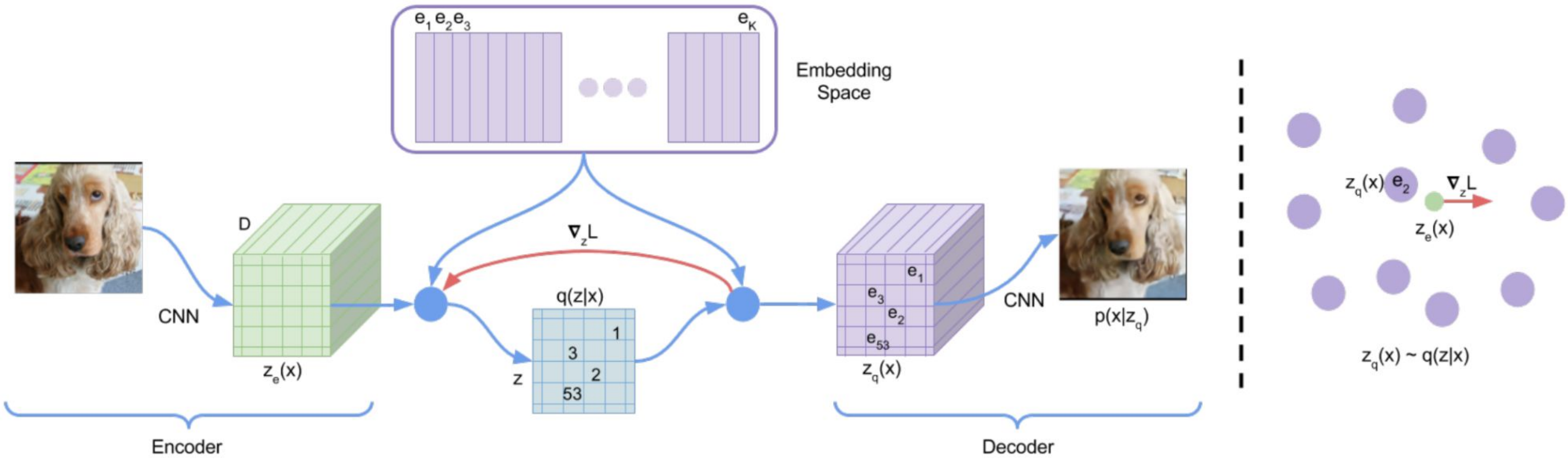


Image source: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

VQ-VAE

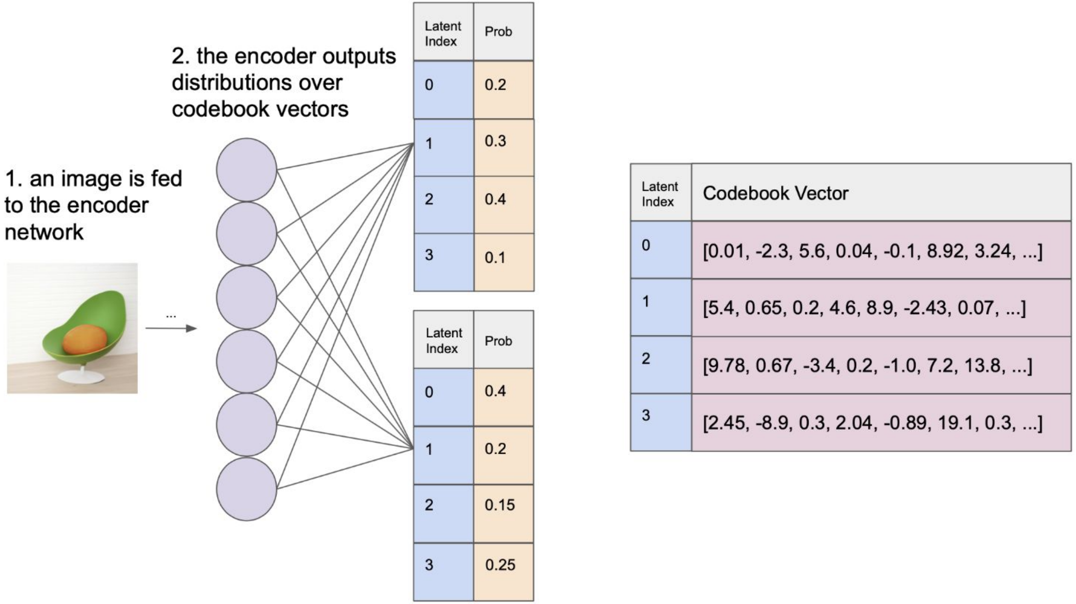


$$\hat{x} = g(f(x))$$

[Neural Discrete Representation Learning]

dVAE

- Main difference - dVAE encoder outputs a distribution over codebook vectors for each latent



