



# HHS Public Access

Author manuscript

*Med Image Comput Assist Interv.* Author manuscript; available in PMC 2021 October 01.

Published in final edited form as:

*Med Image Comput Assist Interv.* 2020 October ; 12262: 529–539.

doi:10.1007/978-3-030-59713-9\_51.

## Joint Modeling of Chest Radiographs and Radiology Reports for Pulmonary Edema Assessment

Geeticka Chauhan<sup>1</sup>, Ruizhi Liao<sup>1</sup>, William Wells<sup>1,2</sup>, Jacob Andreas<sup>1</sup>, Xin Wang<sup>3</sup>, Seth Berkowitz<sup>4</sup>, Steven Horng<sup>4</sup>, Peter Szolovits<sup>1</sup>, Polina Golland<sup>1</sup>

<sup>1</sup>Massachusetts Institute of Technology, Cambridge, MA, USA

<sup>2</sup>Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

<sup>3</sup>Philips Research North America, Cambridge, MA, USA

<sup>4</sup>Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

### Abstract

We propose and demonstrate a novel machine learning algorithm that assesses pulmonary edema severity from chest radiographs. While large publicly available datasets of chest radiographs and free-text radiology reports exist, only limited numerical edema severity labels can be extracted from radiology reports. This is a significant challenge in learning such models for image classification. To take advantage of the rich information present in the radiology reports, we develop a neural network model that is trained on both images and free-text to assess pulmonary edema severity from chest radiographs at inference time. Our experimental results suggest that the joint image-text representation learning improves the performance of pulmonary edema assessment compared to a supervised model trained on images only. We also show the use of the text for explaining the image classification by the joint model. To the best of our knowledge, our approach is the first to leverage free-text radiology reports for improving the image model performance in this application. Our code is available at: [https://github.com/RayRuizhiLiao/joint\\_chestxray](https://github.com/RayRuizhiLiao/joint_chestxray).

### Introduction

We present a novel approach to training machine learning models for assessing pulmonary edema severity from chest radiographs by jointly learning representations from the images (chest radiographs) and their associated radiology reports. Pulmonary edema is the most common reason patients with acute congestive heart failure (CHF) seek care in hospitals [1,9,15]. The treatment success in acute CHF cases depends crucially on effective management of patient fluid status, which in turn requires pulmonary edema quantification, rather than detecting its mere absence or presence.

Chest radiographs are commonly acquired to assess pulmonary edema in routine clinical practice. Radiology reports capture radiologists' impressions of the edema severity in the

---

Co-first authors

The numbers of images of the four severity levels are 2883, 1511, 1709, and 640 respectively.

form of unstructured text. While the chest radiographs possess ground-truth information about the disease, they are often time intensive (and therefore expensive) for manual labeling. Therefore, labels extracted from reports are used as a proxy for ground-truth image labels. Only limited numerical edema severity labels can be extracted from the reports, which limits the amount of labeled image data we can learn from. This presents a significant challenge for learning accurate image-based models for edema assessment. To improve the performance of the image-based model and allow leveraging larger amount of training data, we make use of free-text reports to include rich information about radiographic findings and reasoning of pathology assessment. We incorporate free-text information associated with the images by including them during our training process.

We propose a neural network model that jointly learns from images and free-text to quantify pulmonary edema severity from images (chest radiographs). At training time, the model learns from a large number of chest radiographs and their associated radiology reports, with a limited number of numerical edema severity labels. At inference time, the model computes edema severity given the input image. While the model can also make predictions from reports, our main interest is to leverage free-text information during training to improve the accuracy of image-based inference. Compared to prior work in the image-text domain that fuses image and text features [5], our goal is to decouple the two modalities during inference to construct an accurate image-based model.

