

# A framework for generating research ideas

Maximiliano Casas

19 de octubre de 2022

## Índice

<b>1. Identifying gaps in a research paper</b>	<b>1</b>
<b>2. Generating ideas for building on a research paper</b>	<b>3</b>
<b>3. Iterating on your research ideas</b>	<b>4</b>
<b>4. Examples</b>	<b>5</b>
4.1. Change the task of interest . . . . .	5
4.2. Change the evaluation strategy . . . . .	5
4.3. Change the proposed method . . . . .	5
1. Learn to apply a framework to identify gaps in a research paper, including in the research question, experimental setup and findings	
2. 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	
3. Learn to apply a framework to iterate on your ideas to improve their quality	

## 1. Identifying gaps in a research paper

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.

**Def.** A research hypothesis is a precise, testable statement of what the researchers predict will be the outcome of the study. Not every hypothesis may be explicitly stated. What are hypotheses that have not been tested?

With the CheXero paper

1. Identify gaps in the research question

Example answer:

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

Research hypothesis:

- a) A self-supervised model trained on chest X-ray reports (CheXzero) can perform pathology-classification task with accuracies comparable to those of radiologists
- b) CheXzero can outperform fully supervised models on pathology detection
- c) CheXzero can outperform previous self-supervised approaches (MoCo-CXR, MedAug, and Con-VIRT) on disease classification

Gaps:

- a) Can CheXzero detect diseases that have never been implicitly seen in reports?
  - b) Can CheXzero maintain high-level of performance even when using a small corpus of image-text reports
2. Identify gaps in the experimental setups: Are there shortcomings in the way the methods were evaluated? In the way the comparisons were chosen or implemented? Does the experimental setup test the research hypothesis decisively?

Example answer:

Research hypothesis (with experimental setups):

- a) A self-supervised model trained on chest X-ray reports (CheXzero) can perform pathology-classification task with accuracies comparable to those of radiologists  
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 on the average of 3 board-certified radiologists in the F1 and MCC metrics on 5 diseases.
- b) CheXzero can outperform fully supervised models on pathology detection  
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 that 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.
- c) CheXzero can outperform previous self-supervised approaches (MoCo-CXR, MedAug, and Con-VIRT) on disease classification  
Setup as above.

Gaps:

- a) 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.
- b) 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.
- c) 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.
- d) 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 limitation, implicit and explicit: Results and discussion. We're on the lookout of for expressed limitation of the work. Sometimes the limitations of a method are expressed in the results themselves: where the methods fail.

Example answer:

Gaps:

Explicitly listed:

- a) The self-supervised method still requires repeatedly querying performance on a labeled validation set for hyper-parameter selection and to determine condition-specific probability thresholds when calculating MCC and F1 statistics
- b) 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
- c) 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 taht describe diagnoses such as cancer present in histopathology scans
- d) Lastly, future work should develop approaches to scale this method to larger image sizes to better classify smaller pathologies

Implicit through results:

- a) The model's MCC performance is lower than radiologists on atelectasis and pleural effusion.
- b) The model's AUC performance on Padchest is less than 0.700 on 19 findings out of 57 radiographic findings where n greater than 50.
- c) The CheXzero method severely underperforms on detection of "No Finding" on Padchest, with an AUC of 0.755.

## 2. Generating ideas for building on a research paper

### 1. Change the task of interest

- Can you apply the main ideas to a different modality? Example: Pathology slides often have associated reportes. 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? 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 of 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

- 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.

## 3. 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? They might not be solving a real problem, they might already be published, and they might not be feasible.

1. Search for whether your ideas has been tried: Construct titles for your new paper ideas and see whether google comes up iwth a result. They key 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)

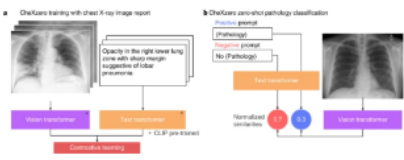
2. Read important related works and follow up works: Often the related work or the discussion might explicitly specify alternative approaches that hold merit: make a list of them. Read through the paper that describes the creation of the dataset that your experiments will use

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. Good ideas will start reinforcing themselves as you read more papers.