Source: https://ml.berkeley.edu/blog/posts/dalle2/

Step 2: Autoregressive transformer on image and text

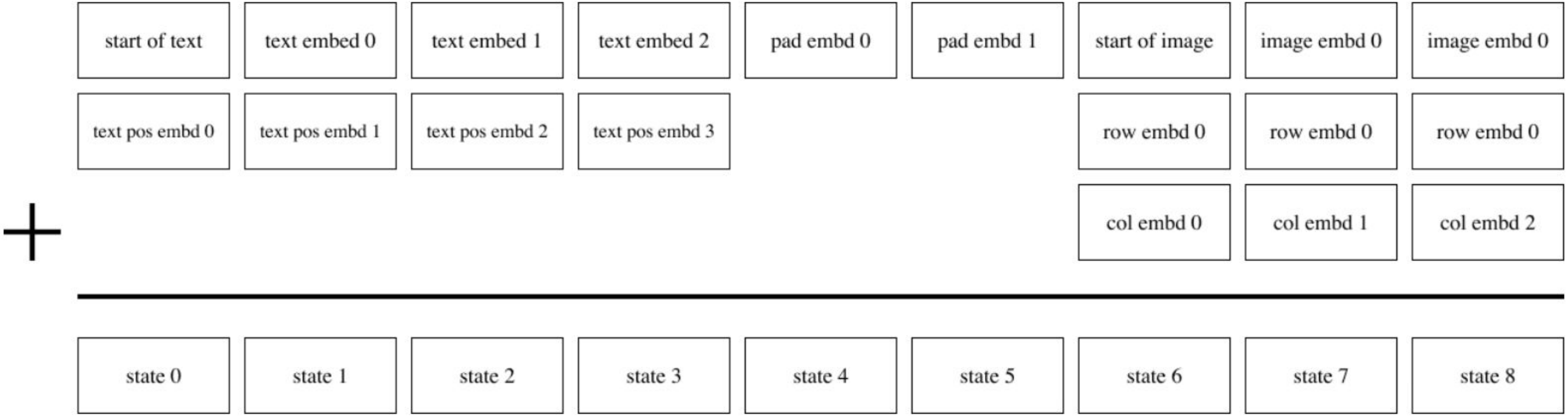
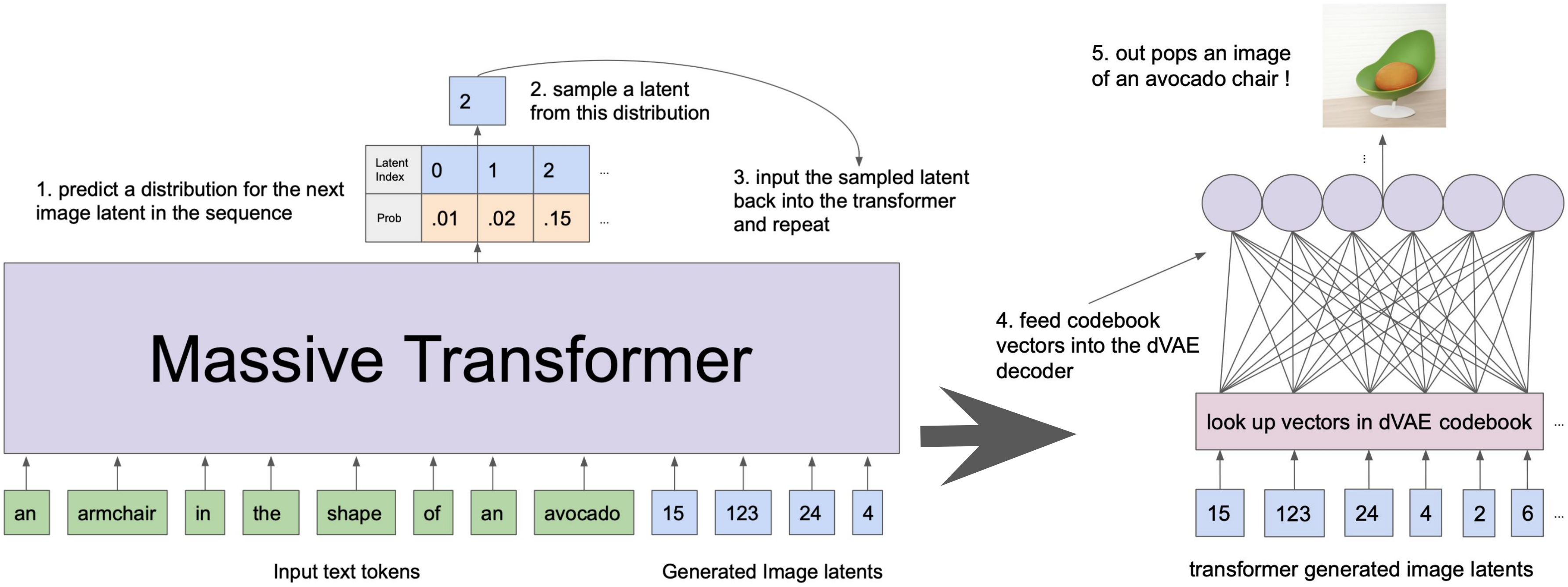


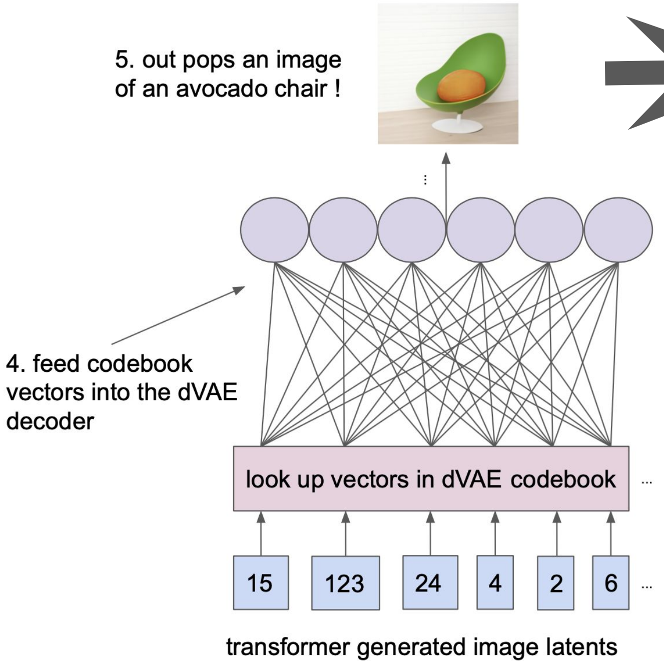
Figure 10. Illustration of the embedding scheme for a hypothetical version of our transformer with a maximum text length of 6 tokens. Each box denotes a vector of size $d_{\text{model}} = 3968$. In this illustration, the caption has a length of 4 tokens, so 2 padding tokens are used (as described in Section 2.2). Each image vocabulary embedding is summed with a row and column embedding.