Prior work in assessing pulmonary edema severity from chest radiographs has focused on using image data only [14,18]. To the best of our knowledge, ours is the first method to leverage the free-text radiology reports for improving the image model performance in this application. Our experimental results demonstrate that the joint representation learning framework improves the accuracy of edema severity estimates over a purely image-based model on a fully labeled subset of the data (supervised). The joint learning framework uses a ranking-based criterion [7,12], allowing for training the model on a larger dataset of unlabeled images and reports. This semi-supervised modification demonstrates a further improvement in accuracy. Additional advantages of our joint learning framework are 1) allowing for the image and text models to be decoupled at inference time, and 2) providing textual explanations for image classification in the form of saliency highlights in the radiology reports.

### **Related Work.**

The ability of neural networks to learn effective feature representations from images and text has catalyzed the recent surge of interest in joint image-text modeling. In supervised learning, tasks such as image captioning have leveraged a recurrent visual attention mechanism using recurrent neural networks (RNNs) to improve captioning performance [28]. The TieNet used this attention-based text embedding framework for pathology detection from chest radiographs [26], which was further improved by introducing a global topic vector and transfer learning [29]. A similar image-text embedding setup has been employed for chest radiograph (image) annotations [20]. In unsupervised learning, training a joint global embedding space for visual object discovery has recently been shown to capture relevant structure [11]. All of these models used RNNs for encoding text features. More

recently, transformers such as the BERT model [8] have shown the ability to capture richer contextualized word representations using self-attention and have advanced the state-of-the-art in nearly every language processing task compared to variants of RNNs. Our setup, while similar to [26] and [11], uses a series of residual blocks [13] to encode the image representation and uses the BERT model to encode the text representation. We use the radiology reports during training only, to improve the image-based model's performance. This is in contrast to visual question answering [2, 3, 19], where inference is performed on an image-text pair, and image/video captioning [16, 22, 24, 28], where the model generates text from the input image.

## 2 Data

For training and evaluating our model, we use the MIMIC-CXR dataset v2.0 [17], consisting of 377,110 chest radiographs associated with 227,835 radiology reports. The data was collected in routine clinical practice, and each report is associated with one or more images. We limited our study to 247,425 frontal-view radiographs.

### Regex Labeling.

We extracted pulmonary edema severity labels from the associated radiology reports using regular expressions (regex) with negation detection [6]. The keywords of each severity level (none=0, vascular congestion=1, interstitial edema=2, and alveolar edema=3) are summarized in the supplementary materials. In order to limit confounding keywords from other disease processes, we limited the label extraction to patients with congestive heart failure (CHF) based on their ED ICD-9 diagnosis code in the MIMIC dataset [10]. Cohort selection by diagnosis code for CHF was previously validated by manual chart review. This resulted in 16,108 radiology reports. Regex labeling yielded 6,710 labeled reports associated with 6,743 frontal-view images<sup>1</sup>. Hence, our dataset includes 247,425 image-text pairs, 6,743 of which are of CHF patients with edema severity labels. Note that some reports are associated with more than one image, so one report may appear in more than one image-text pair.

## 3 Methods

Let  $x^I$  be a 2D chest radiograph,  $x^R$  be the free-text in a radiology report, and  $y \in \{0, 1, 2, 3\}$  be the corresponding edema severity label. Our dataset includes a set of  $N$  image-text pairs  $X = \{x_j\}_{j=1}^N$ , where  $x_j = (x_j^I, x_j^R)$ . The first  $N_1$  image-text pairs are annotated with severity labels  $Y = \{y_j\}_{j=1}^N$ . Here we train a joint model that constructs an image-text embedding space, where an image encoder and a text encoder are used to extract image features and text features separately (Fig. 1). Two classifiers are trained to classify the severity labels independently from the image features and from the text features. This setup enables us to decouple the image classification and the text classification at inference time. Learning the two representations jointly at training time improves the performance of the image model.

