Self-Supervised Models of Vision + Language

CSCI 601 471/671
NLP: Self-Supervised Models

https://self-supervised.cs.jhu.edu/sp2023/
Logistics update

- We recommended few projects for revision. Due on Monday noon.
  - Please talk to us if you need more brainstorming to scope your proposal better.

- Each team is assigned to a mentor (we will email you about the assignment).
- Each team is responsible to schedule meeting time with their mentor.
- Our suggestion: meet with them at least once every 10 days.

- Starting next week: no TA/CA office hours.

- We will have external speakers next week!
Attention Is All You Need

- Transformers are eating the world

- SOTA performance across a range of NLP tasks
  - ...QA
  - ...summarization
  - ... paraphrasing

- Is Attention all you need?
  - Everything thus far has been text
  - What about other modalities, e.g., images?
  - What about multi-modal settings, e.g., captioned images?
Today's Talk

- Resources, datasets, & problems of interest
- Models & architectures
- Challenges & envisioned future
COCO, VQA, CLEVR, oh my

Common Objects In Context (2014)

- Object detection, segmentation, key-point detection, and captioning
- ~330K Images

[Microsoft COCO: Common Objects in Context, 2014]
COCO, VQA, CLEVR, oh my

Visual Question Answering (2017)

- Open-ended questions about images
- ~265K Images, at least 3 questions per image, "gold" answers

[VQA: Visual Question Answering, 2015]
COCO, VQA, CLEVR, oh my

CLEVER (2017)

- "strong biases that models can exploit to correctly answer questions without reasoning"
- ~100K Images
- ~865K Questions
  - Answers to all train/val questions

Q: Are there an equal number of large things and metal spheres?  
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere?  
Q: There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?  
Q: How many objects are either small cylinders or metal things?

[CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, 2016]
Let's Consider Images – How to Encode?
Let's Consider Images - Vision Transformer

* Extra learnable [class] embedding
Let's Consider Images - Vision Transformer
Let's Consider Images - Vision Transformer

Vision Transformer (ViT)

Transformer Encoder

Linear Projection of Flattened Patches

Patch + Position Embedding

* Extra learnable [class] embedding

Class
Bird
Ball
Car
...

MLP Head

[An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020]
What about paired image-text – **How to Encode?**

- **Basilica of St. John Lateran**
- **House with Chimeras – Kiev**
Contrastive Language–Image Pre-training (CLIP)

Given a batch of $N$ (image, text) pairs, predict which of the $N \times N$ possible (image, text) pairings across a batch occurred.

(1) Contrastive pre-training

- **Text Encoder**
- **Image Encoder**

### Code Snippet

```python
# 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 = 12normalize(np.dot(I_f, W_i), axis=1)
T_e = 12normalize(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
```

[Learning Transferable Visual Models From Natural Language Supervision, 2021]
Contrastive Language–Image Pre-training (CLIP)

Consider classifying photos of dogs vs cats - for each image, check if CLIP predicts text description “a photo of a dog” or “a photo of a cat” is more likely to be paired with it.

[Learning Transferable Visual Models From Natural Language Supervision, 2021]
What can CLIP do?

CLIP evaluates associations between image-text pairs:

- **Image Classification**
- **Image Searching**
- **Controllable Caption Generation**
What can't CLIP do?

- No generation capabilities
- Prompting / In-Context Learning
  - Few-shot captioning
Text-to-Image

“a corgi playing a flame throwing trumpet”
• First, a text prompt is input into a **text encoder** that is trained to map the prompt to a representation space.

[Zero-Shot Text-to-Image Generation, 2021]
DALL-E

- First, a text prompt is input into a **text encoder** that is trained to map the prompt to a representation space.

- Next, a **prior** maps the text encoding to a corresponding **image encoding** that captures the semantic information of the prompt contained in the text encoding.

[Zero-Shot Text-to-Image Generation, 2021]
DALL-E

- First, a text prompt is input into a **text encoder** that is trained to map the prompt to a representation space.

- Next, a **prior** maps the text encoding to a corresponding **image encoding** that captures the semantic information of the prompt contained in the text encoding.

- Finally, an **image decoder** stochastically generates an image which is a visual manifestation of this semantic information.
DALL-E

“a corgi playing a flame throwing trumpet”

[Zero-Shot Text-to-Image Generation, 2021]
Multimodal Prompting

- CLIP addresses ImageNet, but what if we introduce novel images / text / classes?

- DALL-E addresses controllable image generation, but still no ICL

- LLMs are transferrable to new tasks via prompting with examples

- Can we do the same thing in a multimodal setting, e.g. vision + language?
Multimodal Few-Shot Learning with Frozen Language Models

- First, consider a pretrained LLM
  - GPT-2-esque (7B)
  - Pretrained on C4

- Next, consider a vision encoder
  - Input images (e.g. WxHxC)
  - Output N tokens

Figure 2: Gradients through a frozen language model’s self attention layers are used to train the vision encoder.
Multimodal Few-Shot Learning with Frozen Language Models

Figure 3: Inference-Time interface for *Frozen*. The figure demonstrates how we can support (a) visual question answering, (b) outside-knowledge question answering and (c) few-shot image classification via in-context learning.
GPT-4
This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.
The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
Multimodal - RLHF

The input could be any modality!

(Improving Multimodal Interactive Agents with Reinforcement Learning from Human Feedback, 2022)
Challenges we still face

Models’ quality depend on frequency of patterns
Challenges we still face - The long tail

*Most things are infrequent*
Envisioning the future

- More modalities — combinations of video (2D, 3D), text, code, etc.
- Large models and more efficient scaling
- More breath — more data and more types of data
- Interaction with physical world — models with hands and actuators
- Better personalization — these agents should serve your needs
- Better human-machine teaming
  - CoPilot for coding
  - CoPilot for writing
  - ...
  - CoPilot for life!!!