3. Get feedback from experts: 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.

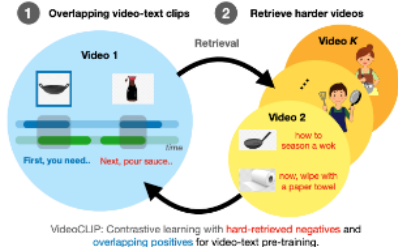

## 4. Examples

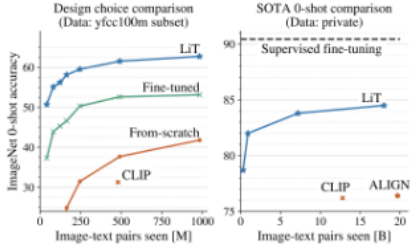
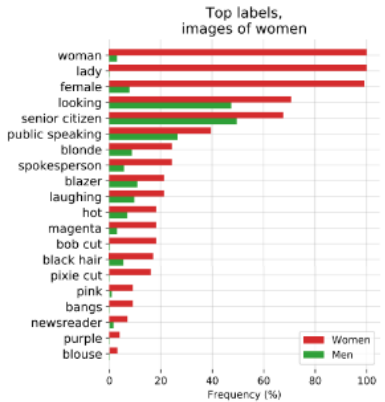
### 4.1. Change the task of interest


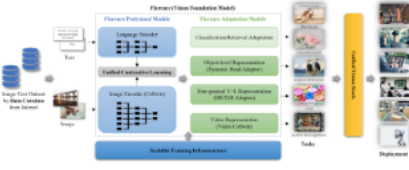
<p><a href="#">CheXZero</a> Expert-level detection of pathologies from unannotated chest X-ray</p>	<ul style="list-style-type: none"> <li>- We demonstrated that we can leverage the pre-trained weights from the CLIP architecture learned from natural images to train a zero-shot model with a domain-specific medical task.</li> <li>- In contrast to CLIP, the proposed procedure allows us to normalize with respect to the negated</li> </ul>	 <p><b>a. Training pipeline.</b> The model learns features from raw radiology reports, which act as a natural source of supervision. <b>b. Prediction of pathologies in a chest X-ray image.</b> 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.</p>
--	---	---

### 4.2. Change the evaluation strategy

### 4.3. Change the proposed method

images via self-supervised learning	version of the same disease classification instead of naively normalizing across the diseases to obtain probabilities from the logits	
<a href="#">VideoCLIP</a> : Contrastive Pre-training for Zero-shot Video-Text Understanding	<ul style="list-style-type: none"> <li>- VideoCLIP trains a transformer for video and text by contrasting temporally overlapping positive video-text pairs with hard negatives from nearest neighbor retrieval.</li> <li>- 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.</li> </ul>	 <p>VideoCLIP: Contrastive learning with <b>hard-retrieved negatives</b> and <b>overlapping positives</b> for video-text pre-training.</p> <p>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 <i>positive</i> pairs, we use video and text clips that are <i>loosely</i> temporarily overlapping instead of enforcing strict start/end timestamp overlap; (2) for <i>negative</i> pairs, we employ a retrieval based sampling technique that uses video clusters to form batches with mutually harder videos.</p>
<a href="#">Florence</a> : A New Foundation Model for Computer Vision	<ul style="list-style-type: none"> <li>- 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).</li> <li>- 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.</li> <li>- For this goal, we add an adaptor Dynamic Head...</li> </ul>	
[your turn]	BASIC, LiT, ALBEF, PaLI, CoCa, Flava	

<p><a href="#">LiT</a>: Zero-Shot Transfer with Locked-image text Tuning</p>	<ul style="list-style-type: none"> <li>- 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.</li> </ul>	 <p>Design choice comparison (Data: yfcc100m subset)</p> <p>SOTA 0-shot comparison (Data: private)</p> <p>Supervised fine-tuning</p>
<p><a href="#">Evaluating CLIP</a>: Towards Characterization of Broader Capabilities and Downstream Implications</p>	<ul style="list-style-type: none"> <li>• 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.</li> <li>• 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.</li> </ul>	 <p>Top labels, images of women</p> <p>Frequency (%)</p> <p>Women Men</p>

<p><b><a href="#">ALIGN</a></b> (Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision)</p>	<ul style="list-style-type: none"> <li>- 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.</li> <li>- 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.</li> </ul>	 <p>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 tasks. 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.</p>
<p><b><a href="#">Florence</a></b>: A New Foundation Model for Computer Vision</p>	<ul style="list-style-type: none"> <li>- Also a task difference (so repeated from above)</li> <li>- 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.</li> </ul>	
<p>[your turn]</p>	<p>BASIC, LiT, ALBEF, PaLI, CoCa, Flava</p>	