### Joint Representation Learning.

We apply a ranking-based criterion [7, 12] for training the image encoder and the text encoder parameterized by  $\theta_E^I$  and  $\theta_E^R$  respectively, to learn image and text feature representations  $I(x^I; \theta_E^I)$  and  $R(x^R; \theta_E^R)$ . Specifically, given an image-text pair  $(x_j^I, x_j^R)$ , we randomly select an impostor image  $x_{s(j)}^I$  and an impostor report  $x_{s(j)}^R$  from  $X$ . This selection is generated at the beginning of each training epoch. Map  $s(j)$  produces a random permutation of  $\{1, 2, \dots, N\}$ .

We encourage the feature representations between a matched pair  $(I_j, R_j)$  to be “closer” than those between mismatched pairs  $(I_{s(j)}, R_j)$  and  $(I_j, R_{s(j)})$  in the joint embedding space. Direct minimization of the distance between  $I$  and  $R$  could end up pushing the image and text features into a small cluster in the embedding space. Instead, we encourage matched image-text features to be close while spreading out all feature representations in the embedding space for downstream classification by constructing an appropriate loss function:

$$\mathcal{F}_E(\theta_E^I, \theta_E^R; x_j, x_{s(j)}) = \max(0, \text{Sim}(I_j, R_{s(j)}) - \text{Sim}(I_j, R_j) + \gamma) + \max(0, \text{Sim}(I_{s(j)}, R_j) - \text{Sim}(I_j, R_j) + \gamma), \quad (1)$$

where  $\text{Sim}(\cdot, \cdot)$  is the similarity measurement of two feature representations in the joint embedding space and  $\gamma$  is a margin parameter that is set to  $|y_j - y_{s(j)}|$  when both  $j \in N_L$  and  $s(j) \in N_L$ ; otherwise,  $\gamma = 0.5$ . The margin is determined by the difference due to the mismatch, if both labels are known: otherwise the margin is a constant.

### Classification.

We employ two fully connected layers (with the same neural network architecture) on the joint embedding space to assess edema severity from the image and the report respectively. For simplicity, we treat the problem as multi-class classification, i.e. the classifiers' outputs  $\hat{y}^I(I; \theta_C^I)$  and  $\hat{y}^R(R; \theta_C^R)$  are encoded as one-hot 4-dimensional vectors. We use cross entropy as the loss function for training the classifiers and the encoders on the labeled data:

$$\mathcal{F}_C(\theta_E^I, \theta_E^R, \theta_C^I, \theta_C^R; x_j, y_j) = \sum_{i=0}^3 y_{ji} \log \hat{y}_i^I(I_j(x_j^I; \theta_E^I); \theta_C^I) + \sum_{i=0}^3 y_{ji} \log \hat{y}_i^R(R_j(x_j^R; \theta_E^R); \theta_C^R), \quad (2)$$

i.e., minimizing the cross entropy also affects the encoder parameters.

### Loss Function.

Combining Eq. (1) and Eq. (2), we obtain the loss function for training the joint model:

$$\mathcal{F}(\theta_E^I, \theta_E^R, \theta_C^I, \theta_C^R; X, Y) = \sum_{j=1}^N \mathcal{F}_E(\theta_E^I, \theta_E^R; x_j, x_s, j) + \sum_{j=1}^N \mathcal{F}_C(\theta_C^I, \theta_C^R, \theta_C^I, \theta_C^R; x_j, y_j) \quad (3)$$

### Implementation Details.

The image encoder is implemented as a series of residual blocks [13], the text encoder is a BERT model that uses the beginner [CLS] token's hidden unit size of 768 and maximum sequence length of 320 [8]. The image encoder is trained from a random initialization, while the BERT model is fine-tuned during the training of the joint model. The BERT model parameters are initialized using pre-trained weights on scientific text [4]. The image features and the text features are represented as 768-dimensional vectors in the joint embedding space. The two classifiers are both 768-to-4 fully connected layers. The neural network architecture is provided in the supplementary materials.

