The University Of Queensland AUSTRALIA | CREATE CHANGE

AI Research Experiences: A Framework for Generating Research Ideas

(Shamelessly based on Pranav Rajpurkar's Harvard CS197 course)

Guido Zuccon

g.zuccon@uq.edu.au
ielab, The University of Queensland, Australia
www.ielab.io

ie lab

What this is about

- Coming up with good research ideas, especially when you're new to a field, is tough
- it requires an understanding of gaps in literature.
- the process of generating research ideas can start after reading a single research paper
- Here: a framework to help you generate your own research ideas

nature biomedical engineering

ARTICLES

https://doi.org/10.1038/s41551-022-00936-9

OPEN

Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning

Ekin Tiu[1,2,4], Ellie Talius[1,2,4], Pujan Patel[1,2,4], Curtis P. Langlotz[3], Andrew Y. Ng[1] and Pranav Rajpurkar[2]

In tasks involving the interpretation of medical images, suitably trained machine-learning models often exceed the performance of medical experts. Yet such a high-level of performance typically requires that the models be trained with relevant datasets that have been painstakingly annotated by experts. Here we show that a self-supervised model trained on chest X-ray images that lack explicit annotations performs pathology-classification tasks with accuracies comparable to those of radiologists. On an external validation dataset of chest X-rays, the self-supervised model outperformed a fully supervised model in the detection of three pathologies (out of eight), and the performance generalized to pathologies that were not explicitly annotated for model training, to multiple image-interpretation tasks and to datasets from multiple institutions.

Identifying Gaps In A Research Paper

1. Identify gaps in the research question
2. Identify gaps in the experimental setups
3. Identify gaps through expressed limitations, implicit and explicit

1. Identify gaps in the research question

1. Write down the central research question of the paper

nature biomedical engineering

ARTICLES

https://doi.org/10.1038/s41551-022-00936-9

OPEN

Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning

Ekin Tiu[1,2,4], Ellie Talius[1,2,4], Pujan Patel[1,2,4], Curtis P. Langlotz[3], Andrew Y. Ng[1] and Pranav Rajpurkar[2]

In tasks involving the interpretation of medical images, suitably trained machine-learning models often exceed the performance of medical experts. Yet such a high-level of performance typically requires that the models be trained with relevant datasets that have been painstakingly annotated by experts. Here we show that a self-supervised model trained on chest X-ray images that lack explicit annotations performs pathology-classification tasks with accuracies comparable to those of radiologists. On an external validation dataset of chest X-rays, the self-supervised model outperformed a fully supervised model in the detection of three pathologies (out of eight), and the performance generalized to pathologies that were not explicitly annotated for model training, to multiple image-interpretation tasks and to datasets from multiple institutions.

1. Identify gaps in the research question

1. Write down the central research question of the paper

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

1. Identify gaps in the research question

1. Write down the central research question of the paper

2. Write down the research hypothesis supporting that central research question

   - research hypothesis: “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

nature biomedical engineering

ARTICLES

https://doi.org/10.1038/s41551-022-00936-9

OPEN

Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning

Ekin Tiu[1,2,4], Ellie Talius[1,2,4], Pujan Patel[1,2,4], Curtis P. Langlotz[3], Andrew Y. Ng[1] and Pranav Rajpurkar[2]

In tasks involving the interpretation of medical images, suitably trained machine-learning models often exceed the performance of medical experts. Yet such a high-level of performance typically requires that the models be trained with relevant datasets that have been painstakingly annotated by experts. Here we show that a self-supervised model trained on chest X-ray images that lack explicit annotations performs pathology-classification tasks with accuracies comparable to those of radiologists. On an external validation dataset of chest X-rays, the self-supervised model outperformed a fully supervised model in the detection of three pathologies (out of eight), and the performance generalized to pathologies that were not explicitly annotated for model training, to multiple image-interpretation tasks and to datasets from multiple institutions.

1. Identify gaps in the research question

1. Write down the central research question of the paper

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

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

1. Identify gaps in the research question

1. Write down the central research question of the paper

2. Write down the research hypothesis supporting that central research question

   - research hypothesis: “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

3. look at gaps between the overall research question and the research hypotheses – what are hypotheses that have not been tested?

1. Identify gaps in the research question

1. Write down the central research question of the paper

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

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

Gaps:

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

2. Identify gaps in the experimental setups

1. Are there shortcomings in the way the methods were evaluated?

2. In the way the comparisons were chosen or implemented?

3. 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.

Research Hypothesis (with Experimental Setups):

Hyp: *A self-supervised model trained on chest X-ray reports (CheXzero) can perform pathology-classification tasks 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 to the average of 3 board-certified radiologists on the F1 and MCC metrics on 5 diseases.

Gap?

Research Hypothesis (with Experimental Setups):

Hyp: *A self-supervised model trained on chest X-ray reports (CheXzero) can perform pathology-classification tasks 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 to the average of 3 board-certified radiologists on the F1 and MCC metrics on 5 diseases.

Gap?

- 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.

Research Hypothesis (with Experimental Setups):

Hyp: *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 “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.”

Gap?

Research Hypothesis (with Experimental Setups):

Hyp: *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 “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.”

Gap?

- 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.

Research Hypothesis (with Experimental Setups):

Hyp: *CheXzero can outperform previous self-supervised approaches (MoCo-CXR, MedAug, and ConVIRT) on disease classification.*

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.”

Research Hypothesis (with Experimental Setups):

Hyp: *CheXzero can outperform previous self-supervised approaches (MoCo-CXR, MedAug, and ConVIRT) on disease classification.*

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.”

Gap?

- 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.
- 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.
- 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

- Look at results, discussion/limitations/conclusions
- lookout for expressed limitations of the work
- sometimes, there’s an explicit limitation section that we can directly use, or we can infer it from statements of future work.
- sometimes the limitations of a method are expressed in the results themselves: where the methods fail.

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.

Explicitly Listed:

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.

Generating Ideas For Building on a Research Paper

1. Change the task of interest
2. Change the evaluation strategy
3. Change the proposed method

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**?
  - 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

- 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.

Iterating on your research ideas

- **Search for whether your idea has been tried**
  - Strategy: 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
- Read Important Related Works and Follow Up Works
- Get feedback from experts

Iterating on your research ideas

- Search for whether your idea has been tried
- **Read Important Related Works and Follow Up Works**
  - make a list of alternative approaches mentioned in Related Works and start working your way through this list.
  - read through the paper that describes the creation of the dataset
  - find the papers that build on the work you are examing (Google Scholar “cited by” function)
- Get feedback from experts

Iterating on your research ideas

- Search for whether your idea has been tried
- Read Important Related Works and Follow Up Works
- **Get feedback from experts**

Examples of how papers have built on gaps

- Paper with gaps: CLIP — “Learning Transferable Visual Models From Natural Language Supervision

- CheXZero: Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning

  - 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 version of the same disease classification instead of naively normalizing across the diseases to obtain probabilities from the logits

Examples of how papers have built on gaps

- Paper with gaps: CLIP — “Learning Transferable Visual Models From Natural Language Supervision

- 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.

Examples of how papers have built on gaps

- Paper with gaps: CLIP — “Learning Transferable Visual Models From Natural Language Supervision
- 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...