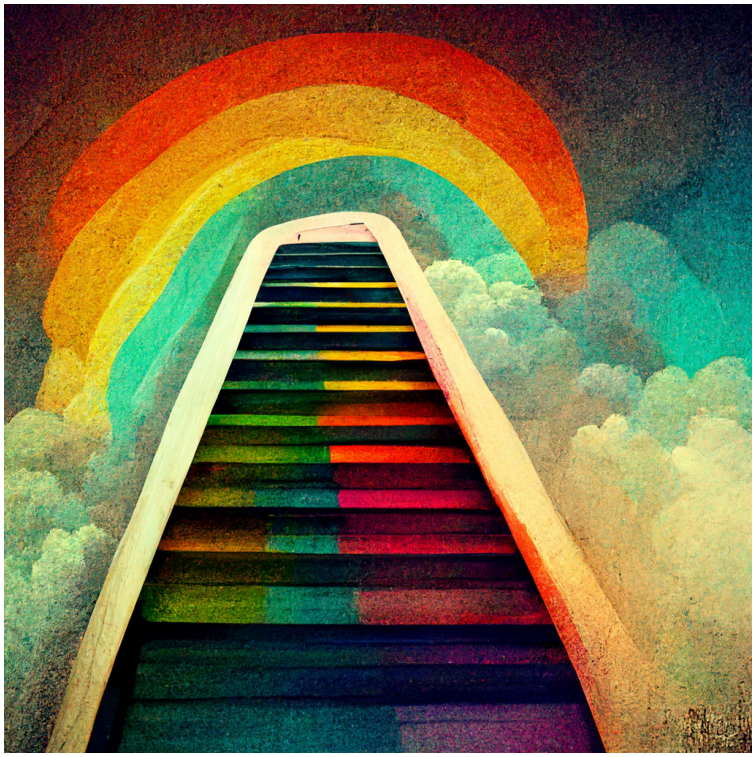
CS197 Harvard: AI Research Experiences

Fall 2022: Lecture 10 & 11 – "I Dreamed a Dream"
A Framework for Generating Research Ideas

Instructed by Pranav Rajpurkar. Website https://cs197.seas.harvard.edu/

Abstract

Coming up with good research ideas, especially when you're new to a field, is tough – it requires an understanding of gaps in literature. However, the process of generating research ideas can start after reading a single research paper. In this lecture, I share a set of frameworks with you to help you generate your own research ideas. First, you will learn to apply a framework to identify gaps in a research paper, including in the research question, experimental setup, and findings. Then, you will learn to apply a framework to generate ideas to build on a research paper, thinking about the elements of the task of interest, evaluation strategy and the proposed method. Finally, you will learn to apply a framework to iterate on your ideas to improve their quality. The lecture is structured such that you first prepare by reading two specified research papers, and we then apply these frameworks to the papers you have read.



*Midjourney Generation: "*a dream of climbing rainbow stairs"

Learning outcomes

- Identify gaps in a research paper, including in the research question, experimental setup, and findings.
- Generate ideas to build on a research paper, thinking about the elements of the task of interest, evaluation strategy and the proposed method.
- Iterate on your ideas to improve their quality.

Starter

Before we start, for this lecture, it is recommended that you read through CheXzero ("Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning") and CLIP ("Learning Transferable Visual Models From Natural Language Supervision") so that you can follow along with the examples referenced throughout the lecture. Refer to our previous notes on how to read a research paper.

A warm up exercise: Write your 3 best ideas for a follow up research paper you would publish to CLIP. After the lecture, come back to this exercise, and see how your answers have changed.

Identifying Gaps In A Research Paper

All research papers have gaps – gaps in the questions that were asked, in the way the experiments were set up, and in the way the paper in with prior work. These gaps often illuminate important directions for future research. I want to share with you some of the ways in which you can identify gaps in a research paper.

I've applied this framework to identifying gaps in the CheXzero paper.