We employ the stochastic gradient-based optimization procedure AdamW [27] to minimize the loss in Eq. (3) and use a warm-up linear scheduler [25] for the learning rate. The model is trained on all the image-text pairs by optimizing the first term in Eq. (3) for 10 epochs and then trained on the labeled image-text pairs by optimizing Eq. (3) for 50 epochs. The mini-batch size is 4. We use dot product as the similarity metric in Eq. (1). The dataset is split into training and test sets. All the hyper-parameters are selected based on the results from 5-fold cross validation within the training set.

## 4 Experiments

### Data Preprocessing.

The size of the chest radiographs varies and is around 3000×3000 pixels. We randomly translate and rotate the images on the fly during training and crop them to 2048×2048 pixels as part of data augmentation. We maintain the original image resolution to capture the subtle differences in the images between different levels of pulmonary edema severity. For the radiology reports, we extract the *impressions*, *ndings*, *conclusion* and *recommendation* sections. If none of these sections are present in the report, we use the *nal report* section. We perform tokenization of the text using ScispaCy [21] before providing it to the BERT tokenizer.

### Expert Labeling.

For evaluating our model, we randomly selected 531 labeled image-text pairs (corresponding to 485 reports) for expert annotation. A board-certified radiologist and two domain experts reviewed and corrected the regex labels of the reports. We use the expert labels for model testing. The overall accuracy of the regex labels (positive predictive value compared against the expert labels) is 89%. The other 6,212 labeled image-text pairs and around 240K unlabeled image-text pairs were used for training. There is no patient overlap between the training set and the test set.

## Model Evaluation.

We evaluated variants of our model and training regimes as follows:

- **image only:** An image-only model with the same architecture as the image stream in our joint model. We trained the image model in isolation on the 6,212 labeled images.
- A joint image-text model trained on the 6,212 labeled image-text pairs only. We compare two alternatives to the joint representation learning loss:
  - **ranking dot, ranking l2, ranking cosine:** the ranking based criterion in Eq. (1) with  $\text{Sim}(I, R)$  defined as one of the dot product  $IR$ , the reciprocal of euclidean distance  $\|IR\|$ , and the cosine similarity  $\frac{I \cdot R}{\|I\| \|R\|}$ ;
  - **dot, l2, cosine:** direct minimization on the similarity metrics without the ranking based criterion.
- **ranking dot semi:** A joint image-text model trained on the 6,212 labeled and the 240K unlabeled image-text pairs in a semi-supervised fashion, using the ranking based criterion with dot product in Eq. (1). Dot product is selected for the ranking-based loss based on cross-validation experiments on the supervised data comparing ranking-dot, ranking-l2, ranking-cosine, dot, l2, and cosine.

All reported results are compared against the expert labels in the test set. The image portion of the joint model is decoupled for testing, and the reported results are predicted from images only. To optimize the baseline performance, we performed a separate hyperparameter search for the image only model using 5-fold cross validation (while holding out the test set).

We use the area under the ROC (AUC) and macro-averaged F1-scores (macro-F1) for our model evaluation. We dichotomize the severity levels and report 3 comparisons (0 vs 1,2,3; 0,1 vs 2,3; and 0,1,2 vs 3), since these 4 classes are ordinal (e.g., (severity = 0 or 1) =  $\hat{y}_0^I + \hat{y}_1^I$ , (severity = 2 or 3) =  $\hat{y}_2^I + \hat{y}_3^I$ ).

## Results.

Table 1 reports the performance statistics for all similarity measures. The findings are consistent with our cross-validation results: the ranking based criterion offers significant improvement when it is combined with the dot product as the similarity metric.

Table 2 reports the performance of the optimized baseline model (image only) and two variants of the joint model (ranking dot and ranking dot semi). We observe that when the joint model learns from the large number of unlabeled image-text pairs, it achieves the best performance. The unsupervised learning minimizes the ranking-based loss in Eq. (1), which does not depend on availability of labels.

It is not surprising that the model is better at differentiating the severity level 3 than other severity categories, because level 3 has the most distinctive radiographic features in the images.

