HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Read our paper here

beamlab

Towards Reliable Zero Shot Classification in Self-Supervised Models with Conformal Prediction

Bhawesh Kumar*, Anil Palepu*, Rudraksh Tuwani, Dr. Andrew Beam

Introduction:

- Self-supervised models trained with a contrastive loss like CLIP can be used for zero-shot classification. However, they require users to come up with captions for classification labels. It is not always easy to know if we have come up with good caption.
- The reliability of zero-shot classification is dependent on the quality of the written caption. We propose Conformal Prediction (CP) framework to assess when a given test caption may be reliably used for zero-shot classification task .

Data:

- Data: MIMIC_CXR
- Clinical Labels: Cardiomegaly, Consolidation, Edema, Pleural Effusion
- Captions: Extracted from the “findings” and “impression” sections of the radiology report for training CLIP models. Query captions for the four labels are obtained from a prior work.

Methods:

- We train two CLIP models: “Findings+Impressions” trained on concatenation of findings and impression sections of the radiology report and “Impressions only” trained on only impression section of the report.
- We propose two novel conformal scores: The first score is essentially a classifier controlling for TPR and second score measures the quality of the captions.
- The first non-conformity score is the cosine distance between a paired caption and image and measures the compatibility between them. The second non-conformity score measures the mean KNN distance (KNND) between a caption x and the set of captions X used to train the model.

Evaluations:

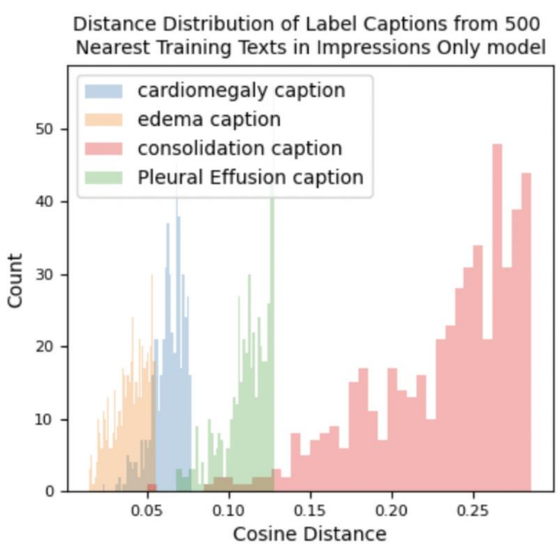
- We evaluate reliability of zero-shot classification task in two ways. Deviation from conformal error rate in zero-shot classification and conformal converge rate required to admit a query caption.
- A higher deviation from conformal error rate means a less-reliable classification task. Further, a higher conformal converge rate required to admit a query caption indicates that query caption for that label is less similar to the captions in training and calibration set.

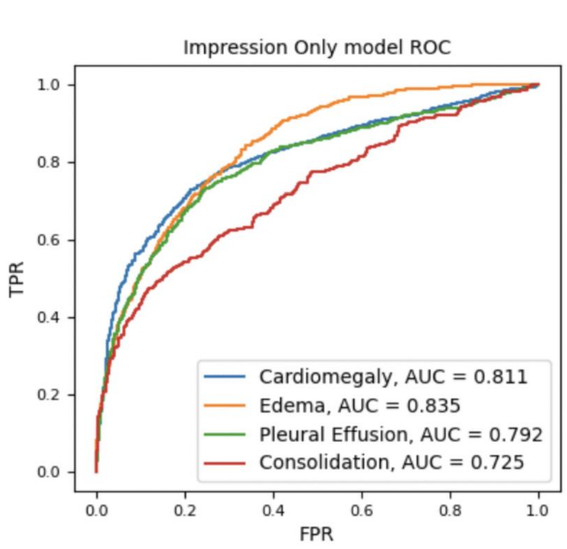
Results:

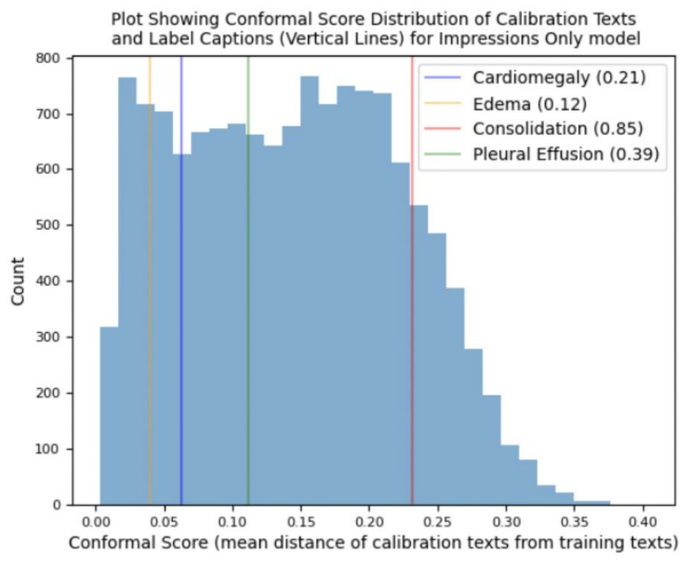
- Neither model achieves the advertised TPR in CP procedure. However, impressions-only model is closer to the correct error rate than the Findings+Impressions model since impression-only model is more similar to the query caption for labels.
- Further, images with single diagnosis achieve better conformal guarantee than images with multiple diagnosis.
- We also find that larger mean KNND from train dataset is associated with poorer performance on zero-shot classification task.

