

Improved Disease Classification in Chest X-rays with Transferred Features from Report Generation

Yuan Xue, Xiaolei Huang*

College of Information Sciences and Technology,
The Pennsylvania State University, University Park, PA

Abstract. Radiology includes using medical images for detection and diagnosis of diseases as well as guiding further interventions. Chest X-rays are commonly used radiological examinations to help spot thoracic abnormalities or diseases, especially lung-related diseases. However, the reporting of chest x-rays requires experienced radiologists who are often in shortage in many regions of the world. In this paper, we first develop an automatic radiology report generation system. Due to the lack of large annotated radiology report datasets and the difficulty of evaluating the generated reports, the clinical value of such systems is often limited. To this end, we train our report generation network on the small IU Chest X-ray dataset then transfer the learned visual features to classification networks trained on the large ChestX-ray14 dataset and use a novel attention guided feature fusion strategy to improve the detection performance of 14 common thoracic diseases. Through learning the correspondences between different types of feature representations, common features learned by both the report generation and the classification model are assigned with higher attention weights and the weighted visual features boost the performance of state-of-the-art baseline thoracic disease classification networks without altering any learned features. Our work not only offers a new way to evaluate the effectiveness of the learned radiology report generation network, but also proves the possibility of transferring different types of visual representations learned on a small dataset for one task to complement features learned on another large dataset for a different task and improve the model performance.

1 Introduction

Deep learning has shown its strength in various types of computer vision tasks such as classification and detection in the last decade. With the rapid development of advanced algorithms and the availability of more annotated datasets, potential applications of deep learning in radiology have become possible and desired by the public. Among such applications, the report generation networks [11, 14, 26] and the disease classification networks [23, 24, 27, 19] are two popular trends. An automated or computer-aided radiology reporting system can provide preliminary findings to radiologists to assist with report writing; A disease classification network can help detect potential abnormalities and diseases shown in the examinations. Both systems can help reduce the workload of expert radiologists as well as make up for staff shortage.

* Xiaolei Huang is the corresponding author. Her email address is: suh972@psu.edu

Current report generation methods mainly follow the encoder-decoder architecture which has been widely used for image captioning [22, 25, 17]. In the radiology report, the *Impression* and the *Findings* can be generated based on the interpretation of images without prior examinations or patient history, which makes the generation tasks a perfect fit for deep learning models. Impression can be a single-sentence conclusion or diagnosis, and findings can be a coherent paragraph containing multiple sentences that describe the radiologist’s observations and findings regarding both normal and abnormal features in the images. While the impression generation can often be handled by an image captioning algorithm, a findings paragraph consists of multiple sentences thus traditional captioning algorithms designed for generating single-sentence description no longer apply. For findings generation, state-of-the-art radiology report generation models use either a hierarchical architecture [11, 14] to generate different sentences based on different topic vectors, or a recurrent architecture [26] to generate one sentence at a time whereby the following sentence is conditioned upon the content of its preceding sentence. Combining the merits of both, we propose a modified recurrent attention model where a succeeding sentence is generated based upon both the prior sentence and a global topic vector encoded by the generated impression. The recurrent attention mechanism guarantees the model to generate diverse and coherent sentences, where the global topic vector forces the model to produce sentences supporting the conclusion or diagnosis.

Although some accomplishments have been made in radiology report generation, there are still several issues remaining unsolved. First, training a general and robust report generation system on a small training set is impractical, as the ground truth report provided in the training set can be biased towards the radiologist’s personal style. Even for the same image, different radiologists can produce entirely different written reports as they only provide information that they think might be important to the potential referees. More importantly, evaluation of the generated report is difficult. Any automatic metrics based on the overlap between the prediction and ground truth cannot capture the words describing negation and uncertainty and lack measurement of semantic similarity, while human evaluation requires extra efforts by human experts and can be error-prone. While the generated report learned by the decoder model can be biased towards the training data, the learned visual features from the encoder model are expected to be more general and robust. Otherwise they are not able to provide enough information to the decoder model to produce diverse sentences describing different regions. Therefore, we introduce a novel evaluation of the trained chest X-ray report generation network, where we only take the encoder part of the model as a visual feature extractor, and transfer these learned features to be used in another large chest X-ray dataset and evaluate their performance for thoracic disease classification.