1. Identify gaps in the research question

Write down the central research question of the paper. Then, write down the research hypothesis supporting that central research question. A research hypothesis is a "precise, testable statement of what the researcher(s) predict will be the outcome of the study." Not every hypothesis may be explicitly stated – you may have to infer this from the experiments that were performed. Now, you can look at gaps between the overall research question and the research hypotheses – what are hypotheses that have not been tested?

```
Example Answer:

Research Question:
How well can an algorithm detect diseases without explicit annotation?

Research Hypothesis:
   1. A self-supervised model trained on chest X-ray reports (CheXzero) can
      perform pathology-classification tasks with accuracies comparable to
      those of radiologists.
   2. CheXzero can outperform fully supervised models on pathology
      detection.
   3. CheXzero can outperform previous self-supervised approaches
```

```
      (MoCo-CXR, MedAug, and ConVIRT) on disease classification.

Gaps:
   1. Can CheXZero detect diseases that have never been implicitly seen in
      reports?
   2. Can CheXZero maintain high-level of performance even when using a
      small corpus of image-text reports.
```

2. Identify gaps in the experimental setups

Now that we have identified the research hypotheses, we can look at the experimental setup – here, we can pay attention to gaps. Are there shortcomings in the way the methods were evaluated? In the way the comparisons were chosen or implemented? Most importantly, does the experimental setup test the research hypothesis decisively? We're not looking at the results of the experiment, but in the setup of the experiment itself.

```
Example Answer:

Research Hypothesis (with Experimental Setups):
   1. A self-supervised model trained on chest X-ray reports (CheXzero) can
      perform pathology-classification tasks with accuracies comparable to
      those of radiologists.
         a. Evaluated on a test set of 500 studies from a single
            institution with a reference standard set by a majority vote -
            similar to what was used by previous studies. Comparison is
            performed to the average of 3 board-certified radiologists on
            the F1 and MCC metrics on 5 diseases. Gaps:
   2. CheXzero can outperform fully supervised models on pathology
      detection.
         a. Evaluated on the AUC metric on the average of 5 pathologies on
            the CheXpert test set (500 studies). Methods evaluated include
            a baseline supervised DenseNet121 model along with the DAM
            method with the reasoning "The DAM supervised method is
            included as a comparison and currently is state-of-the-art on
            the CheXpert dataset. An additional supervised baseline,
            DenseNet121, trained on the CheXpert dataset is included as a
            comparison since DenseNet121 is commonly used in
            self-supervised approaches."
   3. CheXzero can outperform previous self-supervised approaches
      (MoCo-CXR, MedAug, and ConVIRT) on disease classification.
         a. Setup as above.
```

```
Gaps:
   1. On hypothesis 1, The number of radiologists is maybe too small to
      decisively argue for being absolutely comparable to radiologists.
      Maybe the experience/training of the radiologists needs to be
      understood to qualify more precisely what constitutes radiologist
      level performance.
   2. On hypotheses 2/3, The number of pathologies evaluated for were
      limited by the number of samples in the test set. A larger set of
      pathologies evaluated would support the hypotheses more.
   3. On hypothesis 3, the number of self-supervised approaches compared to
      are limited – the choice of label-efficient approaches, ConVIRT,
      MedAug and MoCo-CXR. There are more self-supervised learning
      algorithms which can be compared to.
   4. On hypothesis 3, unclear also whether the comparisons are single
      models or ensemble models, or whether they use the same training
      source.
```