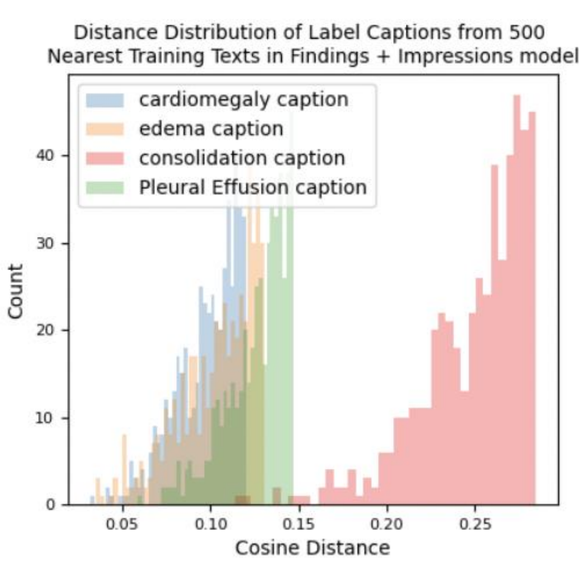
Conclusions:

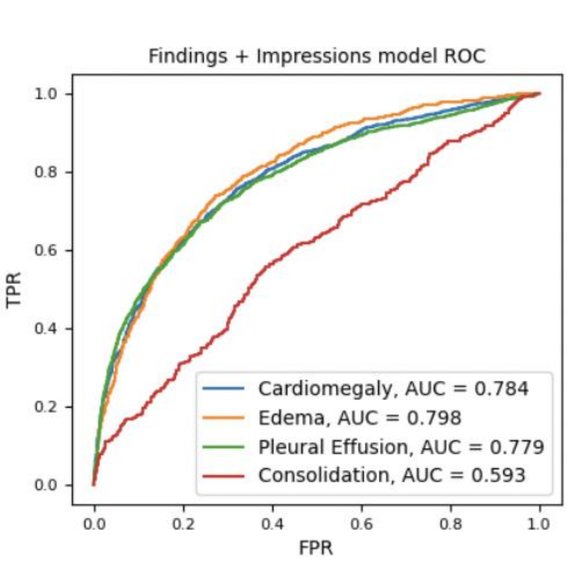
- The CP procedure doesn’t achieve the guaranteed coverage rate due to exchangeability violation in test set caption. The extent of deviation is dependent on texts used to train the model as well as the presence of additional findings in test samples.
- For the KNND non-conformity score, a higher coverage rate required to admit a caption in the CP procedure is predictive of poor AUCROC score in classification task. Thus, our approach provides a useful technique for quantifying quality of captions in zero-shot classification task.

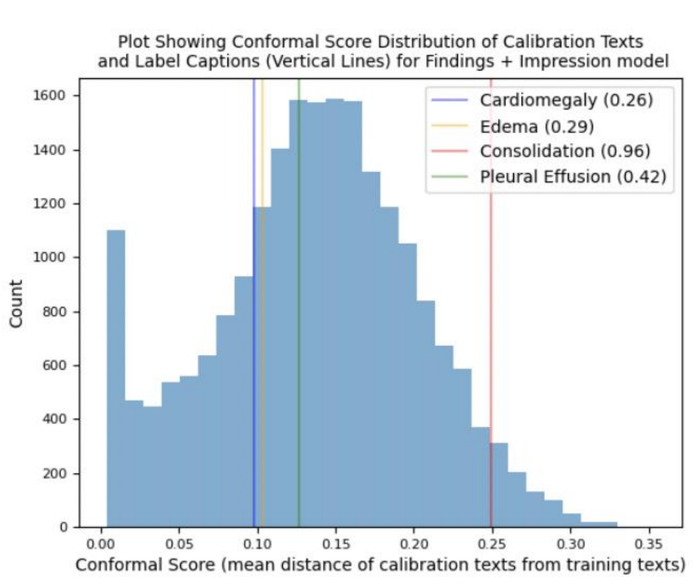












Plots above show histograms of cosine distances from each label caption embedding to their 500 nearest text embeddings in the train set in the two models that we train. On the right are the ROC curves for these label captions corresponding to the zero-shot classification on the test set. Label captions with higher KNN Distance have poorer performance on zero-shot classification as measured by the AUROC score.

Plots above show the conformal score distribution of calibration captions computed from the mean 500 Nearest Neighbor distance from training texts in our two models. The vertical lines correspond to the score for the four label captions. The number in the bracket beside the label names in the plot indicate the coverage rate required to admit the caption corresponding to that label.