Traditional transfer learning benefits from training with a large-scale dataset then transferring the learned knowledge or model weights to another dataset that is typically much smaller for better initialization and faster convergence. However, transfers from a small dataset to a large dataset with cross-task generalization are under-investigated, but such transfers have great research and clinical value since some annotations like complete radiology reports are expensive to obtain due to the difficulty of labeling or privacy concerns. In this work, taking the learned image encoder from the report gen-

eration model, we apply the extracted visual features to another chest X-ray dataset for disease classification. We combine the features from the report generator and the features extracted by a well-trained disease classification network to form an attention map, where common features learned by both networks are emphasized with higher weights. Then we apply the attention map to the original feature learned by the classification network and re-train the transition layer for new predictions. The transferred visual features along with the attention guided feature fusion shows considerable improvements over a baseline classification network, and even achieve further improvements over the state-of-the-art CheXNet [19] without changing any visual features. Our work not only offers a new way to evaluate the effectiveness of learned radiology report generation network, but also proves the possibility of transferring different types of visual representation learned on a small dataset for one task to complement features learned on another large dataset for a different task and improve the model performance.

2 Related Work

Radiology Report Generation. As a joint application of Computer Vision (CV) and Nature Language Processing (NLP) in healthcare, the automatic radiology report generation task has recently attracted considerable attention. Following the encoder-decoder architecture and attention mechanism in image captioning [22, 25, 17], several report generation works have been proposed. To generate long and coherent reports, previous works [11, 14] use a hierarchical architecture [13, 1] to generate a sequence of encoded topic vectors, then each sentence in the report is generated conditioned on the topic vector; Xue *et al.* [26] utilizes a recurrent attention model and brings the contextual information into the loop when predicting next words and sentences. Li *et al.* [14] combines a retrieval module and a generation module to either select a phrase from the template database or generate a new sentence. However, such models still suffer from the bias of training set, and automatic evaluation cannot capture the words describing negation and uncertainty in the predicted report. Taking the merits of previous methods, we modify the recurrent attention model [26] and introduce a global topic vector to guide the generation of findings. Instead of only evaluating the predicted report, we transfer the learned visual features to the ChestX-ray14 dataset and use an attention guided feature fusion scheme to help improve the disease classification performance.

Thoracic Disease Classification. Wang *et al.* [23] introduced the ChestX-ray14 dataset which is currently the largest public repository of radiographs. They also reported a weakly supervised multi-label thoracic disease classification and localization framework in their paper. Yao *et al.* [27] leveraged inter-dependencies among different pathology labels in chest X-rays via LSTMs to improve the disease classification performance. In [19], state-of-the-art disease classification result is reported using a DenseNet-121 [10] backbone; their trained model gets impressive results and exceeds the average radiologist performance on the pneumonia detection task.

A different approach for thoracic disease classification trains the model with multiple tasks to improve classification result. Li *et al.* [15] presents a unified network that simultaneously improves classification and localization with the help of extra bounding boxes indicating disease location. Wang *et al.* [24] uses a CNN-RNN model to generate

a radiology report and encodes the generated report directly to improve the classification result. Different from our model, their report generation and classification models are trained on the same dataset where the ground truth disease labels and reports are inherently correlated since the labels are text-mined from the original report. Since there is no large dataset with annotated radiology report available to the public, it is hard to extend their framework to other domains and tasks.

Transfer Learning for Medical Imaging. While multi-task learning can benefit from sharing features between different tasks to get more general features, transfer learning can leverage the learned features from a source domain to improve the performance in a target domain. Traditional transfer learning use knowledge or model weights learned from tasks for which a large amount of annotated data is available in tasks where only a limited amount of labeled data is available. The large dataset can help generalize the model as training on small datasets can lead to overfitting. Transfer learning has been widely adopted in deep learning such as initializing models with weights pre-trained on ImageNet [4] and has shown performance improvements for medical image classification [20, 9].

However, in medical applications, the amount of labeled data is typically low so using features learned from other medical domains directly is often impractical. To evaluate the visual features learned by our report generator and explore the possibility of transferring visual features learned from a small set to a larger set and improving the cross-task generalization, we propose an attention based feature fusion scheme to modify the already trained classification model. Instead of averaging features in all regions as average pooling in standard models [7, 10], our transfer learning model assigns more weights to the features coexisting in both original feature space and transferred feature space through an attention guided feature fusion mechanism. In this way, features learned in a small set can also help generalize the model for another domain or task.