### Joint Model Visualization.

As a by-product, our approach provides the possibility of interpreting model classification using text. While a method like Grad-CAM [23] can be used to localize regions in the image that are “important” to tire model prediction, it does not identify tire relevant characteristics of tire radiographs, such as texture. By leveraging tire image-text embedding association, we visualize tire heatmap of text attention corresponding to tire last layer of tire [CLS] token in tire BERT model. Tires heatmap indicates report tokens that are important to our model prediction. As shown in Fig. 2, we use Grad-CAM [23] to localize relevant image regions and tire highlighted words (radiographic findings, anatomical structures, etc.) from the text embedding to explain the model’s decision making.

## 5 Conclusion

In this paper, we presented a neural network model that jointly learns from images and text to assess pulmonary edema severity from chest radiographs. The joint image-text representation learning framework incorporates the rich information present in the free-text radiology reports and significantly improves the performance of edema assessment compared to learning from images alone. Moreover, our experimental results show that joint representation learning benefits from the large amount of unlabeled image-text data.

Expert labeling of the radiology reports enabled us to quickly obtain a reasonable amount of test data, but this is inferior to direct labeling of images. The joint model visualization suggests the possibility of using the text to semantically explain the image model, which represents a promising direction for future investigation.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments.

This work was supported in part by NIH NIBIB NAC P41EB015902, Wistron Corporation, Takeda, MIT Lincoln Lab, and Philips. We also thank Dr. Daniel Moyer for helping generate Fig. 1.

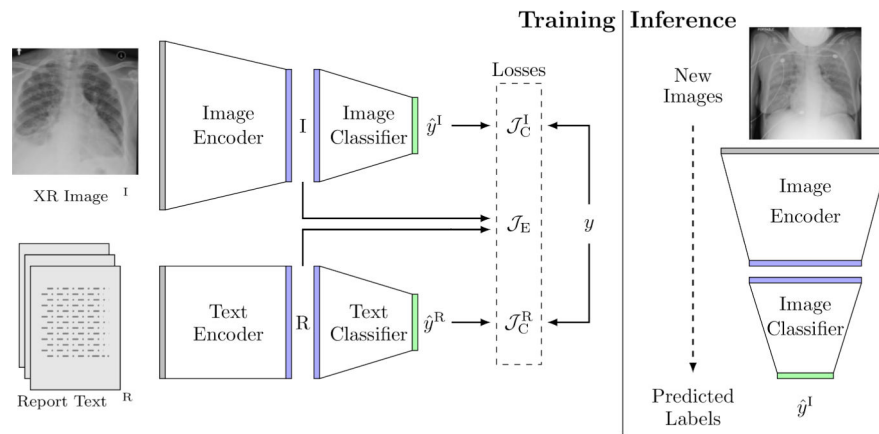
## References

1. Adams KF Jr, Fonarow GC, Emerman CL, LeJemtel TH, Costanzo MR, Abraham WT, Berkowitz RL, Galvao M, Horton DP, Committee ASA, Investigators, et al.: Characteristics and outcomes of patients hospitalized for heart failure in the United states: rationale, design, and preliminary observations from the first 100,000 cases in the acute decompensated heart failure national registry (adhere). *American heart journal* 149(2), 209–216 (2005) [PubMed: 15846257]
2. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L: Bottom-up and top-down attention for image captioning and visual question answering. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 6077–6086 (2018)
3. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Lawrence Zitnick C, Parikh D: Vqa: Visual question answering. In: *Proceedings of the IEEE international conference on computer vision* pp. 2425–2433 (2015)
4. Beltagy I, Lo K, Cohan A: Scibert: A pretrained language model for scientific text. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

