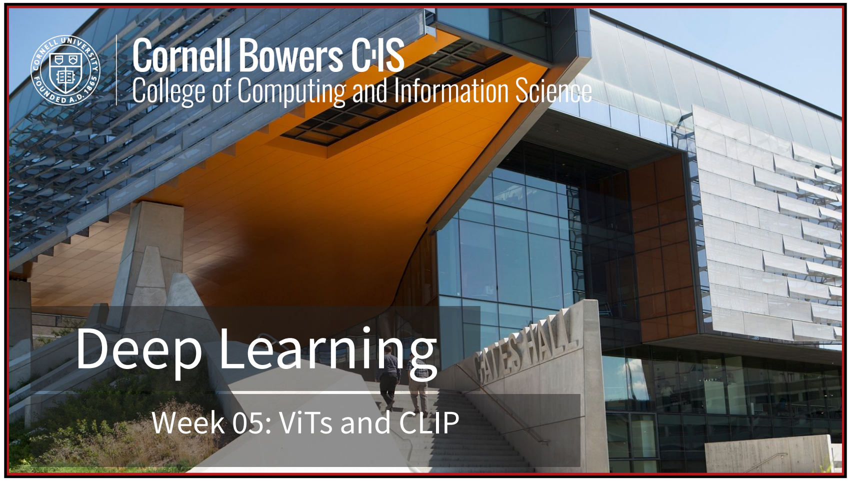


Cornell Bowers C·IS

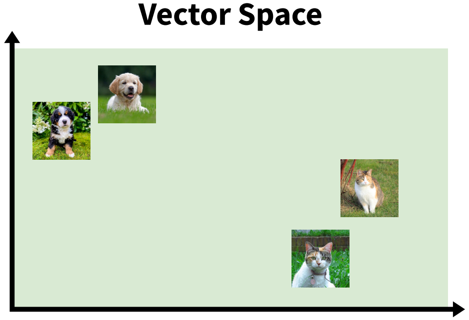
Logistics

- HW3 is due next week
- We have a feedback form (due Friday, March 8th)
- Project proposal due today
  - You can use slip days
  - Please think about compute when you choose a paper for your project
    - Look for reported training hardware and training times
  - Choose papers from ICML, ICLR, NeuRIPS, CVPR, ECCV, ACL, EMNLP, NAACL

Cornell Bowers C·IS

Use self-supervised learning to learn embeddings for images

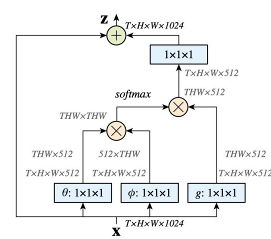
Vector Space



Cornell Bowers C·IS

How to use Attention for Vision Tasks?

Idea #1: Add attention to existing CNNs

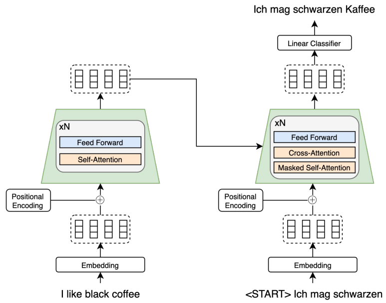


Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794-7803).

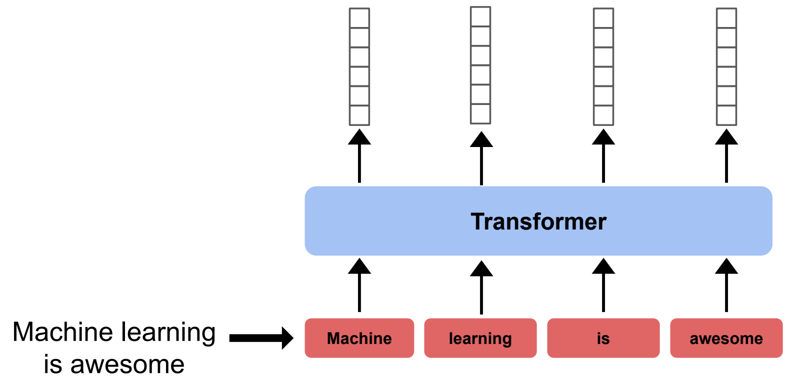
Cornell Bowers CIS

How to use Attention for Vision Tasks?

Idea #2: Adapt standard transformers to image data

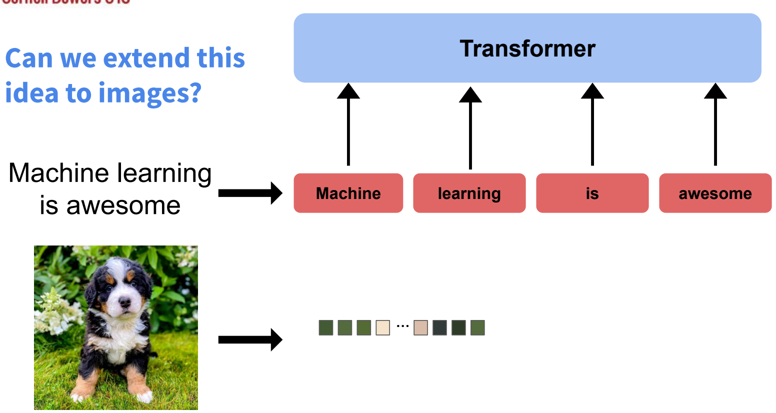


Cornell Bowers CIS



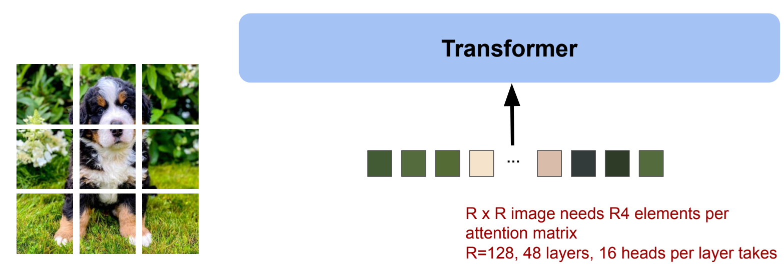
Cornell Bowers CIS

Can we extend this idea to images?



Cornell Bowers CIS

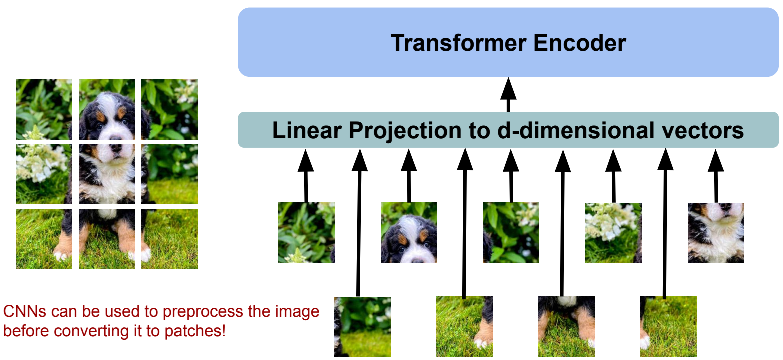
Use pixels as input



R x R image needs R4 elements per attention matrix
R=128, 48 layers, 16 heads per layer takes 768GB of memory for attention matrices for a single example…

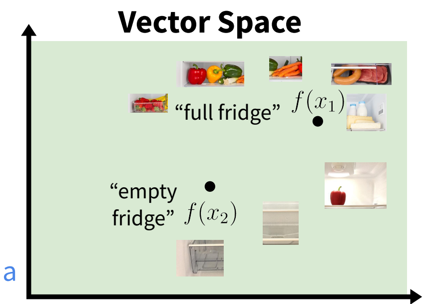
https://web.eecs.umich.edu/~justincj/slides/eecs498/WI2022/598_WI2022_lecture18.pdf

Image patches as “words”



CNNs can be used to preprocess the image before converting it to patches!

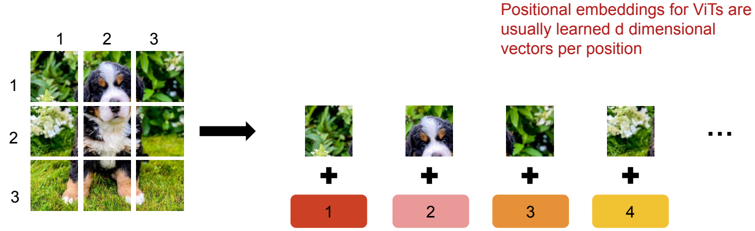
$f(x)$ = Vision Transformers



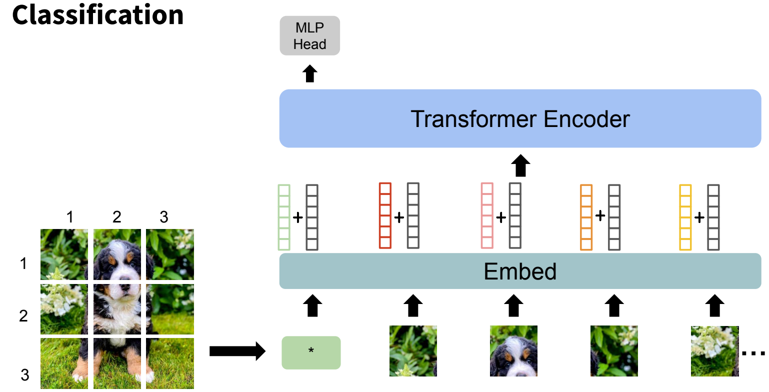
Use image patches like words in a sentence!

Adding positional embeddings

Positional embeddings for ViTs are usually learned d dimensional vectors per position

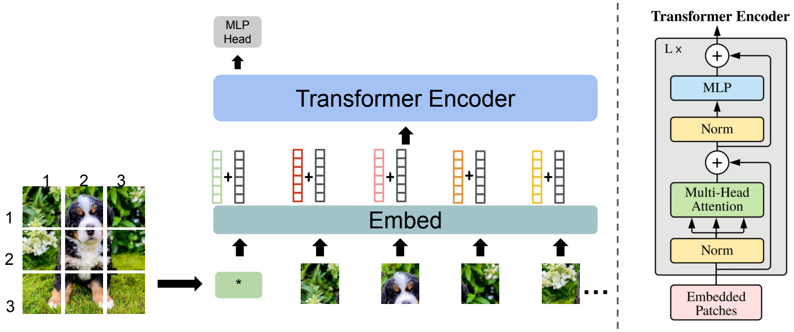


Classification



Cornell Bowers C·IS

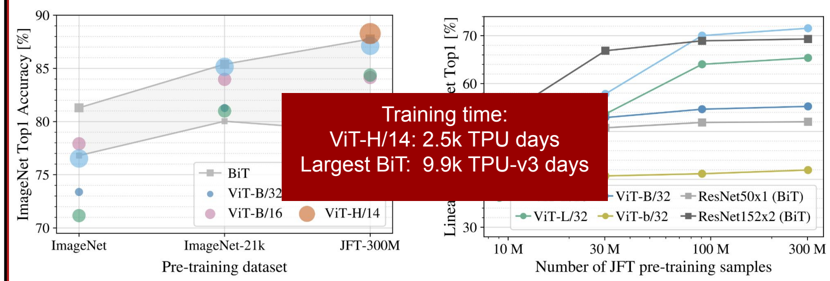
Vision Transformer



Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929.*

Cornell Bowers C·IS

ViT Results



Training time:
ViT-H/14: 2.5k TPU days
Largest BiT: 9.9k TPU-v3 days

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929.*

Cornell Bowers C·IS

ViT Summary

Model:

- Model is almost identical to BERT
- Replace words with PxP pixel image patches, P ∈ {14, 16, 32} (no overlap)
- Each patch is embedded linearly into a vector of size 1024
- 1D positional embeddings

Training:

- For pre-training, optimize for image classification on large supervised dataset (e.g. ImageNet 21K, JFT -300M)
- For fine-tuning, learn a new classification head on a small dataset (e.g. CIFAR-100)

Cornell Bowers C·IS

ACTIVITY: When do ViTs outperform CNNs, and vice versa?

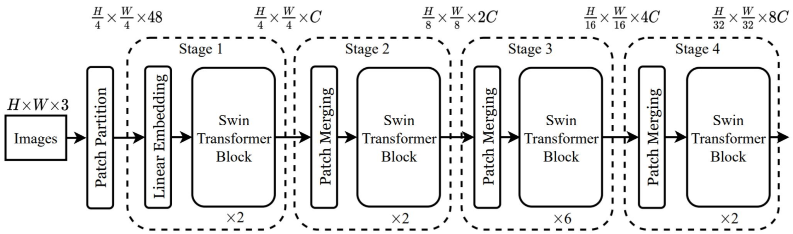
Think about what you know about transformers - what are some of their drawbacks?

When is it “worth it” to use transformers instead of just CNNs?

Cornell Bowers C·IS

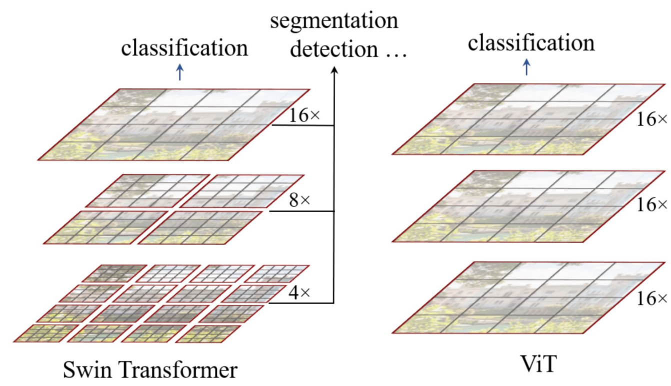
Swim Transformers

Hierarchical architecture that has the flexibility to model at various scales



Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022).

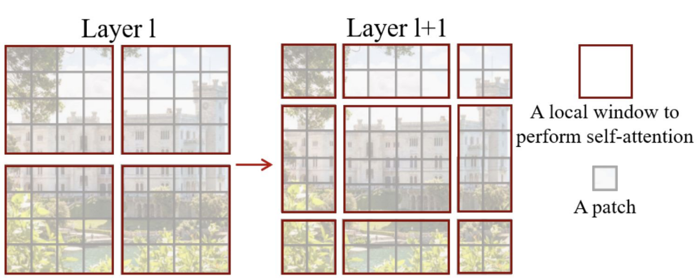
Cornell Bowers C·IS



Cornell Bowers C·IS

Shifted Window attention

Linear computational complexity with respect to image size

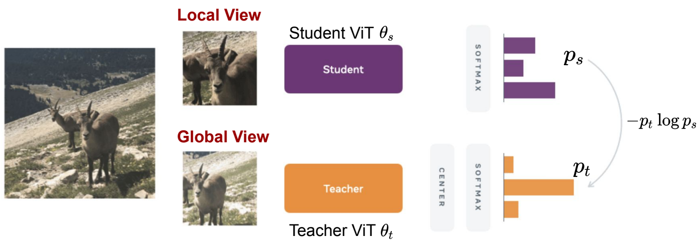


Cornell Bowers C·IS

Performance

method	image size	#param.	FLOPs	throughput (image / s)	ImageNet top-1 acc.
R-101x3 [38]	384^2	388M	204.6G	-	84.4
R-152x4 [38]	480^2	937M	840.5G	-	85.4
ViT-B/16 [20]	384^2	86M	55.4G	85.9	84.0
ViT-L/16 [20]	384^2	307M	190.7G	27.3	85.2
Swin-B	224^2	88M	15.4G	278.1	85.2
Swin-B	384^2	88M	47.0G	84.7	86.4
Swin-L	384^2	197M	103.9G	42.1	87.3

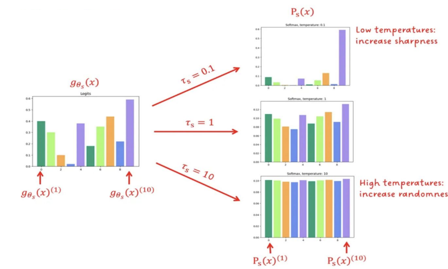
Self-Supervised Vision Transformers (DiNO)



Momentum encoder: $\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$

Centering and sharpening

- Centering prevents one dimension from dominating
- Sharpening prevents learning a uniform distribution



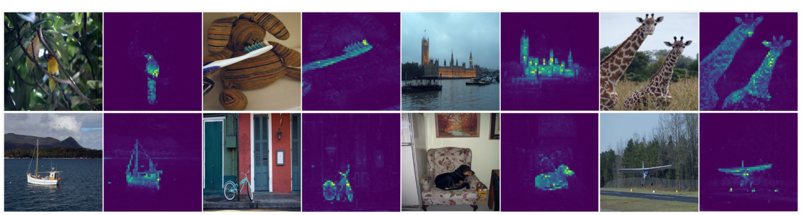


Figure 1: **Self-attention from a Vision Transformer with 8×8 patches trained with no supervision.** We look at the self-attention of the [CLS] token on the heads of the last layer. This token is not attached to any label nor supervision. These maps show that the model automatically learns class-specific features leading to unsupervised object segmentations.

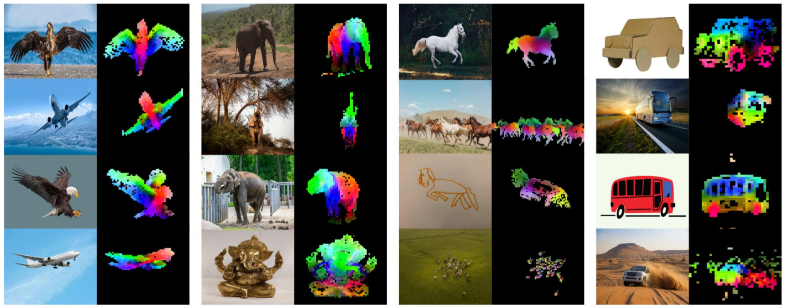
Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9650-9660).

DINO v2

	INet-1k k-NN	INet-1k linear
iBOT	72.9	82.3
+(our reproduction)	74.5 ↑ 1.6	83.2 ↑ 0.9
+LayerScale, Stochastic Depth	75.4 ↑ 0.9	82.0 ↓ 1.2
+128k prototypes	76.6 ↑ 1.2	81.9 ↓ 0.1
+KoLeo	78.9 ↑ 2.3	82.5 ↑ 0.6
+SwiGLU FFN	78.7 ↓ 0.2	83.1 ↑ 0.6
+Patch size 14	78.9 ↑ 0.2	83.5 ↑ 0.4
+Teacher momentum 0.994	79.4 ↑ 0.5	83.6 ↑ 0.1
+Tweak warmup schedules	80.5 ↑ 1.1	83.8 ↑ 0.2
+Batch size 3k	81.7 ↑ 1.2	84.7 ↑ 0.9
+Sinkhorn-Knopp	81.7 =	84.7 =
+Untying heads = DINOv2	82.0 ↑ 0.3	84.5 ↓ 0.2

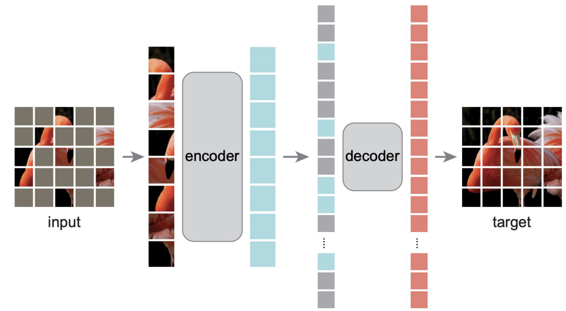
Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193.*

Cornell Bowers C·IS



Cornell Bowers C·IS

Masked Autoencoders (MaE)

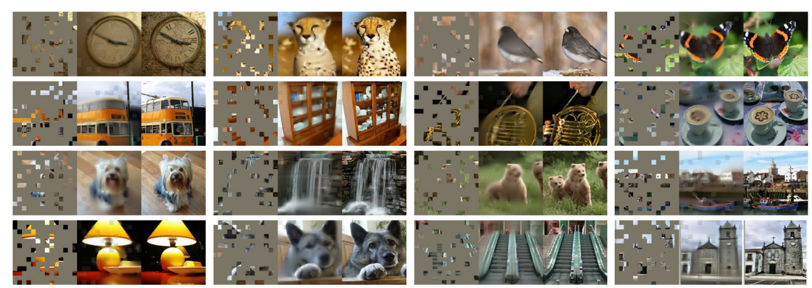


He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16000-16009).

Cornell Bowers C·IS

Discuss: BERT is trained with cross entropy loss. Can you do the same with MaE or should you use a different loss?

Cornell Bowers C·IS



Cornell Bowers C·IS

MaE Results

- Compared to supervised ViTs
  - Requires minimal data augmentation
  - Transfers better to downstream vision tasks
    - Object detection, segmentation

case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

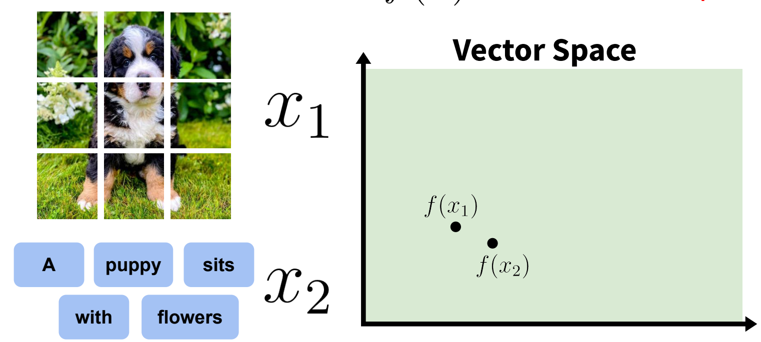
(e) **Data augmentation**. Our MAE works with minimal or no augmentation.

		AP^{box}		AP^{mask}	
method	pre-train data	ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	44.9	47.2

Table 4. **COCO object detection and segmentation** using a ViT Mask R-CNN baseline. All entries are based on our implementation. Self-supervised entries use IN1K data *without* labels. Mask AP follows a similar trend as box AP.

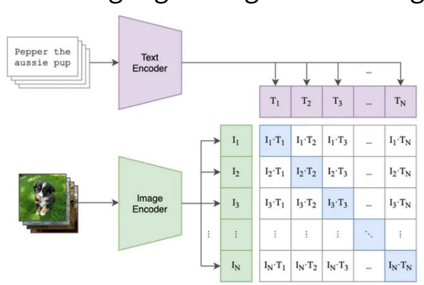
Cornell Bowers C·IS

$f(x)$ = transformer rep.



Cornell Bowers C·IS

CLIP (Contrastive Language–Image Pre-training)



Conde, M. V., & Turgutlu, K. (2021). CLIP-Art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3956-3960).

Cornell Bowers C·IS

Discuss: How can you train this model?

Cornell Bowers C·IS

```
# 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]
```

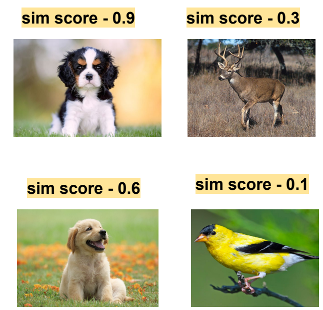
- Trained on 256 V100 GPUs for two weeks on 400 million (image, text pairs)
- On AWS, this would cost at least 200k dollars

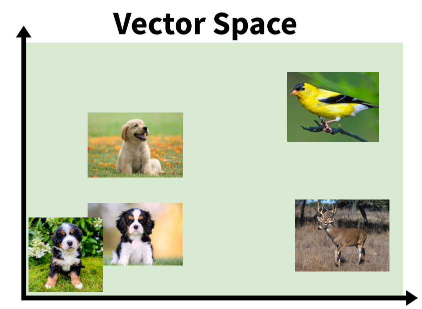
```
# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# 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
```

Cornell Bowers C·IS

Ranking using CLIP



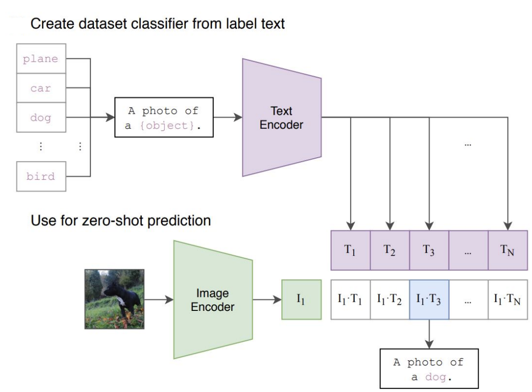


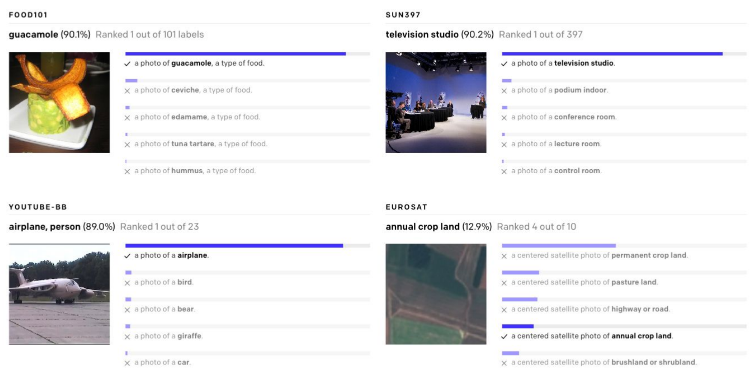
Cornell Bowers C·IS

Clip demo

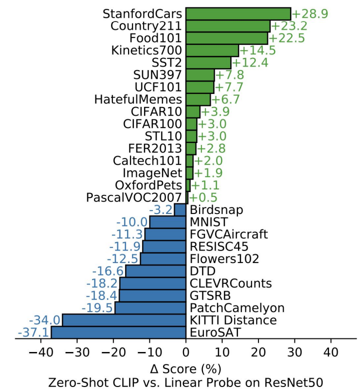
https://huggingface.co/spaces/vivien/clip

Cornell Bowers C·IS

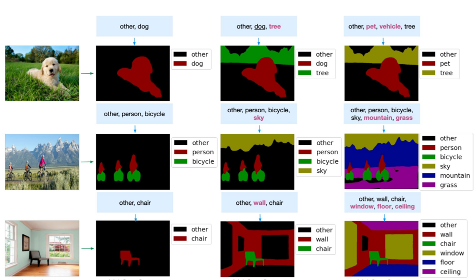




https://openai.com/research/clip

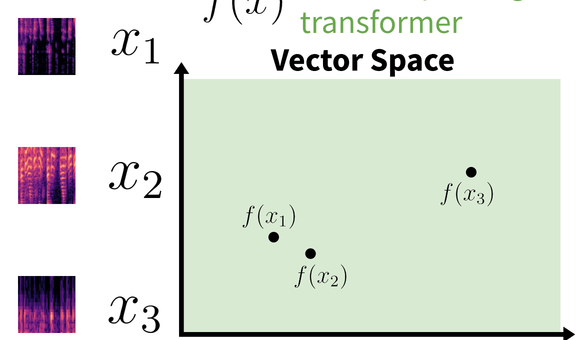


Application of CLIP



https://arxiv.org/pdf/2201.03546.pdf

$f(x)$ = audio spectrogram transformer

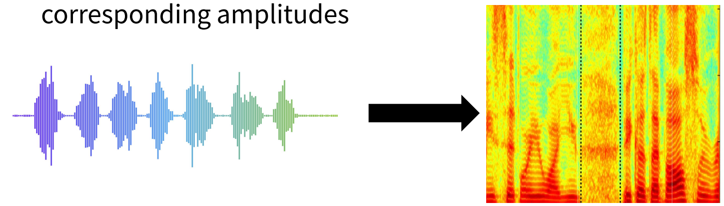


Audio Processing

Audio File

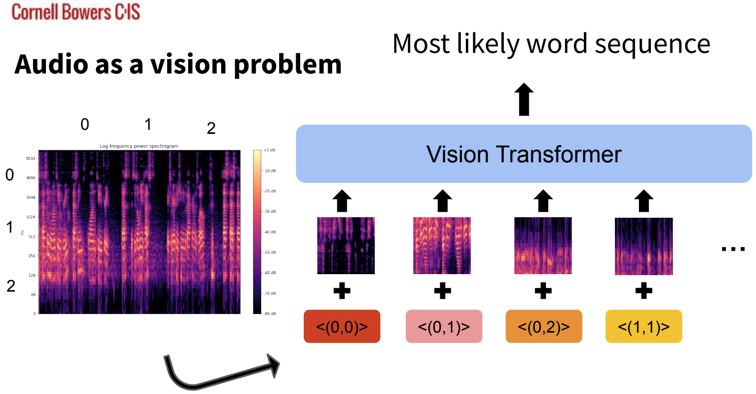
Spectrogram:

- Energy, pitch, fundamental frequency
- Decomposes signal into frequencies and their corresponding amplitudes



Spectrogram source: [Dumpala 2017]

Audio as a vision problem



Review

- Transformers can be used for vision tasks
- Swin transformers can be used for learning features at different scales
- Self-supervised learning is also helpful for transformer backbone vision models
    - Dino and MAE both learn very good embeddings
- Using transformer models for images and text helps build multi-modal models like CLIP