3 Methodology

3.1 Radiology Report Generator

In our report generation model, the estimated probability of generating the i -th sentence with length T is

$$\hat{\mathbb{P}}(S_i = w_1, w_2, \dots, w_T | V, I; \theta) = \mathbb{P}(S_1 | V, I) \prod_{j=2}^{i-1} \mathbb{P}(S_j | V, I, S_{j-1}) \mathbb{P}(w_1 | V, I, S_{i-1}) \prod_{t=2}^T \mathbb{P}(w_t | V, I, S_{i-1}, w_1, \dots, w_{t-1}) . \quad (1)$$

where V is the given frontal view image, I is the impression topic and θ is the model parameter (we omit the θ in the right hand side), S_i represents the i -th sentence and w_t is the t -th token in the i -th sentence. In other words, each succeeding sentence is conditioned upon multimodal inputs including its preceding sentence, the global impression topic and the original image.

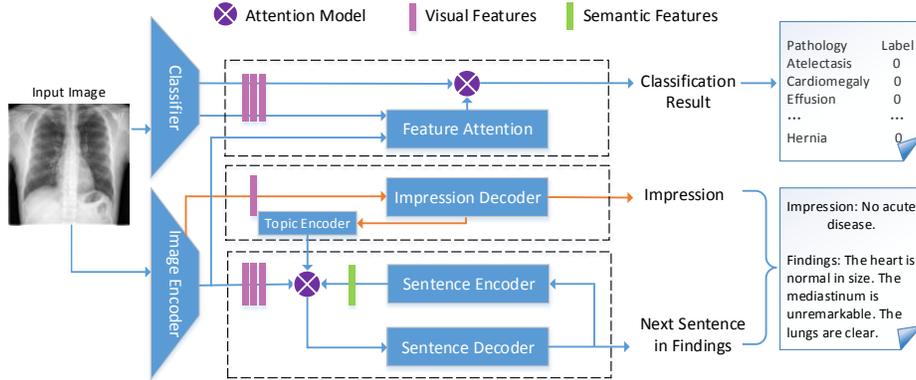


Fig. 1. The architecture of our model. Lower part is the proposed generative model for radiology reports. Upper part is the attention guided feature fusion model for improved thoracic disease classification. Best viewed in color.

The training is done by Maximum Log-likelihood Estimate (MLE) and minimizing the cross entropy loss as

$$\theta^* = \operatorname{argmax}_{\theta} \sum_{i=1}^L \log \hat{\mathbb{P}}(S_i = G_i | V, I; \theta), \quad (2)$$

where G_i is the ground truth for the i -th sentence in the findings paragraph.

The overall architecture of our framework that takes frontal image views as input and generates a radiology report with impression and findings is shown in the lower part of Fig. 1. The report generator first takes the frontal view image as input and generates an impression sentence via a CNN-RNN model, then a Bi-directional Long Short-Term Memory (Bi-LSTM) [6] with mean-pooling over the sequence length is used as a topic encoder to produce a global topic vector. The generation of the first sentence in the findings paragraph is conditioned upon both the image input and the topic vector. Then the decoder model recurrently takes the multimodal inputs of visual features from the image encoder, semantic representation of the preceding sentence from the sentence encoder, and the global topic vector of the impression from the topic encoder and generates the next sentence. The report decoding model can be regarded as a combination of the hierarchical model and the recurrent model. More details are shown in the following subsections.

Image Encoder. Similar to [26], a Convolutional Neural Network (CNN) based image encoder is first applied to extract both global and regional visual features from the input images. To get a better initialization, the image encoder is built upon the pre-trained ResNet-152 [7] network pre-trained on ImageNet [4]. We resize the input images to 224×224 to be consistent with the pre-trained ResNet-152 image encoder. To get a better attention map, the local feature matrix $f \in \mathbb{R}^{2048 \times 49}$ (reshaped from $2048 \times 7 \times 7$) is extracted from the last convolution layer of ResNet-152. Each column of f is one regional feature vector. Thus each image has $k = 49$ sub-regions. Meanwhile, we extract the global feature vector $f \in \mathbb{R}^{2048}$ from the last average pooling layer of ResNet-152.

The image encoder is fine tuned on the frontal view chest X-rays so that the learned visual features can be transferred to another classification task.