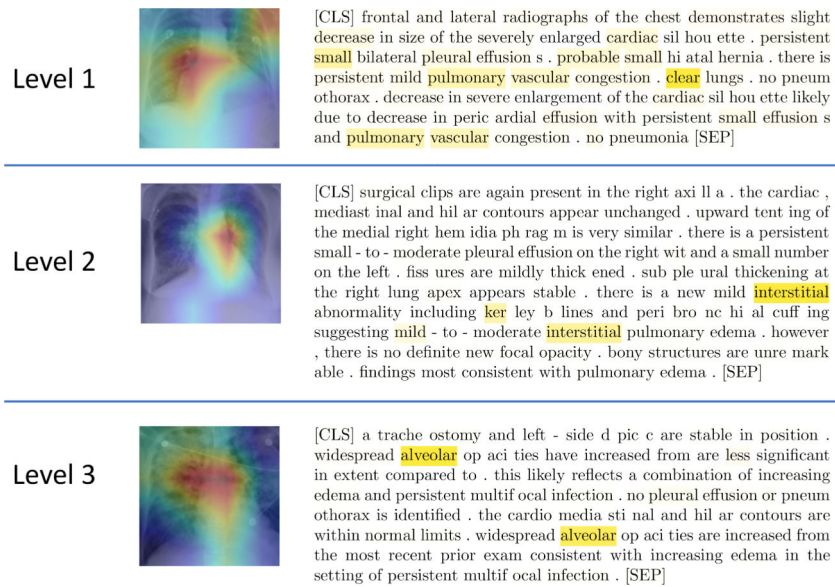
- International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) pp. 3606–3611 (2019)
5. Ben-Younes H, Cadene R, Cord M, Thome N: Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision pp. 2612–2620 (2017)
  6. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG: A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics* 34(5), 301–310 (2001) [PubMed: 12123149]
  7. Chechik G, Sharma V, Shalit U, Bengio S: Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research* 11 (Mar), 1109–1135 (2010)
  8. Devlin J, Chang MW, Lee K, Toutanova K: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
  9. Gheorghiade M, Follath F, Ponikowski P, Barsuk JH, Blair JE, Cleland JG, Dickstein K, Drazner MH, Fonarow GC, Jaarsma T, et al.: Assessing and grading congestion in acute heart failure: a scientific statement from the acute heart failure committee of the heart failure association of the european society of cardiology and endorsed by the european society of intensive care medicine. *European journal of heart failure* 12(5), 423–433 (2010) [PubMed: 20354029]
  10. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE: Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation* 101(23), e215–e220 (2000) [PubMed: 10851218]
  11. Harwath D, Recasens A, Suris D, Chuang G, Torralba A, Glass J: Jointly discovering visual objects and spoken words from raw sensory input. In: Proceedings of the European Conference on Computer Vision (ECCV) pp. 649–665 (2018)
  12. Harwath D, Torralba A, Glass J: Unsupervised learning of spoken language with visual context. In: Advances in Neural Information Processing Systems. pp. 1858–1866 (2016)
  13. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition pp. 770–778 (2016)
  14. Horng S, Liao R, Wang X, Dalal S, Golland P, Berkowitz SJ: Deep learning to quantify pulmonary edema in chest radiographs. arXiv preprint arXiv:2008.05975 (2020)
  15. Hunt SA, Abraham WT, Chin MH, Feldman AM, Francis GS, Ganiats TG, Jessup M, Konstam MA, Mancini DM, Michl K, et al.: 2009 focused update incorporated into the acc/aha 2005 guidelines for the diagnosis and management of heart failure in adults: a report of the american college of cardiology foundation/american heart association task force on practice guidelines developed in collaboration with the international society for heart and lung transplantation. *Journal of the American College of Cardiology* 53(15), e1–e90 (2009) [PubMed: 19358937]
  16. Jing B, Xie P, Xing E: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)
  17. Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C.y., Mark RG, Horng S: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* 6 (2019)
  18. Liao R, Rubin J, Lam G, Berkowitz S, Dalal S, Wells W, Horng S, Golland P: Semi-supervised learning for quantification of pulmonary edema in chest x-ray images. arXiv preprint arXiv:1902.10785 (2019)
  19. Lu J, Yang J, Batra D, Parikh D: Hierarchical question-image co-attention for visual question answering. In: Advances in neural information processing systems. pp. 289–297 (2016)
  20. Moradi M, Madani A, Gur Y, Guo Y, Syeda-Mahmood T: Bimodal network architectures for automatic generation of image annotation from text. In: International Conference on Medical Image Computing and Computer-Assisted Intervention pp. 449–456. Springer (2018)
  21. Neumann M, King D, Beltagy I, Ammar W: ScispaCy: Fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task pp. 319–327. Association for Computational Linguistics, Florence, Italy (8 2019). 10.18653/v1/W19-5034, <https://www.aclweb.org/anthology/W195034>