Step 2: Autoregressive transformer on image and text

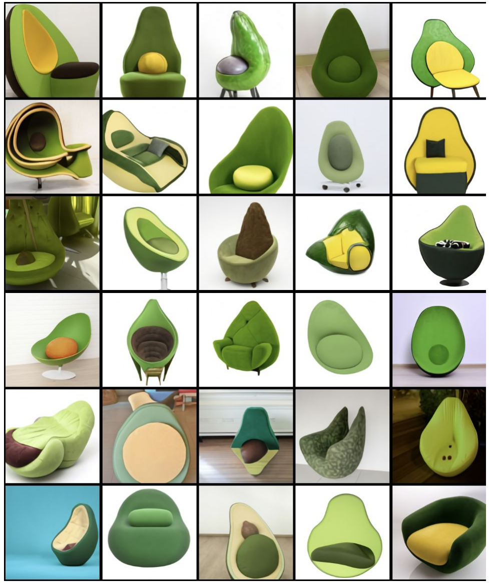


Source: https://ml.berkeley.edu/blog/posts/dalle2/

Step 3: Rerank results



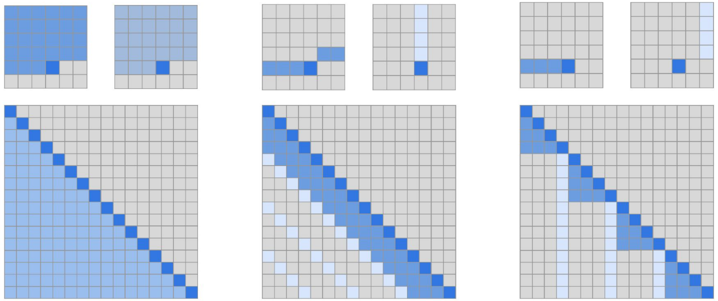
Rerank with CLIP



Source: https://ml.berkeley.edu/blog/posts/dalle2/

Training and engineering

- Same dataset as CLIP
- data augmentation to the images before encoding them
- 10% BPE dropout
- Axial attention - get around O(n^2) attention mask computation
- Gradient compression
- 1024, 16 GB NVIDIA V100 GPUs



(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)

Dall-E objective

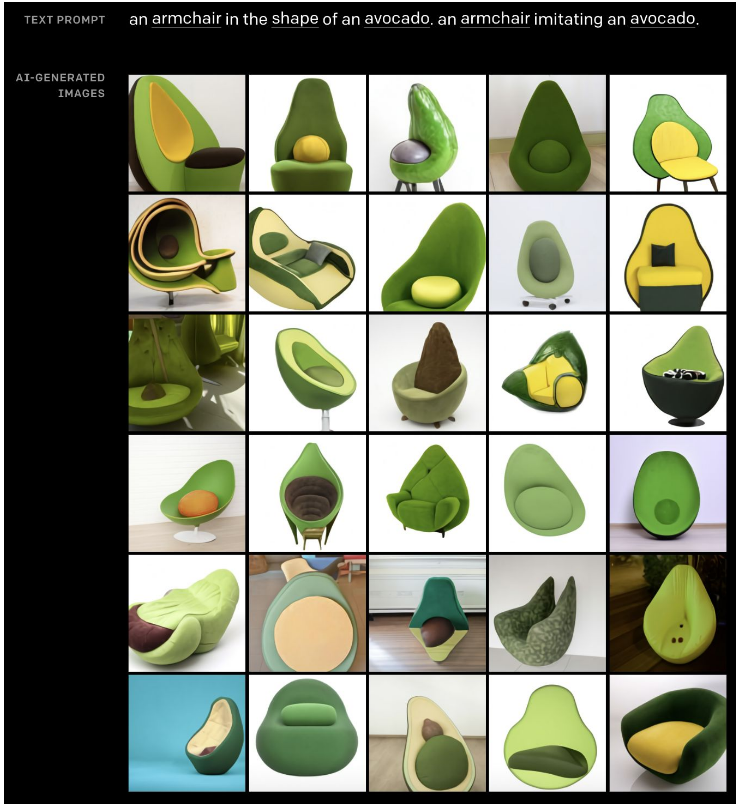
$$p_{\theta,\psi}(x, y, z) = p_\theta(x \mid y, z) p_\psi(y, z)$$