Report Decoder. The decoder network is responsible for generating both the impression and the findings paragraph. The impression decoder first takes the global visual features learned by the image encoder as input and a single layer Long Short-Term Memory (LSTM) [8] is used for sentence decoding. Following [2], we adopt a Bi-LSTM that reads the generated impression in two opposite directions along with a mean pooling over each dimension of the hidden units to get a global topic vector. The sentence decoder takes regional visual features and the topic vector to generate the first sentence. After that, each sentence is generated by taking regional visual features, the previously generated sentence and the topic vector as a multimodal input. The sentence decoder is a stacked 2-layer LSTM. All hidden and embedding dimensions are fixed to 512 for all of the LSTM models discussed herein. Both regional and global visual features are converted into channel dimension 512 to match the embedding size before being fed into any LSTMs. The regional visual features are converted as input to the sentence decoder. The global topic vector is used as the initialization of the sentence decoder, while the learned encoding of the preceding sentence is combined with the visual representations to generate an attention map. The attention weights for the regional visual features are computed as follows:

$$\mathbf{a} = \mathbf{W}_{\text{att}}((\mathbf{v}; \mathbf{s}\mathbf{1}^k)) + \mathbf{b}_{\text{att}} , \quad (3)$$

where $\mathbf{v} \in \mathbb{R}^{512 \times 49}$ are the regional visual features learned by the image encoder, $\mathbf{s} \in \mathbb{R}^{512 \times 1}$ represents the encoding of the preceding sentence, $\mathbf{1}^k \in \mathbb{R}^{1 \times 49}$ is a vector with all ones. \mathbf{v} and $\mathbf{s}\mathbf{1}^k$ are concatenated along the embedding dimension as the input to the attention network. $\mathbf{W}_{\text{att}} \in \mathbb{R}^{1 \times 1024}$, and $\mathbf{b}_{\text{att}} \in \mathbb{R}^{1 \times 49}$ are parameters of the attention network. Then the weights are normalized over all regions to get the attention distribution and applied to the original visual representation as:

$$\mathbf{v}_{\text{att}} = \sum_{i=1}^k \frac{\exp(a_i)}{\sum_i \exp(a_i)} \mathbf{v}_i , \quad (4)$$

where i refers to the i -th region in the regional visual representation.

The generation process is repeated until an empty sentence is generated or a maximum number of sentences is reached, which indicates the end of the paragraph. Our model combines the recurrent and hierarchical architecture in the decoder network. While the recurrent attention mechanism forces the model to focus on different regions of the input image to generate more diverse sentences and keep intra-paragraph coherence, the global topic vector adds an additional constraint so that the generated sentences can support the theme of the entire report. An example of the generated report by our report generator is shown in Fig. 2.

Our proposed report generator is trained by the Adam optimizer [12]. The initial learning rate is set to be 1e-4, and learning rate decay is 0.5 for every 5 epochs. The batchsize is 32 for training. During inference, the greedy search is adopted for generating words and sentences in every timestep. The maximum number of sentences is set to be 7. Although the impression decoder and the findings decoder can be trained jointly, we separate the training process to get more diverse results. The final model transferred to the classification model for feature fusion is trained on all training data. Although

Input Image	Predicted Report	Original Report
	<p>Findings: The cardiac contours are normal. The lungs are clear. There is no pleural effusion or pneumothorax. There is no focal air space opacity to suggest a pneumonia. Degenerative changes of the thoracic spine.</p> <p>Impression: No acute cardiopulmonary finding.</p>	<p>Findings: Heart size and mediastinal contour normal. Lungs are clear. Pulmonary vascularity normal. No pleural effusions or pneumothoraces. Minimal degenerative changes thoracic spine.</p> <p>Impression: No acute cardiopulmonary process.</p>

Fig. 2. An example report generated by our model in comparison with the original written report provided by the radiologist. The notable words in bold font in the findings indicate that our model is capable of capturing some of the abnormalities in the input image.

the generated report can be biased towards the training data, we believe that the learned visual features should be more general than the semantic features and can be transferred to other domains.

3.2 Thoracic Disease Classifier and Attention Guided Feature Fusion

The basic thoracic disease classification network is a CNN that takes frontal view radiographs as input and generates multiple disease labels. We first train two different baseline classification models to get the visual features. We start from the ResNet-18 model and later move to the DenseNet-121 model with more layers. For the ResNet baseline, the final fully connected layer is replaced with one that has 14 outputs for 14 disease categories, after which we apply a sigmoid nonlinearity as in [19]. The training is done by minimizing the binary cross entropy loss. The weights of the network are initialized with weights from models pre-trained on ImageNet [4]. We pick the model with the lowest validation loss as the final model. After the training is completed, we discard the transition block including the final pooling layer, the final fully connected layer and the sigmoid nonlinearity to keep only the local visual features. Then, we run the image encoder of our report generator on the classification dataset. The extracted local visual features serve as a high-level linguistic abstract of the input images here. Both the local visual features from the classification network and the report generation network with $k = 49$ sub-regions are then fed into the feature attention module. The feature attention module can be interpreted as:

$$\mathbf{a}' = \mathbf{W}'_{\text{att}}((\mathbf{v}^r; \mathbf{v}^c)) + \mathbf{b}'_{\text{att}}, \quad (5)$$

where \mathbf{v}^r are the regional visual features from the report generator, \mathbf{v}^c are the visual features from the original classification network. \mathbf{v}^r and \mathbf{v}^c are concatenated along the channel/embedding dimension as the input to the feature attention network. \mathbf{W}'_{att} and \mathbf{b}'_{att} are parameters of the attention network. Then the modified visual representation is computed as:

$$\mathbf{v}^c_{\text{att}} = \sum_{i=1}^k \frac{\exp(a'_i)}{\sum_i \exp(a'_i)} \mathbf{v}^c_i, \quad (6)$$

where i refers to the i -th region in the regional visual representation. The detailed process is shown in the upper part of Fig. 1. Local features from the image encoder and the classification model are concatenated and fed into the feature attention module. After calculating an attention map over all regions, we apply the attention weights to the original visual features learned from the classification network. We fix the visual features

and re-train the transition block to get a new classification prediction. While visual features remain unchanged, the feature attention module can discover the correspondences between features trained on two separate domains and tasks, and emphasize more on the features that coexist in both representations. We replace the average pooling operation in the original classification model with learned attention weights for feature fusion. Compared with average pooling, the attention guided feature fusion can achieve better generalization and get potentially higher classification accuracy without re-training the feature extraction model or changing any learned visual features.

To better illustrate the strength of our feature attention module, we also apply the attention guided feature fusion on the state-of-the-art CheXNet [19] which uses a DenseNet-121 backbone model. The training process is similar to the ResNet-18 baseline model; the only difference is that the output channel dimension in DenseNet-121 is 1024 while it is 512 in ResNet-18. Since the CheXNet is the state-of-the-art model and already gets very high accuracy on the ChestX-ray14 [23] dataset, it is hard to get further improvements. We did experiments that performed a naive concatenation of the two feature representations, but failed to get any significant improvements over the original classification model. Using our attention guided feature fusion mechanism, however, we are able to achieve improvements (as shown in Table 2) with the help of visual features learned on a small report generation dataset and thus demonstrate the effectiveness of our model. More details are explained in Section 4.

4 Experiments

Our report generator is trained on the Indiana University Chest X-Ray Collection [3]. The IU X-ray dataset contains 3,955 radiology reports from 2 large hospital systems within the Indiana Network for Patient Care database, and 7,470 associated chest X-rays from the hospitals picture archiving systems. Each report is associated with a pair of images which are the frontal and lateral views, and contains comparison, indication, findings, and impression sections. Since the transferred visual features will be utilized in the ChestX-ray14 dataset containing only frontal view X-rays, we further filter out reports without frontal view images or without complete sections of findings and impression, resulting in 3,331 reports with associated frontal view images.

For data preprocessing, we tokenize and lowercase all the words that appear more than twice in the findings and impression sections of all reports and obtain 1,357 unique words. To mitigate the small size of the dataset and generalize our model, dropout of 0.2 is added in both the encoder and decoder networks. Moreover, considering that images in the IU X-ray and ChestX-ray14 datasets can look different, we further resize the original input images to size 256×256 then randomly crop them to size 224×224 and randomly change the brightness, contrast and saturation of input images with rate 0.1 for data augmentation. All images are normalized based on the mean and standard deviation of images in the ImageNet training set. To provide some insights to the performance of our report generator, we report BLEU [18], METEOR [5], ROUGE [16] and CIDEr [21] scores as for image captioning tasks and in previous works [11, 14, 26]. The automatic evaluation results reported in Table. 1 are done on the test set with 300 randomly picked reports. We tokenize and lowercase all words in both the predicted report and the ground

Table 1. Evaluation of generated reports on our test set using BLEU, METEOR, ROUGE and CIDEr metrics. For findings generation, we compare our model with two baseline models [26, 14]. We also provide a comparison with one baseline model [26] for the impression and findings generation.

Data	Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
Findings	Recurrent-Attn [26]	0.441	0.320	0.231	0.181	0.220	0.366	0.243
	HRGR-Agent [14]	0.438	0.298	0.208	0.151	-	0.322	0.343
	Ours	0.477	0.332	0.243	0.189	0.223	0.380	0.320
Findings + Impression	Recurrent-Attn [26]	0.465	0.332	0.244	0.190	0.224	0.480	0.495
	Ours	0.489	0.340	0.252	0.195	0.230	0.478	0.565

truth report, and all punctuation are considered as independent tokens. We compare with baseline model [26] on both the findings generation and the combination of findings and impression. The result of [14] is from the original paper and they use different train/test split only for findings generation; baseline model of [26] is re-trained using our train/test split with only frontal view radiographs. During the evaluation, we observe that higher scores do not necessarily indicate better generation performance and reports with more sentences describing normal findings typically get higher scores but may miss more crucial abnormalities. The combination results of findings and impression always get higher scores since most impression sentences are normal such as “no acute cardiopulmonary findings”. Due to the small size of the training dataset and the nature of the radiology report generation problem, we believe that metrics designed for image captioning are not appropriate evaluations for report generation, as they cannot capture the words describing negation and uncertainty in the report. Even for the CIDEr [21] which has shown a better correlation with human judgments than other metrics for captioning tasks, it is based on term frequency-inverse document frequency (TF-IDF). However, the corpus of radiology reports is very different from other text corpora so CIDEr may not work as well as in other tasks. Thus, our goal is not to evaluate our model with these metrics and compare directly with other models. Rather, the results are only illustrated to show that our report generation model is among the state-of-the-arts and should be able to learn meaningful features by the image encoder. The learned visual features are formally evaluated on the ChestX-ray14 dataset via feature transferring and feature fusion.

ChestX-ray14 [23] is currently the largest public repository of radiographs containing 112,120 frontal view chest X-rays of 30,805 unique patients. Each image is annotated with up to 14 different thoracic pathology labels text-mined from the associated report. The labels are expected to have accuracy higher than 90%. The dataset is randomly split into training (70%), validation (10%), and test (20%) sets as in previous work on ChestX-ray14 [23, 27, 19]. The data preprocessing is the same as in the report generation model, including the random crop and color jitter for data augmentation and consistent with the training of the report generator. Training of both classification models and the feature fusion are done via the Adam optimizer [12] and minimizing the binary cross entropy (BCE) loss. The initial learning rate is also set to be $1e-4$, and learning rate decay is 0.5 every time the validation loss has not been decreased for 5 epochs. The batchsize is 32 for training. We first train two baseline classification net-

Table 2. Comparison of AUCs of ROC curve for classification of 14 disease categories in ChestX-ray14 test set. R18 and D121 represent the ResNet-18 and the DenseNet-121 baseline models, respectively. TF denotes the model with transferred features and attention guided feature fusion. Note that the state-of-the-art baseline model of CheXNet [19] with DenseNet-121 backbone is re-implemented and re-trained since we use different data preprocessing.

Pathology	[23]	[24]	[27]	[15]	R18	R18-TF	D121[19]*	D121-TF
Atelectasis	0.716	0.732	0.772	0.80	0.797	0.819	0.814	0.822
Cardiomegaly	0.807	0.844	0.904	0.81	0.895	0.884	0.902	0.892
Effusion	0.784	0.793	0.859	0.87	0.874	0.875	0.878	0.881
Infiltration	0.609	0.666	0.695	0.70	0.694	0.709	0.706	0.710
Mass	0.706	0.725	0.792	0.83	0.817	0.813	0.838	0.841
Nodule	0.671	0.685	0.717	0.75	0.732	0.789	0.782	0.794
Pneumonia	0.633	0.72	0.713	0.67	0.766	0.770	0.774	0.767
Pneumothorax	0.806	0.847	0.841	0.87	0.857	0.852	0.845	0.870
Consolidation	0.708	0.701	0.788	0.80	0.795	0.810	0.806	0.813
Edema	0.835	0.829	0.882	0.88	0.89	0.898	0.895	0.898
Emphysema	0.815	0.865	0.829	0.91	0.883	0.905	0.910	0.922
Fibrosis	0.769	0.796	0.767	0.78	0.821	0.844	0.838	0.851
Pleural Thickening	0.708	0.735	0.79	0.772	0.763	0.780	0.779	0.788
Hernia	0.767	0.876	0.914	0.77	0.923	0.929	0.951	0.946
Average	0.738	0.772	0.802	0.80	0.822	0.834	0.837	0.842