3. Identify gaps through expressed limitations, implicit and explicit

Now that we have identified gaps in the experimental setup, we make our way to the results and discussion. Here, we're on the lookout for expressed limitations of the work. Part of this work is easy: sometimes, there's an explicit limitation section that we can directly use, or we can infer it from statements of future work. However, sometimes the limitations of a method are expressed in the results themselves: where the methods fail.

```
Example Answer:

Gaps:

Explicitly Listed:
   1. "the self-supervised method still requires repeatedly querying
      performance on a labelled validation set for hyperparameter selection
      and to determine condition-specific probability thresholds when
      calculating MCC and F1 statistics.
   2. "the self-supervised method is currently limited to classifying image
      data; however, medical datasets often combine different imaging
      modalities, can incorporate non-imaging data from electronic health
      records or other sources, or can be a time series. For instance,
```

```
      magnetic resonance imaging and computed tomography produce
      three-dimensional data that have been used to train other
      machine-learning pipelines.
   3. “On the same note, it would be of interest to apply the method to
      other tasks in which medical data are paired with some form of
      unstructured text. For instance, the self-supervised method could
      leverage the availability of pathology reports that describe
      diagnoses such as cancer present in histopathology scans”
   4. “Lastly, future work should develop approaches to scale this method
      to larger image sizes to better classify smaller pathologies.”

Implicit through results:
   1. The model’s MCC performance is lower than radiologists on atelectasis
      and pleural effusion.
   2. The model’s AUC performance on Padchest is < 0.700 on 19 findings out
      of 57 radiographic findings where n > 50.
   3. The CheXzero method severely underperforms on detection of “No
      Finding” on Padchest, with an AUC of 0.755.
```

Exercise: Repeat the above application of the framework to identify gaps with CLIP.

Generating Ideas For Building on a Research Paper

Above, we have used a framework to identify gaps in a research paper. These gaps give us ideas for opportunities for improvement, but it may not always be clear how to tackle a gap. The following framework is designed to help you think about three axes on which you can build on a research paper. Again, we apply this framework to the CheXzero example.

1. Change the task of interest

- Can you apply the main ideas to a different modality?
    - Example: Pathology slides often have associated reports. Can you pair pathology slides with reports and do disease detection?
- Can you apply the main ideas to a different data type?
    - Example: Maybe the report doesn’t have to be text – maybe we can pair medical (e.g. pathology slide) images with available genomic alterations and perform similar contrastive learning.
- Can you apply the method or learned model to a different task?
    - Example: Maybe the CheXzero model could be applied to do object detection or semantic segmentation of images? Or maybe to medical image question answering.
- Can you change the outcome of interest?

→ Robustness or Data Efficiency

    - Example: Rather than accuracy, we can examine robustness properties of the CheXzero contrastive learning method. Or consider data efficiency of the method, or its performance on different patient subgroups compared to fully supervised methods.

2. Change the evaluation strategy

- Can you evaluate on a different dataset?
    - Example: CheXzero only considers CheXpert, MIMIC-CXR, and Padchest. However, there are other datasets that include very different types of patients or disease detection tasks, like the Shenzhen dataset which includes tuberculosis detection, or Ranzcr CLIP, which includes a line positioning task.
- Can you evaluate on a different metric?
    - Example:The AUC metric is used to evaluate the discriminative performance, but it doesn't give us insight into the calibration of the model (are the probability outputs reflective of the long-run proportion of disease outcomes), which could be measured by a calibration curve.
- Can you understand why something works well / breaks?
    - Example: It's unexplored whether there's a relationship between the frequency of disease-specific words occurring in the reports and performance on the different pathologies. This relationship could be empirically explored to explain the high-performance on some categories on padchest and low performance on others.
- Can you make different comparisons?
    - Example: There are many open comparisons we can address, including the comparison of radiologists to the model on Padchest, which would require the collection of further radiologist annotations.

3. Change the proposed method

(Caveat: This set of questions might best apply to deep learning method papers. However, I've found analog sets of questions in other research subdomains.)

- Can you change the training dataset or data elements?
    - Example: CheXzero trains on MIMIC-CXR, which is one of the few datasets that has both images and reports. A couple of things however which can change is that training could be augmented using IU-Xray dataset (OpenI), or the training can use another section of the radiology report (the findings section).
- Can you change the pre-training/training strategy?
    - Example: CheXZero leverages starting with a pre-trained OpenAI model, but there are newer checkpoints available that are trained on a larger dataset

(LAION-5B). In addition, there are training strategies that modify the loss functions including masked-language modeling in combination with the image-text contrastive losses, which are all areas of exploration for future work.

- Can you change the deep learning architecture?
    - Example: Rather than have a unimodal encoder for the image and text, a multimodal encoder could be used; this would take in both an image/image-embedding, and the text/text-embedding. This idea comes from advances in vision-language modeling/pretraining.
- Can you change the problem formulation?
    - Example: Right now, the CheXZero problem formulation is limited to take in one input, whereas typically a report can be paired with a set of more than one chest x-ray image. The formulation could thus be extended to take one or more available images (views) as input.

Exercise: Repeat the above application of the framework to identify ideas for extending CLIP.

Iterating on your research ideas

Ideas you come up with are going to get much better with iteration. Why might an idea not be a good idea? Reasons include: they might not be solving a real problem, they might already be published, and they might not be feasible. So how do we work with ideas to assess whether they are *good*?

1. Search for whether your idea has been tried:

It's possible your proposed new idea has already been tried, especially if the paper you're planning to build on is not recent. An exercise I do here to find out whether this is the case is to construct titles for your new paper ideas and see whether google comes up with a result. The key sometimes is to know multiple ways to refer to the same concept, which requires getting an understanding of related work.

```
Example: if I am interested in the application of a CheXzero-like approach
to other kinds of data, I might search for:
   -  contrastive learning histopathology text (no relevant results)
   -  contrastive learning histopathology genomic alteration (returned a
      match)
```

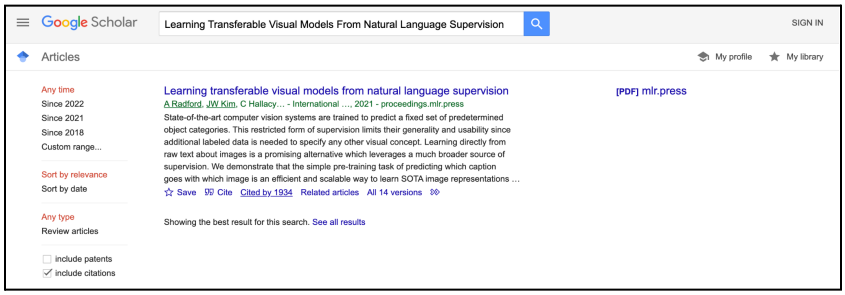
2. Read Important Related Works and Follow Up Works:

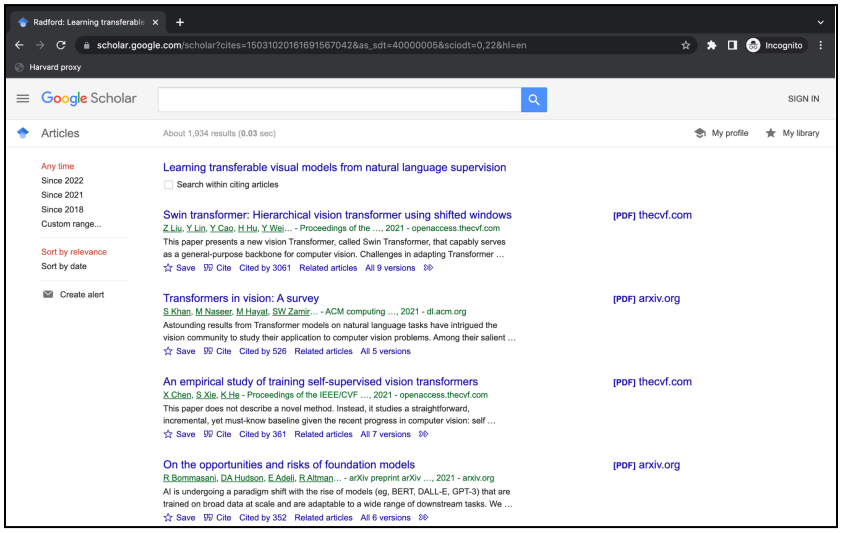
Often the related work or the discussion might explicitly specify alternative approaches that hold merit: make a list of these and start working your way through this list. You might benefit from reading through the paper that describes the creation of the dataset that your experiments will use.

If the paper you're building on has been around for long enough, you can find the papers that build on the work by using Google Scholar 'cited by', searching through abstracts on ArXiv, or searching explicitly for a task of interest to see the associated benchmark. Maintain a reading list like we used in Lecture 3. I think that good ideas will start reinforcing themselves as you read more papers in this reading list.

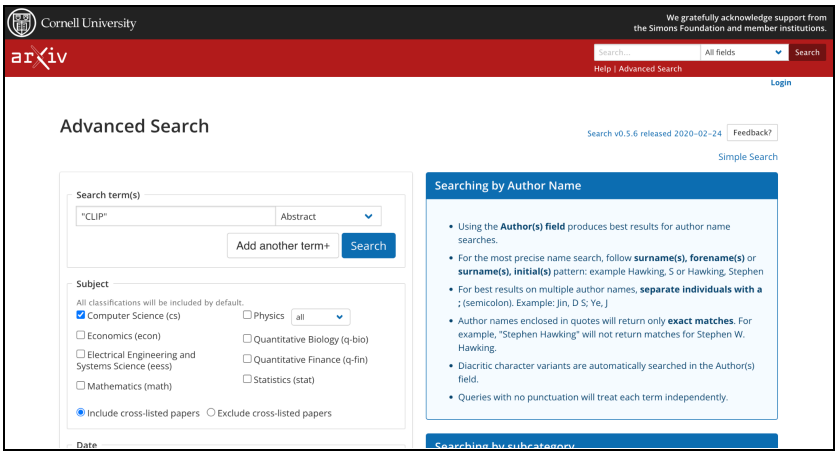
Example below shown for the CLIP paper:

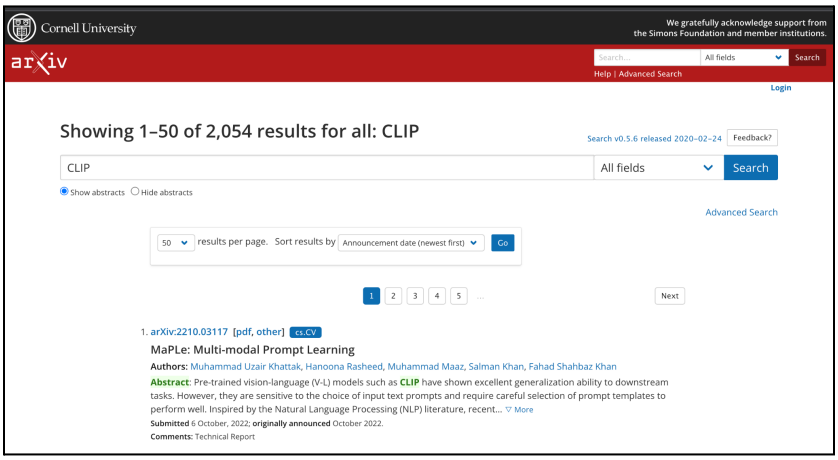
Google Scholar Cited By



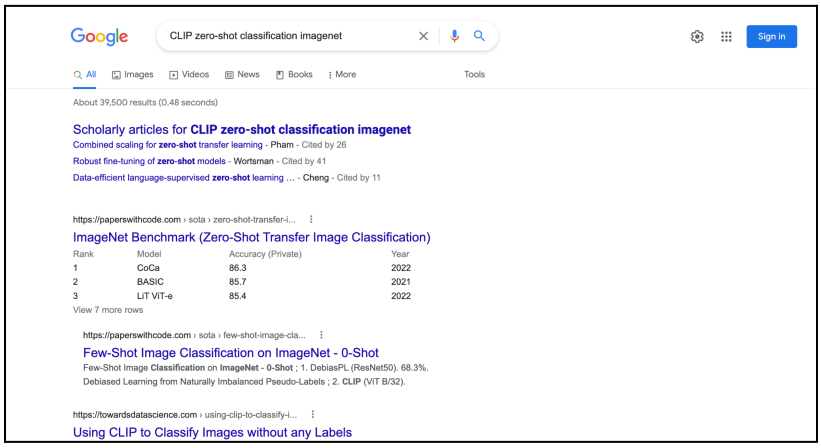


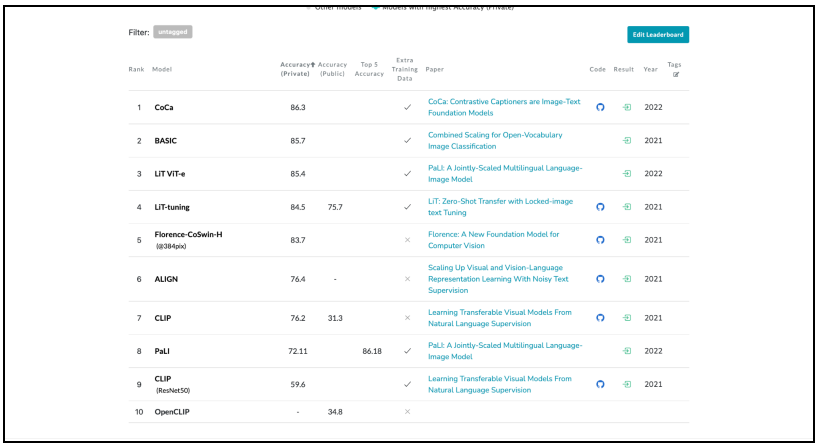
ArXiv Search





Google specific task





Filter: untagged — Edit Leaderboard

Rank	Model	Accuracy (Private)	Accuracy (Public)	Top 5 Accuracy	Extra Training Data	Paper	Code	Result	Year	Tags
1	CoCa	86.3			✓	CoCa: Contrastive Captioners are Image-Text Foundation Models			2022	
2	BASIC	85.7			✓	Combined Scaling for Open-Vocabulary Image Classification			2021	
3	LIT ViT-e	85.4			✓	PaLI: A Jointly-Scaled Multilingual Language-Image Model			2022	
4	LIT-tuning	84.5	75.7		✓	LiT: Zero-Shot Transfer with Locked-image text Tuning			2021	
5	Florence-CoSwin-H (@384pix)	83.7			×	Florence: A New Foundation Model for Computer Vision			2021	
6	ALIGN	76.4	-		×	Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision			2021	
7	CLIP	76.2	31.3		×	Learning Transferable Visual Models From Natural Language Supervision			2021	
8	PaLI	72.11		86.18	✓	PaLI: A Jointly-Scaled Multilingual Language-Image Model			2022	
9	CLIP (ResNet50)	59.6			✓	Learning Transferable Visual Models From Natural Language Supervision			2021	
10	OpenCLIP	-	34.8		×					

3. Get feedback from experts

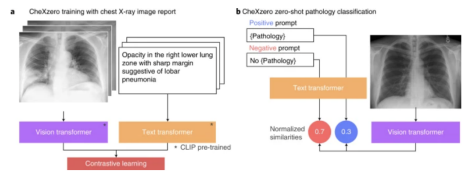
Once you have drafted up your idea in written form, I encourage you to try to get feedback from domain experts. You can write an email to the authors of the work that you're building on, sharing your idea and plan, and ask them what they think about your idea and approach. Sometimes, you might hear back from these experts (and sometimes you will not hear back, and that's okay; try reaching out to someone else)!

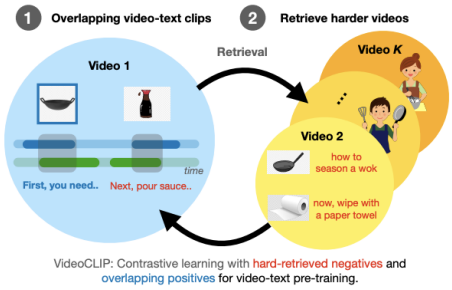
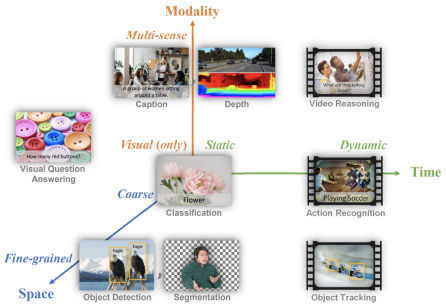
Exercise: Now, apply the Take your best idea for building on top of CLIP and google it. Write down below what you find.

Example: Framework in Action

Now that you've seen how you would begin to identify gaps, propose ideas and iterate on them, let's see how people have identified gaps in CLIP and built on top of them in the last 2 years.

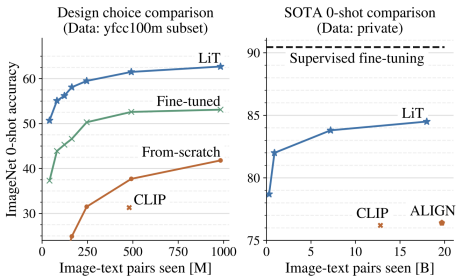
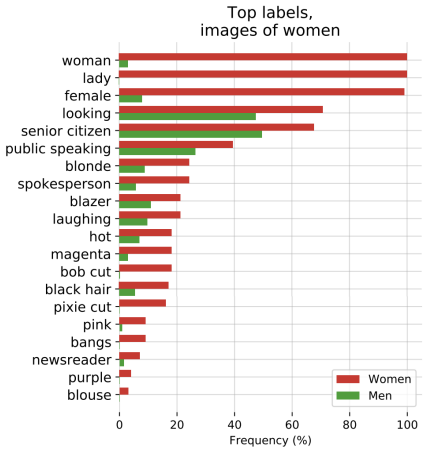
Change the task of interest

CheXZero Expert-level detection of pathologies from unannotated chest X-ray	- We demonstrated that we can leverage the pre-trained weights from the CLIP architecture learned from natural images to train a zero-shot model with a domain-specific medical task. - In contrast to CLIP, the proposed procedure allows us to normalize with respect to the negated	 **a**, Training pipeline. The model learns features from raw radiology reports, which act as a natural source of supervision. **b**, Prediction of pathologies in a chest X-ray image. For each pathology, we generated a positive and negative prompt (such as 'consolidation' versus 'no consolidation'). By comparing the model output for the positive and negative prompts, the self-supervised method computes a probability score for the pathology, and this can be used to classify its presence in the chest X-ray image.

images via self-supervised learning	version of the same disease classification instead of naively normalizing across the diseases to obtain probabilities from the logits	
VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding	- VideoCLIP trains a transformer for video and text by contrasting temporally overlapping positive video-text pairs with hard negatives from nearest neighbor retrieval. - Our effort aligns with the latter line of work [CLIP], but is the first to transfer a pre-trained discriminative model to a broad range of tasks in multi-modal video understanding.	 VideoCLIP: Contrastive learning with hard-retrieved negatives and overlapping positives for video-text pre-training. Figure 1: VideoCLIP aims for zero-shot video understanding via learning fine-grained association between video and text in a transformer using a contrastive objective with two key novelties: (1) for *positive* pairs, we use video and text clips that are *loosely* temporarily overlapping instead of enforcing strict start/end timestamp overlap; (2) for *negative* pairs, we employ a retrieval based sampling technique that uses video clusters to form batches with mutually harder videos.
Florence: A New Foundation Model for Computer Vision	- While existing vision foundation models such as CLIP (Radford et al., 2021) ... focus mainly on mapping images and textual representations to a cross-modal shared representation, we introduce a new computer vision foundation model, Florence, to expand the representations from coarse (scene) to fine (object), from static (images) to dynamic (videos), and from RGB to multiple modalities (caption, depth). - We extend the Florence pretrained model to learn finegrained (i.e. , object-level) representation, which is fundamental to dense prediction tasks such as object detection. - For this goal, we add an adaptor Dynamic Head...	
[your turn]	BASIC, LiT, ALBEF, PaLI, CoCa, Flava	

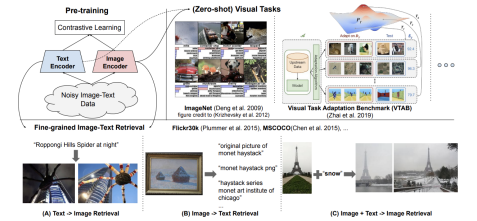
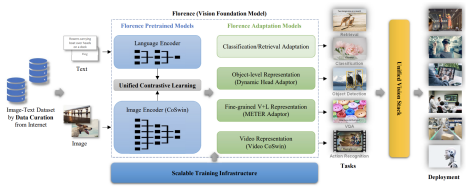
Exercise: Read through your selection of paper above. Share how it changed the task.

Change the evaluation strategy

LiT: Zero-Shot Transfer with Locked-image text Tuning	- We evaluate the resulting model's multilingualism in two ways, both of which have limitations discussed in Appendix J. First, we translate the ImageNet prompts into the most common languages using an online translation service and perform zero-shot classification in each of them... Second, we use the Wikipedia based Image Text (WIT) dataset [54] to perform T → I retrieval across more than a hundred languages.	
Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications	• First, we find that the way classes are designed can heavily influence model performance when deployed, pointing to the need to provide users with education about how to design classes carefully. Second, we find that CLIP can unlock certain niche tasks with greater ease, given that CLIP can often perform surprisingly well without task-specific training data. • When we studied the performance of ZS CLIP on 'in the wild' celebrity identification using the CelebA dataset...we found that the model had 59.2% top-1 accuracy out of 100 possible classes for 'in the wild' 8k celebrity images. However, this performance dropped to 43.3% when we increased our class sizes to 1k celebrity names.	
[your turn]	BASIC, ALBEF, PaLI, CoCa, Flava, Florence	

Exercise: Read through your selection of paper above. Share how it changed the evaluation.

Change the proposed method

ALIGN (Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision)	- We leverage a noisy dataset of over one billion image alt-text pairs, obtained without expensive filtering or post-processing steps in the Conceptual Captions dataset. - ALIGN follows the natural distribution of image-text pairs from the raw alt-text data, while CLIP collects the dataset by first constructing an allowlist of high-frequency visual concepts from English Wikipedia.	 *Figure 1.* A summary of our method, ALIGN. Visual and language representations are jointly learned from noisy image alt-text data. The representations can be used for vision-only or vision-language task transfer. Without any fine-tuning, ALIGN powers zero-shot visual classification and cross-modal search including image-to-text search, text-to-image search and even search with joint image+text queries.
Florence: A New Foundation Model for Computer Vision	- Also a task difference (so repeated from above) - Our Florence pretrained model uses a two-tower architecture: a 12-layer transformer (Vaswani et al., 2017) as language encoder, similar to CLIP (Radford et al., 2021), and a hierarchical Vision Transformer as the image encoder. The hierarchical Vision Transformer is a modified Swin Transformer (Liu et al., 2021a) with convolutional embedding, called CoSwin Transformer.	
[your turn]	BASIC, LiT, ALBEF, PaLI, CoCa, Flava	

Exercise: Read through your selection of paper above. Share how it changed the proposed method.