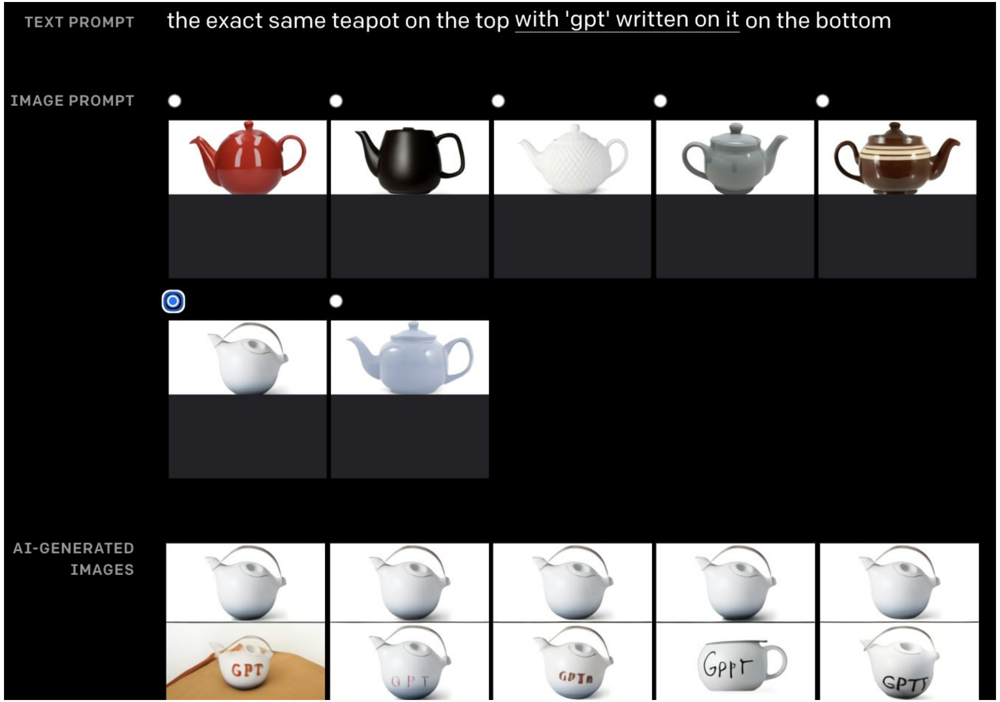
$$\ln p_{\theta,\psi}(x, y) \geqslant \mathop{\mathbb{E}}_{z \sim q_\phi(z \mid x)} \Big(\ln p_\theta(x \mid y, z) - \beta \, D_{\mathrm{KL}}(q_\phi(y, z \mid x), p_\psi(y, z)) \Big), \quad (1)$$

where:

- q_ϕ denotes the distribution over the 32×32 image tokens generated by the dVAE encoder given the RGB image x[2];
- p_θ denotes the distribution over the RGB images generated by the dVAE decoder given the image tokens; and
- p_ψ denotes the joint distribution over the text and image tokens modeled by the transformer.

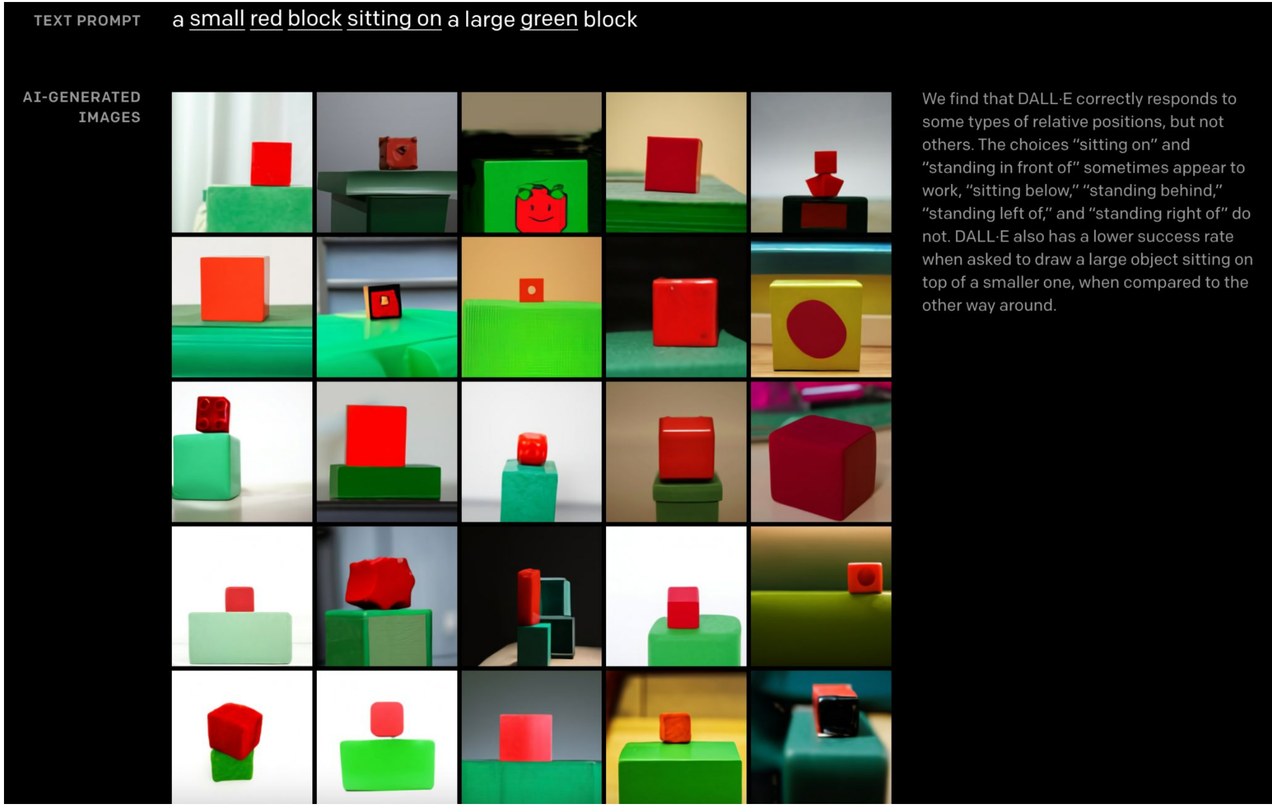
Cool things Dall-E can generate





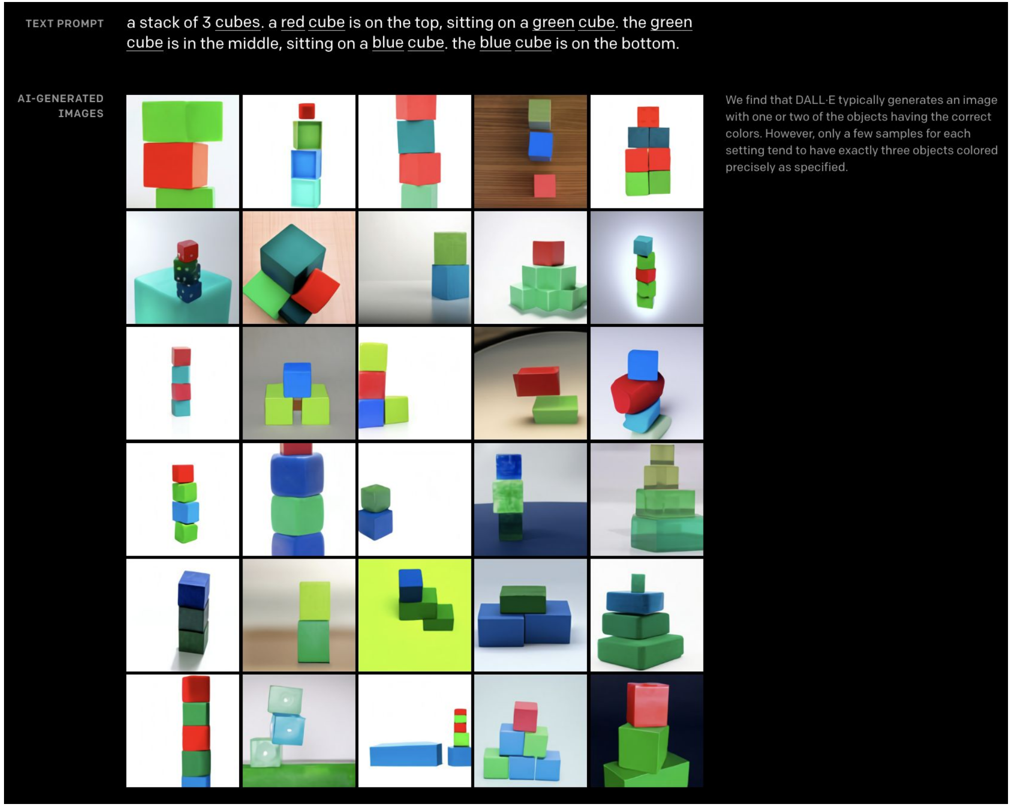
Spatial understanding - Dall-E is imperfect

- Adjectives to correct object (large, small)
- Inversion of relation
- Misidentification of relation
- Mischaracterization of arguments (blocks appear as weird shapes)



Spatial understanding - Dall-E is imperfect

- Adjectives to correct object (large, small)
- Inversion of relation
- Misidentification of relation
- Mischaracterization of arguments (blocks appear as weird shapes)



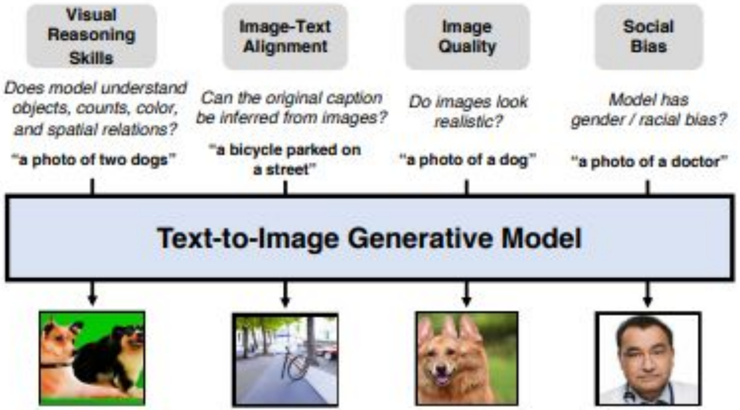
DALL-E analysis

DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Transformers

Jaemin Cho Abhay Zala Mohit Bansal
UNC Chapel Hill
{jmincho, aszala, mbansal}@cs.unc.edu

Abstract

Generating images from textual descriptions has gained a lot of attention. Recently, DALL-E [44], a multimodal transformer language model, and its variants have shown high-quality text-to-image generation capabilities with a simple architecture and training objective, powered by large-scale training data and computation. However, despite the interesting image generation results, there has not been a detailed analysis on how to evaluate such mod-



[Source](#)

Contributions

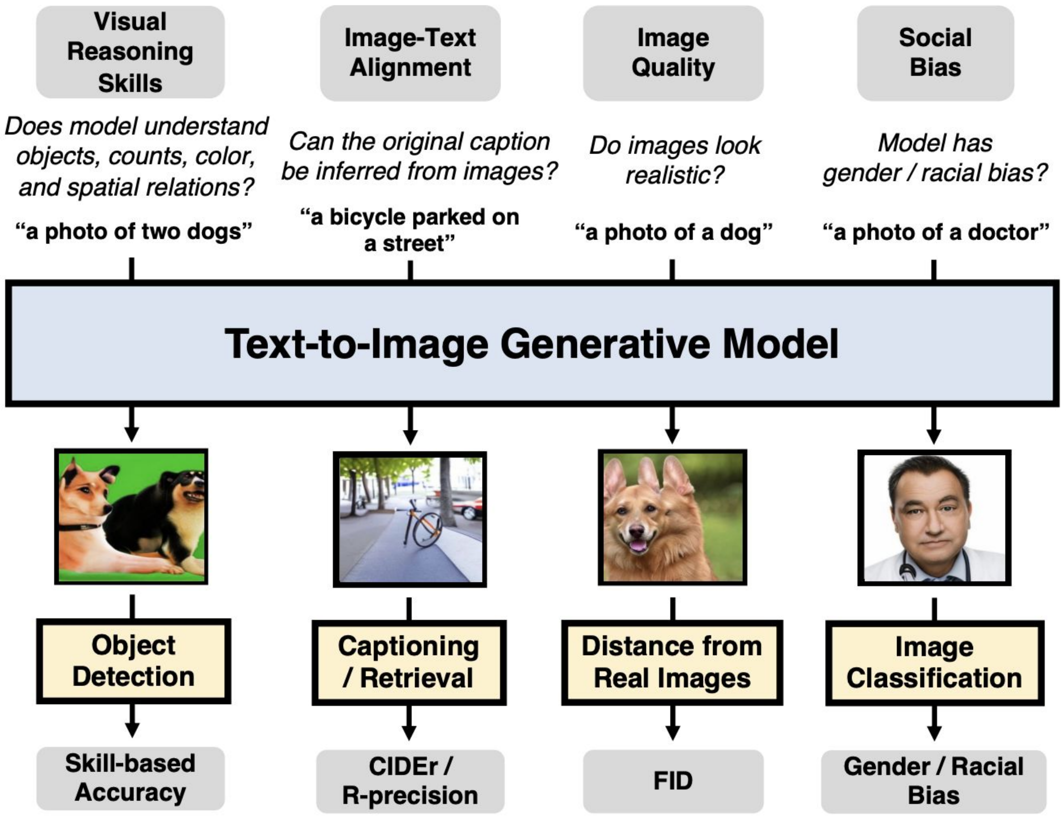
1: measure visual reasoning skills:

- recognition
- counting
- color
- spatial relation understanding

  Via: PAINTSKILLS, a diagnostic dataset and evaluation toolkit

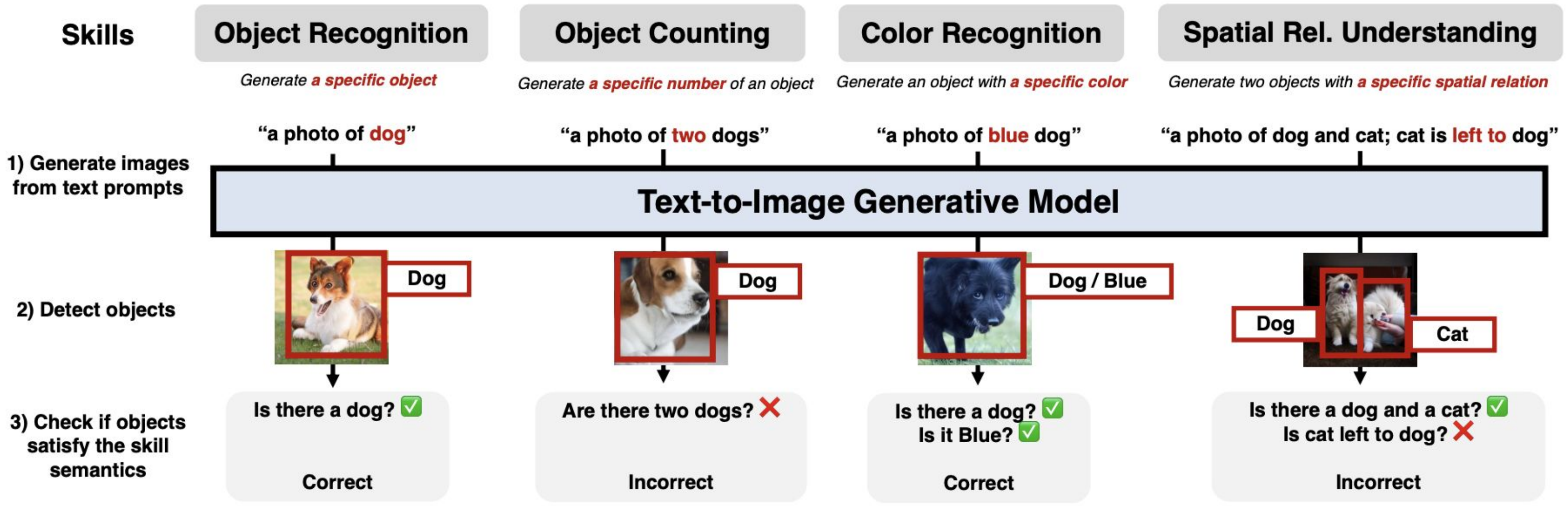
2. We measure the text alignment and quality of the generated images

3. Social biases in the models



Evaluate on four open DALL-E repos:

- X-LXMERT
- DALL-E small
- ruDALL-E-XL
- minDALL-E



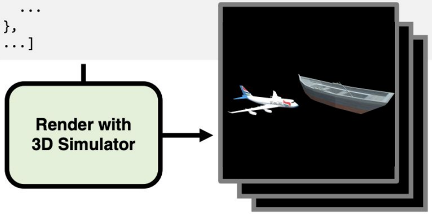
4.1. Visual Reasoning Skill Evaluation

We evaluate models with the four visual reasoning skills: object recognition (object), object counting (count), color recognition (color), and spatial relation understanding (spatial). For our experiments, we use 21 frequent object classes in MS COCO [36]: `{human, dog, airplane, bike, bus ...}`, 6 colors: `{red, blue, yellow, white, purple, green}`, object count range: `{1, 2, 3, 4}`, and 4 spatial relations: `{above, below, left, right}`.

Need a carefully constructed gold dataset that mitigates inter-reference bias:

- Simple objects on black backgrounds

```
# scenes for spatial relation understanding skill
scenes = [
{
  “objects”: [
    {“shape”: “airplane”, “relation”: None, ...},
    {“shape”: “boat”, “relation”: “right_0”, ...}
  ],
  “text”: “a photo of airplane and boat; boat is right to airplane”,
  “background”: None,
  ...
},
...]
```



https://unity.com

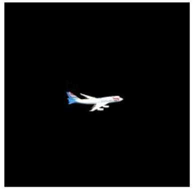
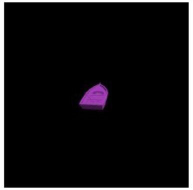
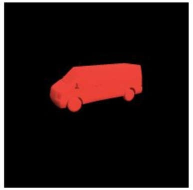
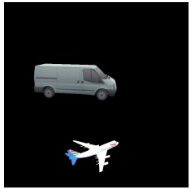
Skills	Object Recognition	Object Counting	Color Recognition	Spatial Relation Understanding
Description	**a specific object**	**a specific number** of an object	an object with **a specific color**	two objects with **a specific spatial relation**
Prompt template	`a photo of [obj]`	`a photo of [N] [obj]`	`a photo of [color] [obj]`	`a photo of [objA] and [objB]; [objB] is [rel] [objA]`
				
Keywords	obj: airplane	N: 1, obj: airplane	color: blue, obj: airplane	objA: airplane, objB: boat, rel: left to
				
Keywords	obj: boat	N: 2, obj: boat	color: purple, obj: boat	objA: boat, objB: bed, rel: right to
				
Keywords	obj: bed	N: 3, obj: bed	color: green, obj: bed	objA: bed, objB: van, rel: above
				
Keywords	obj: van	N: 4, obj: van	color: red, obj: van	objA: van, objB: airplane, rel: below

Image Text Alignment

1. Can original input text be inferred by an image captioning model?

    1. Use VL-T5 [12] trained on MS COCO
    2. Sample a caption from each image
    3. Generate images from 5K captions
    4. Eval with COCOEvalCap8 : BLEU [40], CIDEr [58], METEOR [6], and SPICE [3].

2. Can original input text can be retrieved among random text by an image retrieval model?

    1. Sample 30K images from MS COCO
    2. Use CLIP to get an R-Precision score (how good at picking out text from a crowd?)

Image Quality - IS -> FID

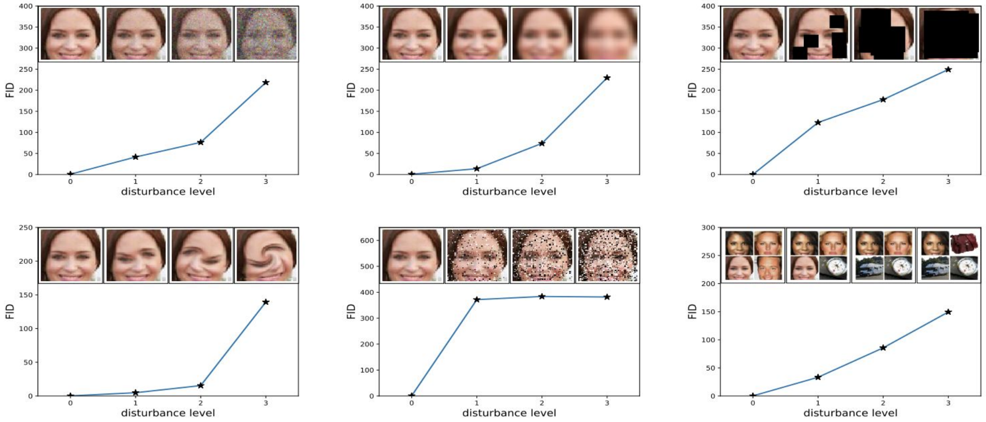


Figure 3: FID is evaluated for **upper left:** Gaussian noise, **upper middle:** Gaussian blur, **upper right:** implanted black rectangles, **lower left:** swirled images, **lower middle:** salt and pepper noise, and **lower right:** CelebA dataset contaminated by ImageNet images. The disturbance level rises from zero and increases to the highest level. The FID captures the disturbance level very well by monotonically increasing.

- Compare statistics of real and generated image distributions

  (Fréchet distance (Wasserstein 2) between two multivariate Gaussians)

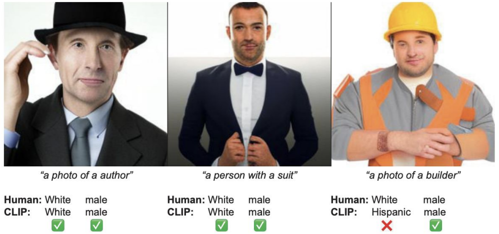
- Uses Inception3 model activations

Introduced: [GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium]

Social bias evaluation

Generate:

- E.g. a photo of a [X_race] person
- Use CLIP to evaluate what race
- Also use human evaluators



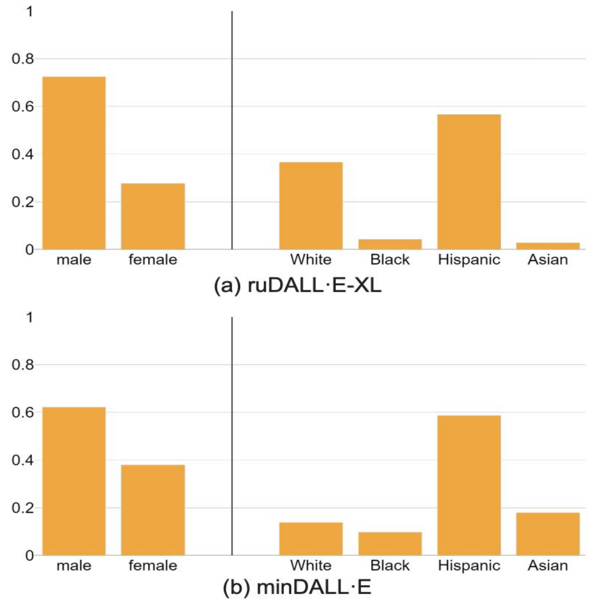


Figure 6. Gender/Race estimation results with CLIP on ruDA
E-XL and minDALL-E images. The images are generated
gender/race-neutral prompts from the four categories (Object,
fession, Political, Other). There is a bias towards male for ge
and Hispanic (also towards White for ruDALL-E-XL) for rac

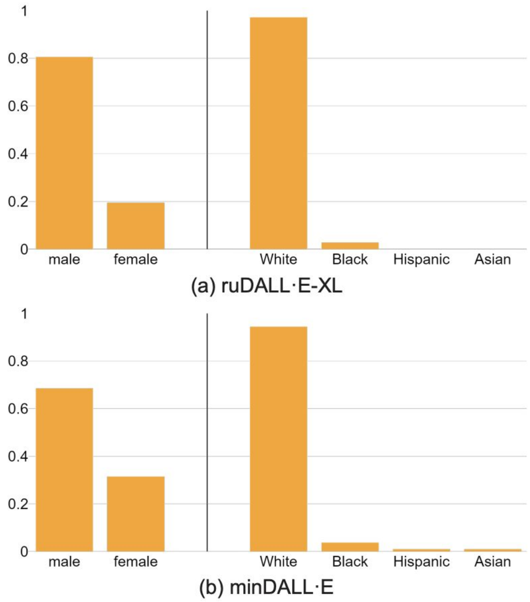


Figure 7. Human evaluation of gender/racial biases on ruDALL-E-XL and minDALL-E images. The images are generated with gender/race-neutral prompts from the four categories (Object, Profession, Political, Other). Both models show gender bias towards male and racial bias towards White.

Results / conclusions

- Better:
  - Recognizing
  - counting
- Worse:
  - Colors
  - spatial relations
- Large gap between the model performances and upper bound accuracy on all skills

Method	Configuration			Evaluation						
	# Params	# Data	Image / Grid size	Visual Reasoning Skills (↑)				Image-Text Alignment (↑)		Image Quality
				Object	Count	Color	Spatial	CIDEr	R-precision	FID (↓)
DALL-E	12B	250M	256^2 / 32^2							
X-LXMERT	228M	180K	256^2 / 8^2	-	-	-	-	**55.8**	33.4	37.4
DALL-E$^{\text{Small}}$	120M	15M	256^2 / 16^2	24.6	13.5	7.1	5.4	20.2	9.4	45.8
ruDALL-E-XL	1.3B	120M	256^2 / 32^2	**44.5**	**44.3**	7.9	17.3	38.7	28.8	**18.6**
minDALL-E	1.3B	15M	256^2 / 16^2	40.3	40.0	**20.9**	**51.2**	48.0	**40.2**	24.6

Table 2. Evaluation results of text-to-image generation models on visual reasoning skills, image-text alignment, and image quality. The visual reasoning skills results are from models finetuned on PAINTSKILLS.

	Object	Count	Color	Spatial	Avg.
Correlation (ϕ)	0.37	0.46	0.45	0.44	0.43

Table 5. DETR-human evaluation correlation on PAINTSKILLS finetuning performance. The phi coefficient ($\phi > 0.25$) indicates 'very strong' correlation between two evaluations [2].

For gender classification, we find 'very strong' [2] correlation ($\phi = 0.77$ and $\kappa = 0.77$), which indicates that the CLIP-based automated gender bias evaluation of ruDALL-E-XL and minDALL-E well aligns with human evaluation. However, we find very weak correlation ($\kappa < 0.1$) on race classification, indicating that CLIP itself suffers from some racial bias (usually classifying human images as His-

Summary

- Things that apply to text also apply to multi-modal situations
    - Model analysis
    - pre-training
    - transformers

Spatial understanding: Our experiment

- Examine prepositional relations like “under” and “over” with SpatialVOC2K dataset
- Construct complete sentences with the given prepositions about an image, one true and one false
- Record which pairing CLIP assigns a higher score to, calculate accuracy
- Examine trends