works: ResNet-18 [7] and CheXNet [19] using a DenseNet-121 [10] backbone, then evaluate the disease classification performance of the feature fusion, using the area under ROC curve (AUC) score for 14 different diseases. During feature fusion, all visual features learned by the image encoder of report generator and by the original classification model are fixed to ensure the improvements are not from the re-training of the model. Table 2 illustrates the per-class and average AUCs comparison of 14 diseases on the test set. Unlike [19], random horizontal flipping is not implemented as we believe the input frontal view chest X-rays are not symmetrical (e.g., cardiac abnormalities always appear in the left side of the chest) and it is not reasonable to flip the input images. For baseline classification models, we change the last fully connected layer accordingly to fit the number of categories which is 14 for the ChestX-ray14 dataset.

As we can see in Table 2, the ResNet-18 model with transferred features and attention guided feature fusion outperforms the baseline ResNet-18 classification model considerably on almost all diseases except for Mass. Remind that we do not re-train any of the visual features learned by the original classification network and the performance boost comes from better utilization of the learned features. To better illustrate the effectiveness of feature transferring and the attention guided feature fusion, we also apply our model to the state-of-the-art CheXNet [19]. With different data preprocessing, we re-train the CheXNet with a DenseNet-121 backbone as a baseline model. After feature fusion, we observe clear improvements on 12 diseases except for Cardiomegaly and Hernia, without altering any visual features learned by the original CheXNet. Moreover, our DenseNet-121 model with transferred features achieves highest AUC scores on 11 out of 14 classes, and has the highest average AUC score among all methods. The experimental results show that the image encoder in the report generator indeed

learned meaningful features during training, and the attention guided feature fusion is capable of improving the classification result through a better utilization of features and an emphasis on features that generalize across tasks.

5 Conclusion

In summary, we have proposed an improved recurrent attention model for radiology report generation along with an attention guided feature transfer and feature fusion model for thoracic disease classification. The report generation model is first trained on a small chest X-ray dataset with written reports provided by radiologists, then the learned visual representations including some high-level abstract of the input radiographs are transferred to a larger chest X-ray dataset with multiple disease labels. The features are combined under the guidance of a feature attention module. After applying the attention weights to the features extracted by the original classification model, we successfully improve the disease classification result without changing any visual features on the state-of-the-art baseline classification network. The experimental results on ChestX-ray14 dataset demonstrate that the proposed transferring of visual representations learned for different tasks on different datasets and the attention guided feature fusion can improve the model performance even on a large dataset. We believe that, by utilizing feature representations from different domains or tasks in a complementary manner, such feature transfer and fusion models have great potential and can be extended to other medical imaging applications where training data are limited so as to generalize the original model and enhance performance.

References

1. Chatterjee, M., Schwing, A.G.: Diverse and coherent paragraph generation from images. arXiv preprint arXiv:1809.00681 **2** (2018)
2. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)
3. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* **23**(2), 304–310 (2015)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. pp. 248–255. IEEE (2009)
5. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: *Proceedings of the ninth workshop on statistical machine translation*. pp. 376–380 (2014)
6. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks* **18**(5-6), 602–610 (2005)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 770–778. IEEE (2016)

8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* **9**(8), 1735–1780 (1997)
9. Hoo-Chang, S., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging* **35**(5), 1285 (2016)
10. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269. IEEE (2017)
11. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3337–3345. IEEE (2017)
14. Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. arXiv preprint arXiv:1805.08298 (2018)
15. Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.J., Li, F.F.: Thoracic disease identification and localization with limited supervision. arXiv preprint arXiv:1711.06373 (2017)
16. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* (2004)
17. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3242–3250. IEEE (2017)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
19. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
20. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging* **35**(5), 1299–1312 (2016)
21. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4566–4575. IEEE (2015)
22. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3156–3164. IEEE (2015)
23. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3462–3471. IEEE (2017)
24. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: 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)
25. 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)

26. Xue, Y., Xu, T., Long, L.R., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multimodal recurrent model with attention for automated radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 457–466. Springer (2018)
27. Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K.: Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.10501 (2017)