22. Plummer BA, Brown M, Lazebnik S: Enhancing video summarization via vision-language embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition pp. 5781–5789 (2017)
23. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision pp. 618–626 (2017)
24. Vasudevan AB, Gygli M, Volokitin A, Van Gool L: Query-adaptive video summarization via quality-aware relevance estimation. In: Proceedings of the 25th ACM international conference on Multimedia pp. 582–590 (2017)
25. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L.u., Polosukhin I: Attention is all you need In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc (2017), <http://papers.nips.cc/paper/7181attentionisallyouneed.pdf>
26. Wang X, Peng Y, Lu L, Lu Z, Summers RM: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition pp. 9049–9058 (2018)
27. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, et al.: Huggingface’s transformers: State-of-the-art natural language processing. ArXiv pp. arXiv-1910 (2019)
28. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning pp. 2048–2057 (2015)
29. Xue Y, Huang X: Improved disease classification in chest x-rays with transferred features from report generation. In: International Conference on Information Processing in Medical Imaging pp. 125–138. Springer (2019)

**Fig. 1:**

The architecture of our joint model, along with an example chest radiograph  $x^I$  and its associated radiology report  $x^R$ . At training time, the model predicts the edema severity level from images and text through their respective encoders and classifiers, and compares the predictions with the labels. The joint embedding loss  $\mathcal{J}_E$  associates image embeddings  $I$  with text embeddings  $R$  in the joint embedding space. At inference time, the image stream and the text stream are decoupled and only the image stream is used. Given a new chest radiograph (image), the image encoder and classifier compute its edema severity level.



**Fig. 2:** Joint model visualization. Top to bottom: (Level 1) The highlight of the Grad-CAM image is centered around the right hilar region, which is consistent with findings in pulmonary vascular congestion as shown in the report. (Level 2) The highlight of the Grad-CAM image is centered around the left hilar region which shows radiating interstitial markings as confirmed by the report heatmap. (Level 3) Grad-CAM highlights bilateral alveolar opacities radiating out from the hila and sparing the outer lungs. This pattern is classically described as “batwing” pulmonary edema mentioned in the report. The report text is presented in the form of sub-word tokenization performed by the BERT model, starting the report with a [CLS] token and ending with a [SEP].

**Table 1:**

Performance statistics for all similarity measures.

Method	AUC (0 v 1,2,3)	AUC (0,1 v 2,3)	AUC (0,1,2 v 3)	macro-F1
12	0.78	0.76	0.83	0.42
ranking 12	0.77	0.75	0.80	0.43
cosine	0.77	0.75	0.81	0.44
ranking cosine	0.77	0.72	0.83	0.41
dot	0.65	0.63	0.61	0.15
<b>ranking dot</b>	<b>0.80</b>	<b>0.78</b>	<b>0.87</b>	<b>0.45</b>

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 2:**

Performance statistics for the two variants of our joint model and the baseline image model.

Method	AUC (0 v 1,2,3)	AUC (0,1 v 2,3)	AUC (0,1,2 v 3)	macro-F1
image only	0.74	0.73	0.78	0.43
ranking dot	0.80	0.78	0.87	0.45
<b>ranking dot semi</b>	<b>0.82</b>	<b>0.81</b>	<b>0.90</b>	<b>0.51</b>

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript