38th Vienna Deep Learning Meetup

17th February 2021

#VDLM

virtual edition

Vienna
Deep Learning
Meetup

The Organizers:

Thomas Lidy
Musimap

Alex Schindler
AIT & TU Wien

Jan Schlüter
JKU Linz

René Donner
contextflow

Agenda for Today

Vienna
Deep Learning
Meetup

Welcome & Introduction

Announcements

"OpenAI: CLIP & DALL·E" *by Michael Pieler, contextflow*

"Coordinate-based Neural Representations" *by Jan Schlüter, JKU*

("Neural Architecture Search / AutoML" *by René Donner, contextflow* ***postponed*)**

Networking in Breakout-Rooms

Announcements

Machine Learning Prague

FEBRUARY 26 - 28, 2021

Online practical conference about ML, AI and Deep Learning applications

1 000+ ATTENDEES

45 SPEAKERS

10 WORKSHOPS

5+ MASTERMIND SESSIONS

1 HACKATHON

Code 20% off - vdlmeetup

www.mlprague.com

Machine Learning Prague

Machine Learning Prague

FEBRUARY 26 - 28, 2021

Online practical conference about ML, AI and Deep Learning applications

1 000+	45	10	5+	1
ATTENDEES	SPEAKERS	WORKSHOPS	MASTERMIND SESSIONS	HACKATHON

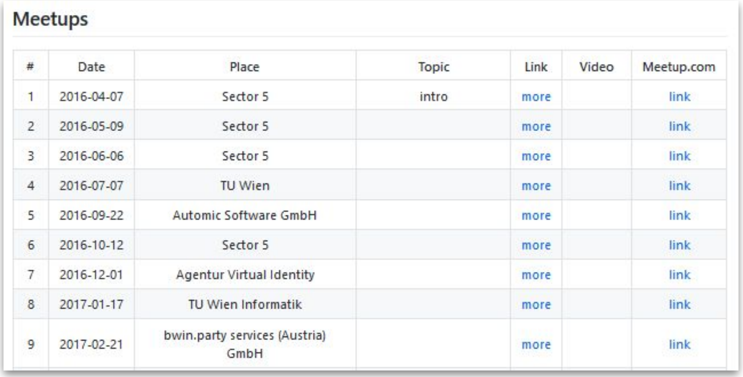
2 Winners of free tickets: Nathanya and Ingrid

Machine Learning Prague

VDLM on Github

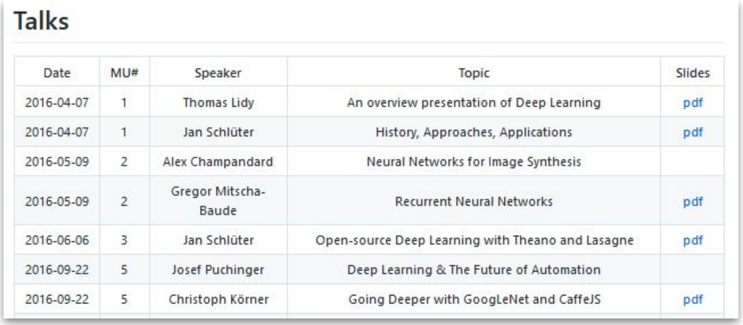
https://github.com/vdlm/meetups

- all talks
- slides
- photos
- videos
- Wiki



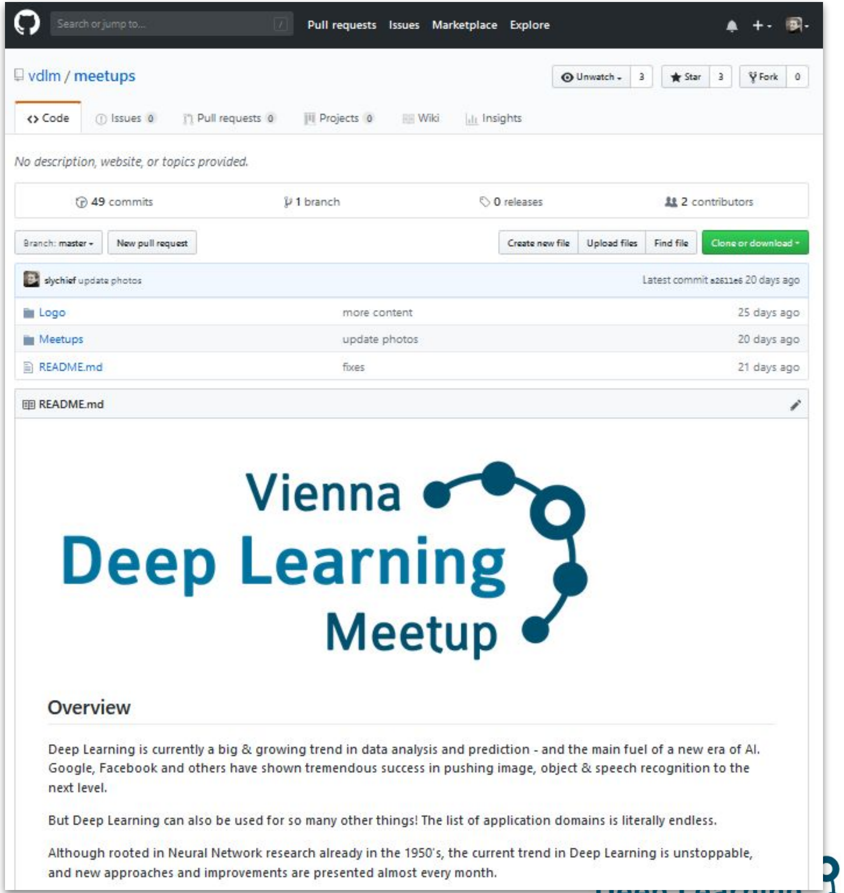
Meetups

#	Date	Place	Topic	Link	Video	Meetup.com
1	2016-04-07	Sector 5	intro	more		link
2	2016-05-09	Sector 5		more		link
3	2016-06-06	Sector 5		more		link
4	2016-07-07	TU Wien		more		link
5	2016-09-22	Automic Software GmbH		more		link
6	2016-10-12	Sector 5		more		link
7	2016-12-01	Agentur Virtual Identity		more		link
8	2017-01-17	TU Wien Informatik		more		link
9	2017-02-21	bwin.party services (Austria) GmbH		more		link



Talks

Date	MU#	Speaker	Topic	Slides
2016-04-07	1	Thomas Lidy	An overview presentation of Deep Learning	pdf
2016-04-07	1	Jan Schlüter	History, Approaches, Applications	pdf
2016-05-09	2	Alex Champandard	Neural Networks for Image Synthesis	
2016-05-09	2	Gregor Mitscha-Baude	Recurrent Neural Networks	pdf
2016-06-06	3	Jan Schlüter	Open-source Deep Learning with Theano and Lasagne	pdf
2016-09-22	5	Josef Puchinger	Deep Learning & The Future of Automation	
2016-09-22	5	Christoph Körner	Going Deeper with GoogLeNet and CaffeJS	pdf



vdlm / meetups

No description, website, or topics provided.

49 commits · 1 branch · 0 releases · 2 contributors

slychief update photos — Latest commit a2611e6 20 days ago

Logo	more content	25 days ago
Meetups	update photos	20 days ago
README.md	fixes	21 days ago

README.md

Vienna Deep Learning Meetup

Overview

Deep Learning is currently a big & growing trend in data analysis and prediction - and the main fuel of a new era of AI. Google, Facebook and others have shown tremendous success in pushing image, object & speech recognition to the next level.

But Deep Learning can also be used for so many other things! The list of application domains is literally endless.

Although rooted in Neural Network research already in the 1950's, the current trend in Deep Learning is unstoppable, and new approaches and improvements are presented almost every month.

Agenda for Today

Welcome & Introduction

Announcements

"OpenAI: CLIP & DALL·E" *by Michael Pieler, contextflow*

"Coordinate-based Neural Representations" *by Jan Schlüter, JKU*

("Neural Architecture Search / AutoML" *by René Donner, contextflow* ***postponed*)**

Networking in Breakout-Rooms

CLIP

https://openai.com/blog/clip/

Introduction

- A model that learns visual concepts from natural language supervision.

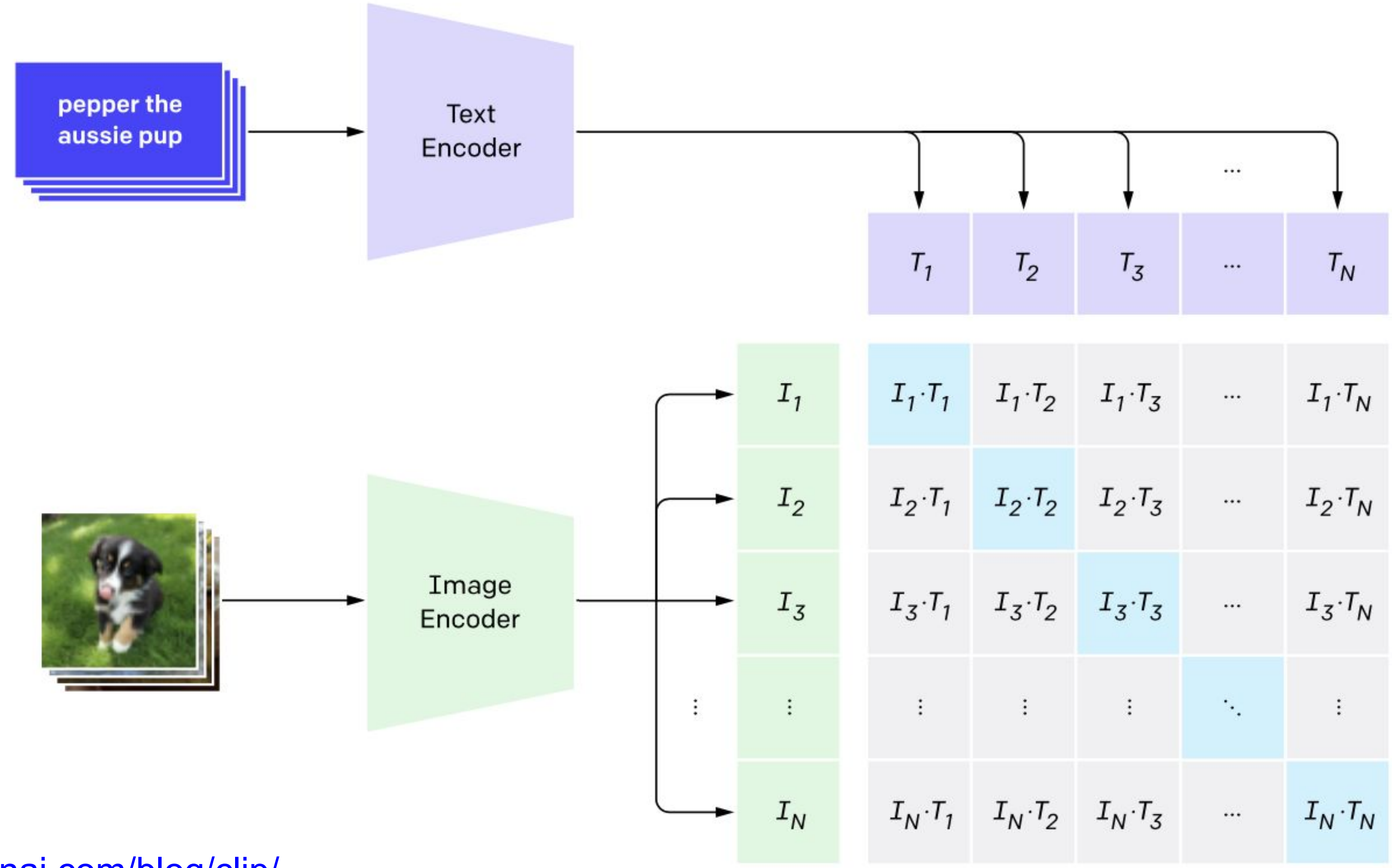
- Can be applied to any visual classification benchmark (similar to the “zero-shot” capabilities of GPT-2 and GPT-3).

- By not directly optimizing for the benchmark, results become much more representative.

→ Architecture!

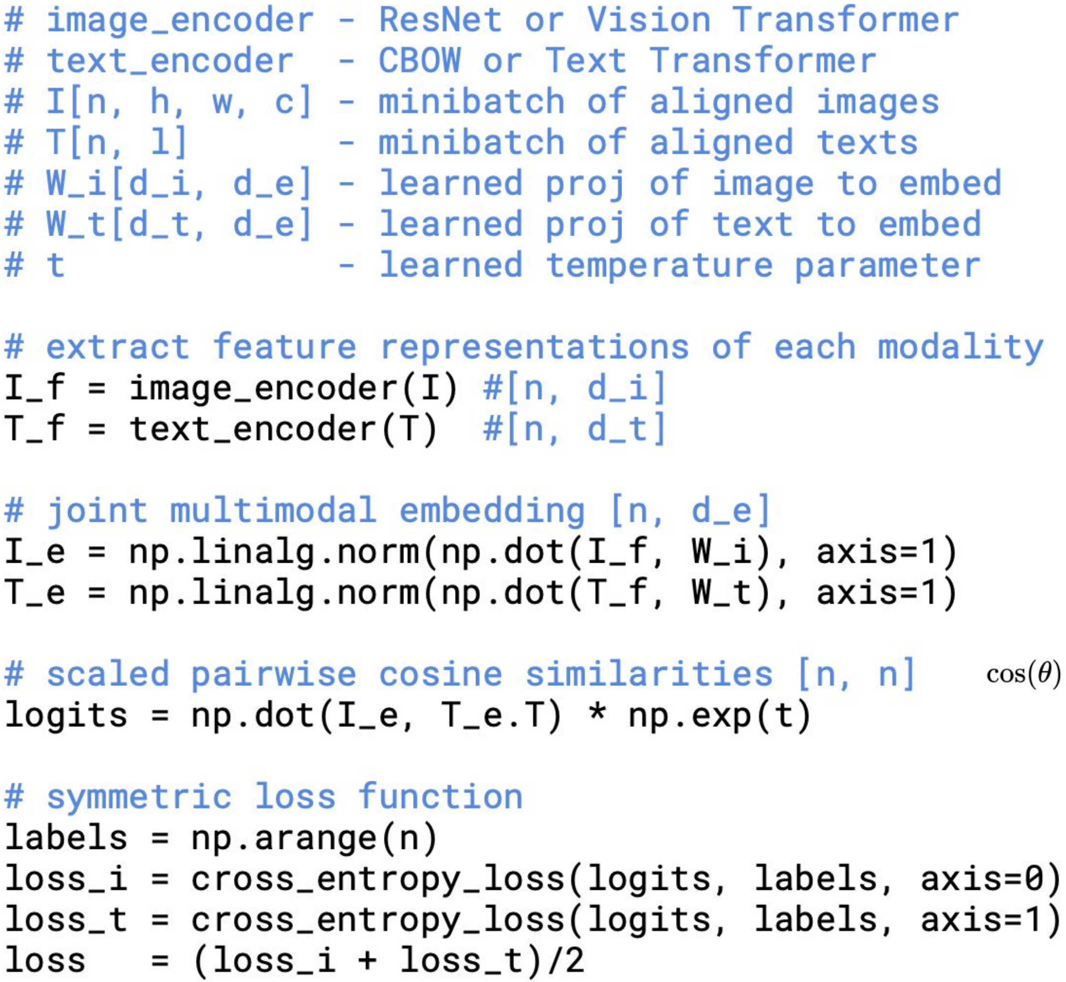
https://openai.com/blog/clip/

1. Contrastive pre-training



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

https://openai.com/blog/clip/

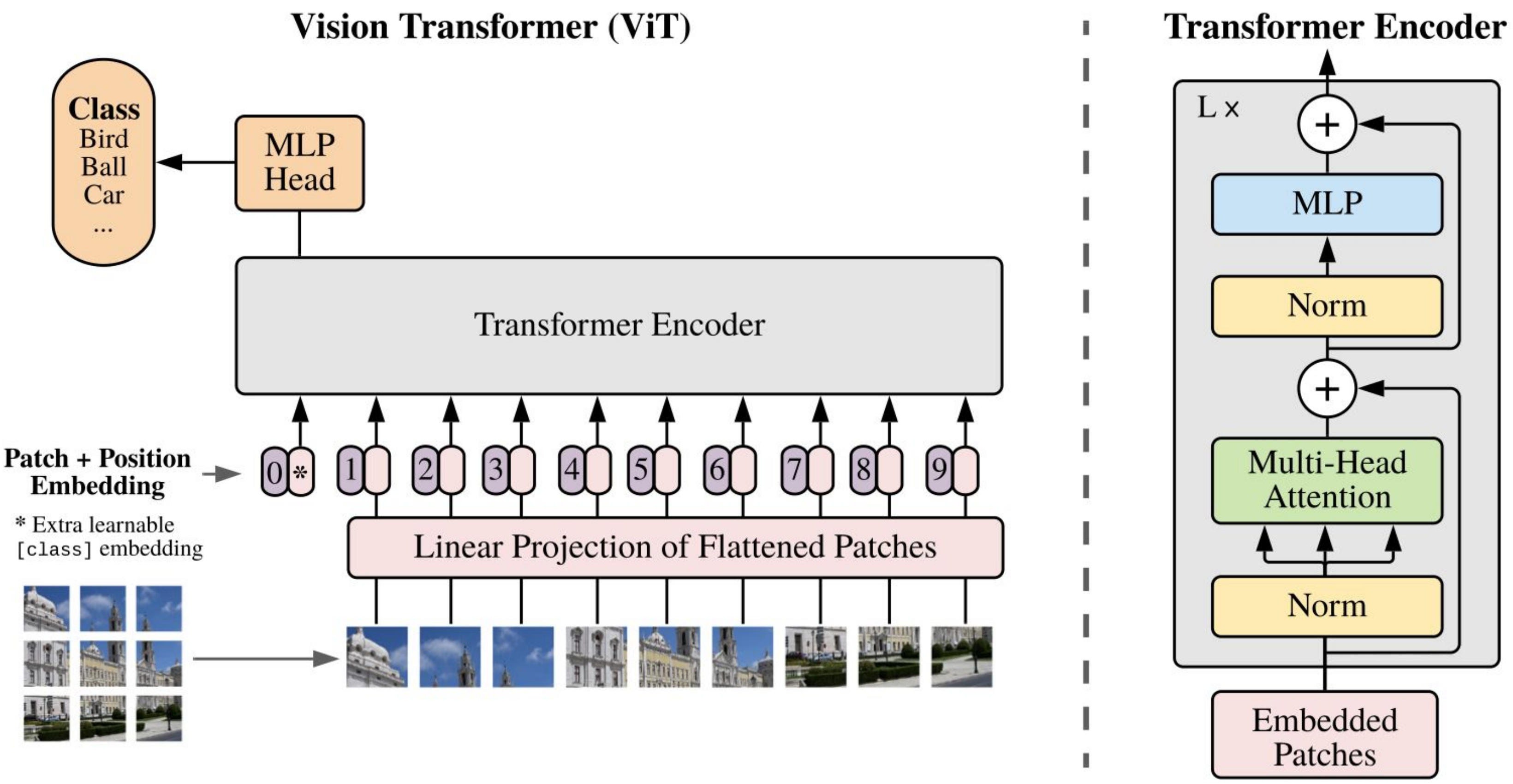


```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = np.linalg.norm(np.dot(I_f, W_i), axis=1)
T_e = np.linalg.norm(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

https://openai.com/blog/clip/

Vienna Deep Learning Meetup

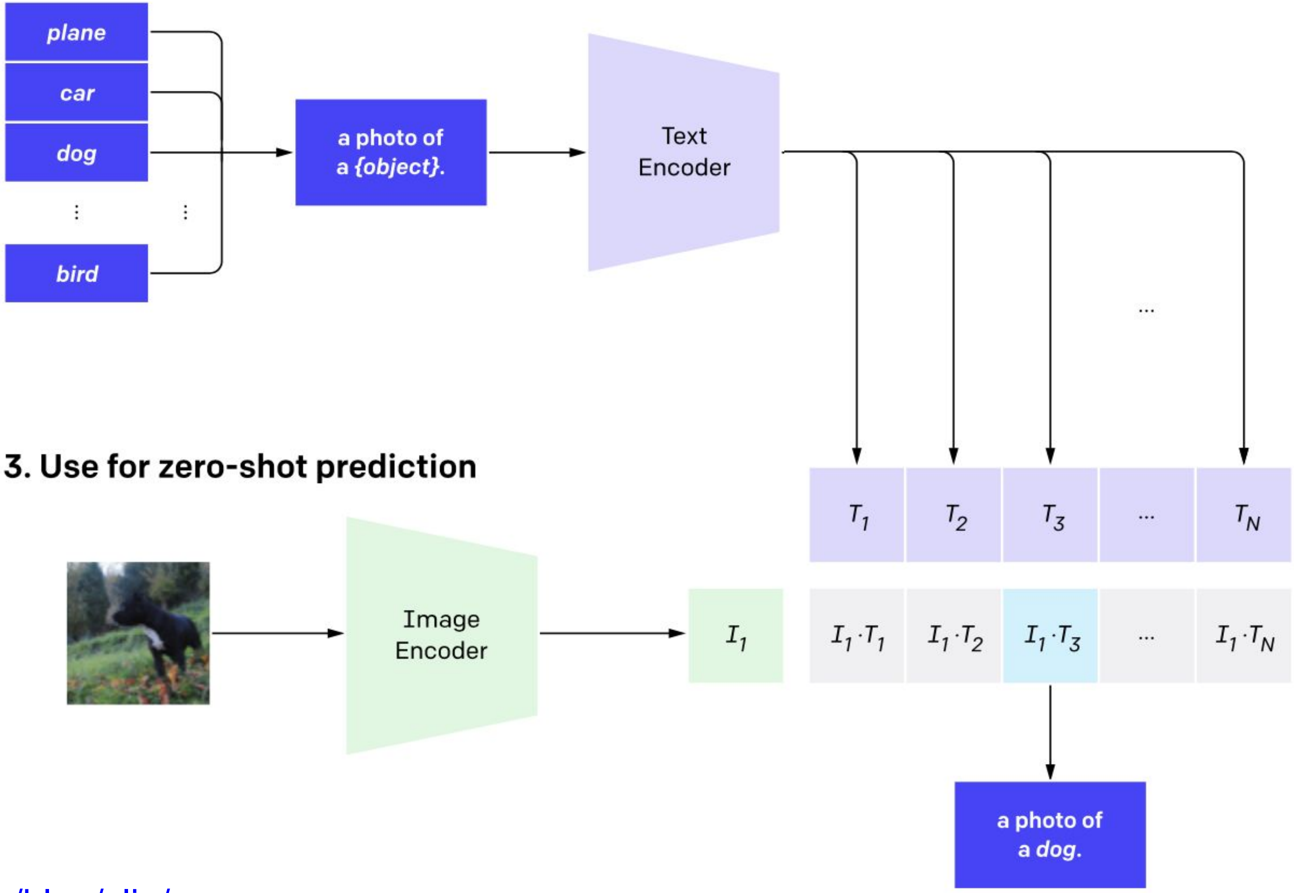


https://openreview.net/pdf?id=YicbFdNTTy

Vienna Deep Learning Meetup

2. Create dataset classifier from label text

3. Use for zero-shot prediction



plane

car

dog

⋮

bird

a photo of a *{object}*.

Text Encoder

T_1 T_2 T_3 … T_N

Image Encoder

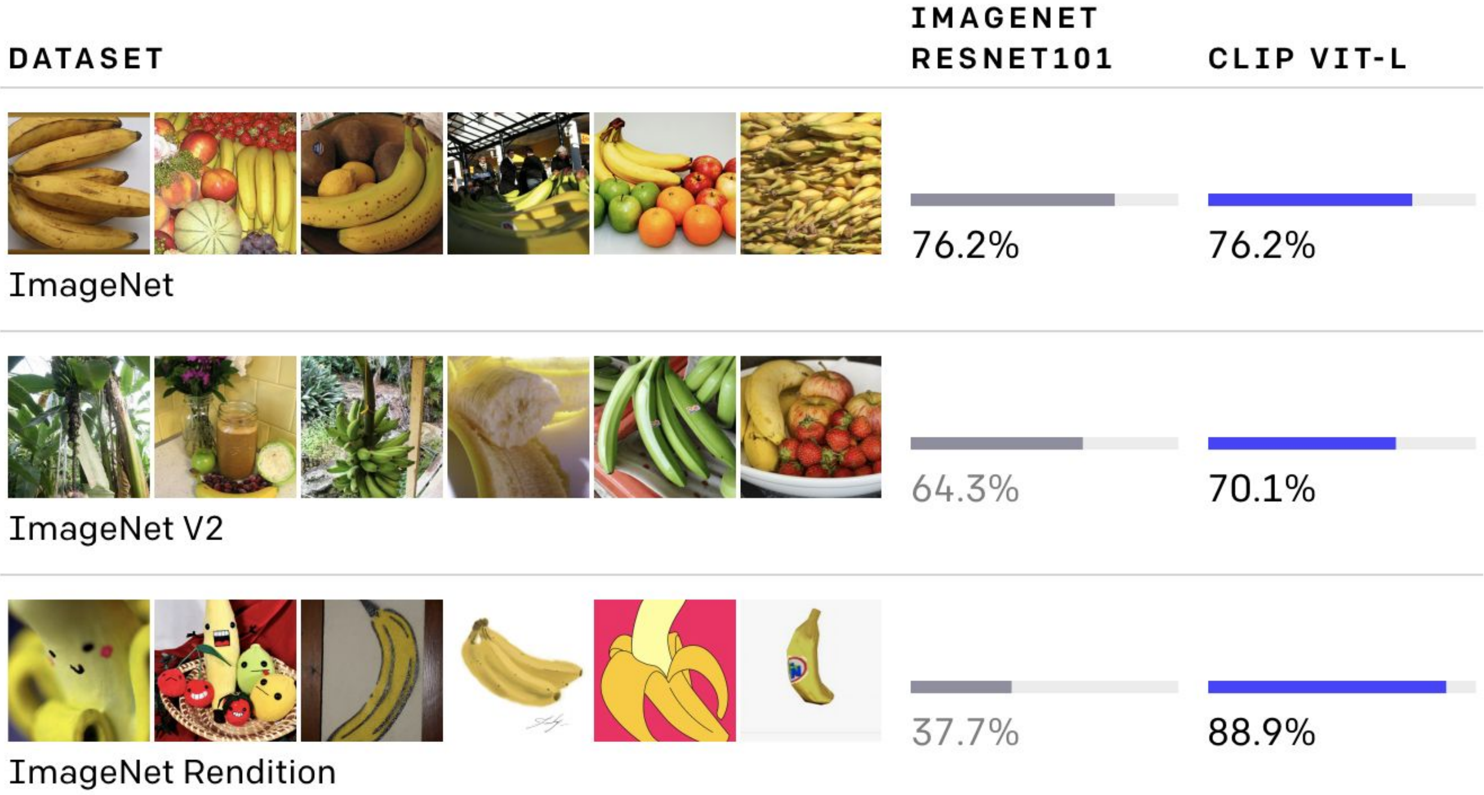
I_1

$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ … $I_1 \cdot T_N$

a photo of a *dog*.

https://openai.com/blog/clip/

Results

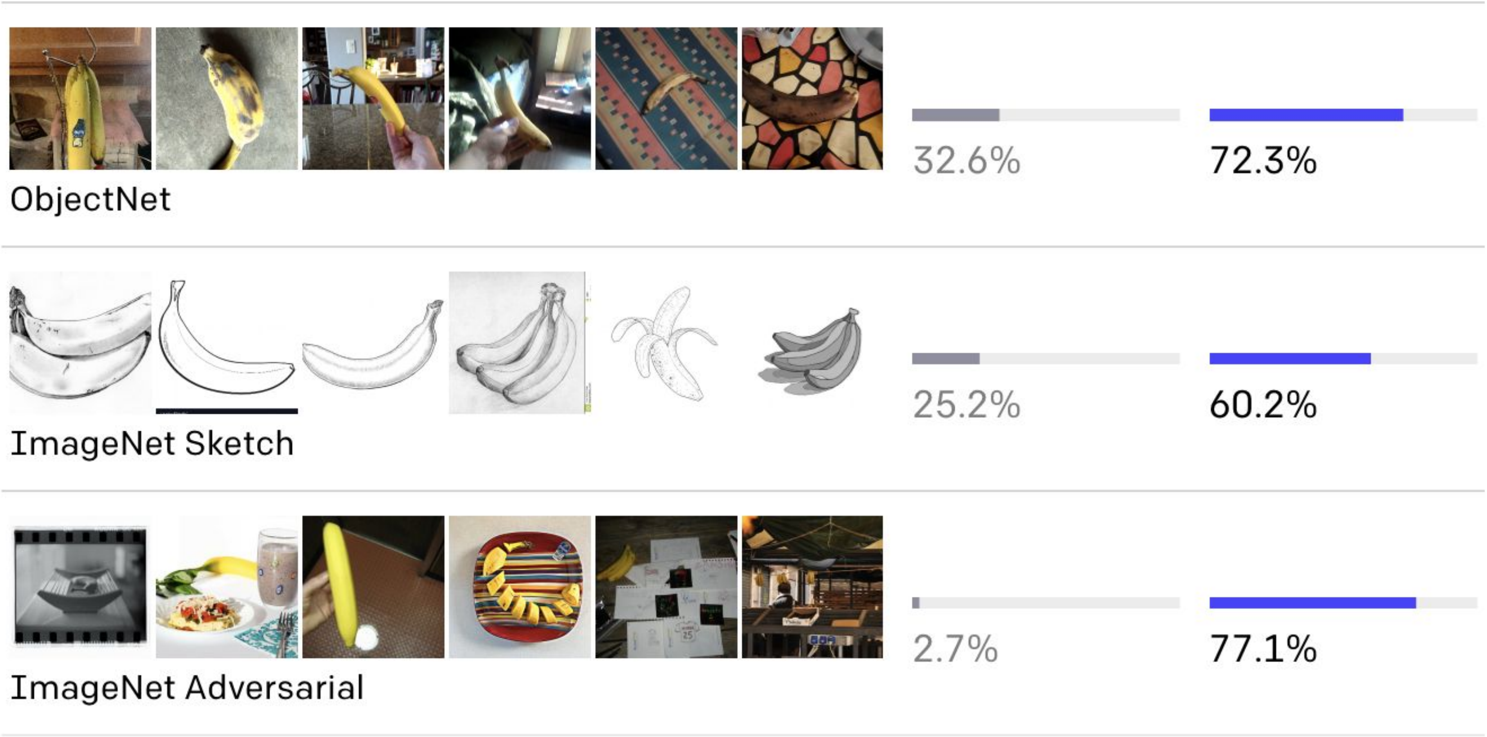


DATASET	IMAGENET RESNET101	CLIP VIT-L
ImageNet	76.2%	76.2%
ImageNet V2	64.3%	70.1%
ImageNet Rendition	37.7%	88.9%

https://openai.com/blog/clip/

Vienna Deep Learning Meetup

Results



ObjectNet	32.6%	72.3%
ImageNet Sketch	25.2%	60.2%
ImageNet Adversarial	2.7%	77.1%

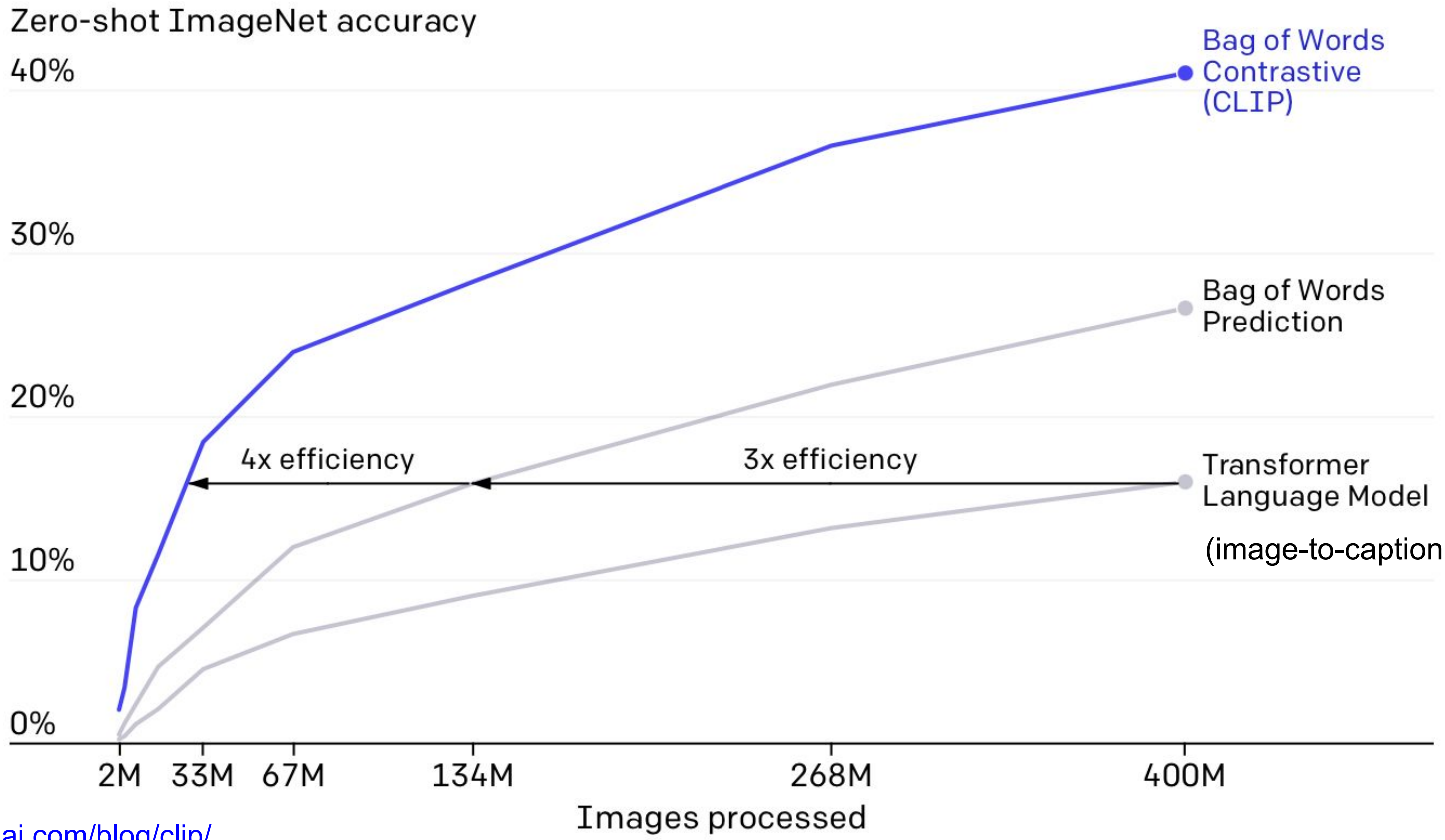
https://openai.com/blog/clip/

Key takeaways

- CLIP is highly efficient:
  - Adoption of contrastive objective for connecting text with images is 4x to 10x more efficient at zero-shot ImageNet classification.
  - Adoption of the ViT gave a further 3x gain in compute efficiency over a standard ResNet.
  - (The best performing CLIP model trains on 256 GPUs for 2 weeks which is similar to existing large scale image models.)

https://openai.com/blog/clip/

Key takeaways



(image-to-caption language models)

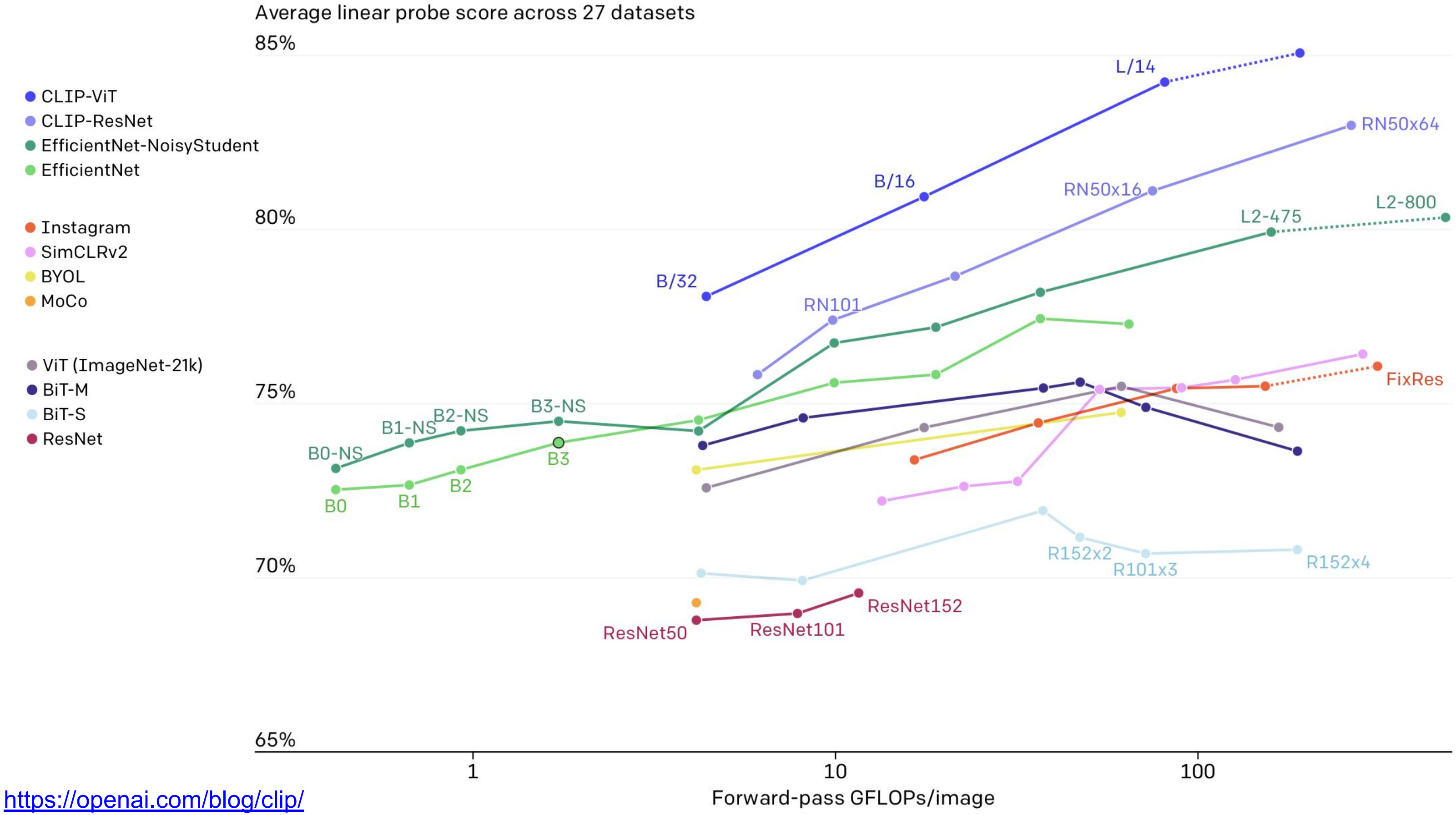
https://openai.com/blog/clip/

Key takeaways

- CLIP is flexible and general:
    - Because CLIP learns a wide range of visual concepts directly from natural language, it is significantly more flexible and general than existing ImageNet models.
    - Performs a wide set of tasks during pre-training including OCR, geo-localization, action recognition, and many others.
    - CLIP is able to zero-shot perform many different tasks, e.g.,
    - CLIP's zero-shot performance was validated on over 30 different datasets (see next slide).

https://openai.com/blog/clip/

Key takeaways



https://openai.com/blog/clip/

Vienna Deep Learning Meetup

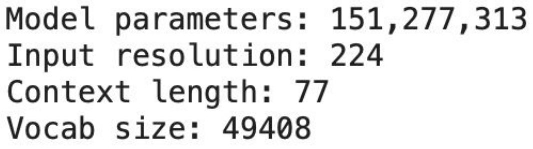
Limitations

- CLIP usually performs well on recognizing common objects, it struggles on more abstract or systematic tasks.
- Zero-shot CLIP struggles compared to task specific models on very fine-grained classification.
- CLIP has poor generalization to images not covered in its pre-training dataset.
- CLIP's zero-shot classifiers can be sensitive to wording or phrasing and sometimes require trial and error "prompt engineering" to perform well.

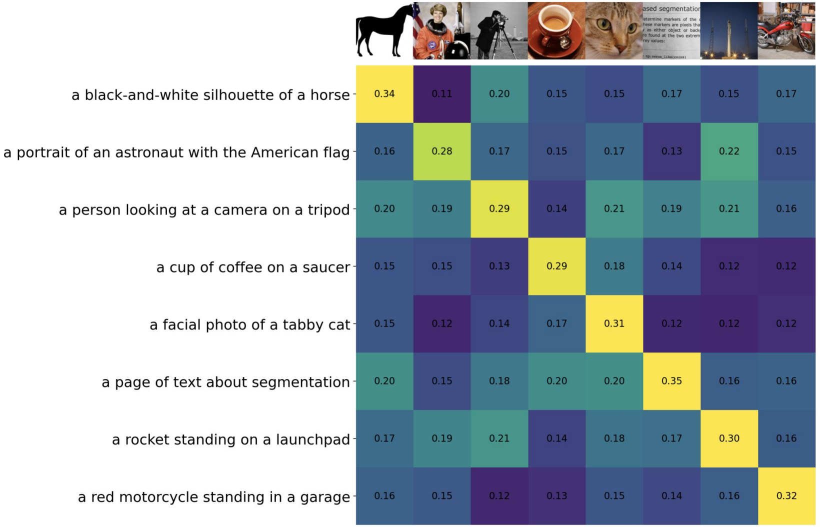
https://openai.com/blog/clip/

Demo

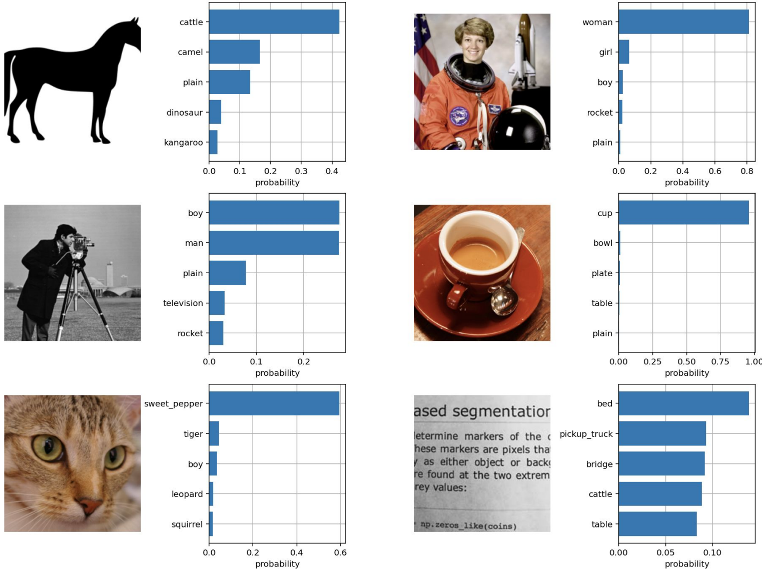
- https://colab.research.google.com/github/openai/clip/blob/master/notebooks/Interacting_with_CLIP.ipynb



```
Model parameters: 151,277,313
Input resolution: 224
Context length: 77
Vocab size: 49408
```



a black-and-white silhouette of a horse	0.34	0.11	0.20	0.15	0.15	0.17	0.15	0.17
a portrait of an astronaut with the American flag	0.16	0.28	0.17	0.15	0.17	0.13	0.22	0.15
a person looking at a camera on a tripod	0.20	0.19	0.29	0.14	0.21	0.19	0.21	0.16
a cup of coffee on a saucer	0.15	0.15	0.13	0.29	0.18	0.14	0.12	0.12
a facial photo of a tabby cat	0.15	0.12	0.14	0.17	0.31	0.12	0.12	0.12
a page of text about segmentation	0.20	0.15	0.18	0.20	0.20	0.35	0.16	0.16
a rocket standing on a launchpad	0.17	0.19	0.21	0.14	0.18	0.17	0.30	0.16
a red motorcycle standing in a garage	0.16	0.15	0.12	0.13	0.15	0.14	0.16	0.32



Vienna Deep Learning Meetup

Conclusion

- CLIP allows people to design their own classifiers and removes the need for task-specific training data.

- CLIP does not need task-specific training data, therefore, it can unlock certain niche tasks with greater ease.

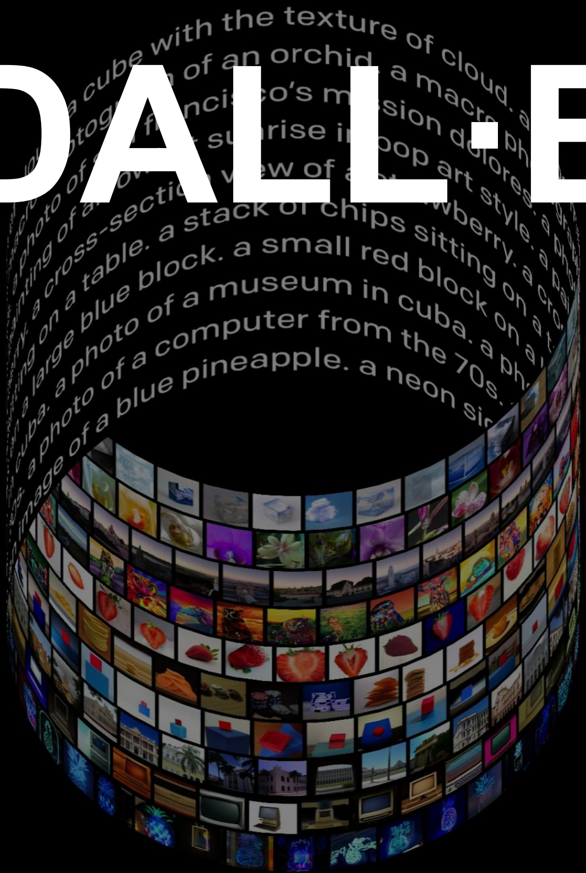
  → We will look at some additional interesting applications at the end! :-)

https://openai.com/blog/clip/

Interesting CLIP applications

- General: https://twitter.com/quasimondo/status/1351191660059832320 & https://twitter.com/Buntworthy/status/1348346412208189441
- Filter noisy fotos: https://twitter.com/l4rz/status/1352630033832140800
- Natural language video search: https://github.com/haltakov/natural-language-youtube-search
- Visual search engine: https://same.energy

DALL·E


a cube with the texture of cloud. a
of an orchid. a macro ph
francisco's mission dolores
sunrise in pop art style.
view of a strawberry. a ph
a cross-section
a table. a stack of chips sitting on a cro
a large blue block. a small red block on a t
a photo of a museum in cuba. a ph
a photo of a computer from the 70s.
image of a blue pineapple. a neon sig

https://openai.com/blog/dall-e/

Vienna Deep Learning Meetup

Introduction

- Is able to create plausible images for a great variety of sentences that explore the compositional structure of language.

- 12-billion parameter version of GPT-3
    - trained to generate images from text descriptions
    - using a dataset of text–image pairs

- Receives both the text and the image as a single stream of data containing up to 1280 tokens.

- Trained using MLE to generate all of the tokens, one after another.

https://openai.com/blog/dall-e/

Architecture & examples

- Standard causal mask for the text tokens.
- Sparse attention for the image tokens with either a row, column, or convolutional attention pattern, depending on the layer.
- Reranking with CLIP (kind of language-guided search).
- OpenAI plans to provide more details about the architecture and training procedure in an upcoming paper.

https://openai.com/blog/dall-e/

Architecture

- (Inofficial) implementation: https://github.com/lucidrains/DALLE-pytorch
  - Pretrained discrete VAE network
  - DALL·E network: Transformer and pretrained discrete VAE
  - CLIP (can be used) for ranking the generated images

Interesting DALL·E applications

- Examples from the official blog post: https://openai.com/blog/dall-e/
- Illustration of a baby shark in a wizard hat wielding a blue light saber: https://twitter.com/sama/status/1346543962652246017
- https://github.com/lucidrains/big-sleep & https://github.com/lucidrains/deep-daze

Sources

https://openai.com/blog/clip/

https://cdn.openai.com/papers/Learning_Transferable_Visual_Models_From_Natural_Language.pdf

https://github.com/openai/CLIP

https://colab.research.google.com/github/openai/clip/blob/master/notebooks/Interacting_with_CLIP.ipynb

https://www.reddit.com/r/MachineLearning/comments/ldc6oc/p_list_of_sitesprogramsprojects_that_use_openais/

https://www.reddit.com/r/MachineLearning/comments/lcjizm/p_evertrove_we_made_a_usable_mlpowered_image/

https://www.reddit.com/r/MachineLearning/comments/lbwtvb/r_p_generating_images_from_caption_and_vice_versa/

https://openai.com/blog/dall-e/

https://github.com/lucidrains/DALLE-pytorch

https://github.com/EleutherAI/DALLE-mtf

Vienna Deep Learning Meetup

Next Meetup: March 17th, 2021

www.meetup.com/Vienna-Deep-Learning-Meetup