Feature Difference Makes Sense:
A Medical Image Captioning Model Exploiting Feature Difference and Tag Information

Hyeryun Park[1], Kyungmo Kim[1], Jooyoung Yoon[1], Seongkeun Park[2], Jinwook Choi[2,3]

[1]Interdisciplinary Program for Bioengineering, Graduate School, Seoul National University
[2]Department of Biomedical Engineering, College of Medicine, Seoul National University
[3]Institute of Medical and Biological Engineering, MRC, Seoul National University
{helena.park, medinfoman, joo0yoon, ilj, jinchoi}@snu.ac.kr

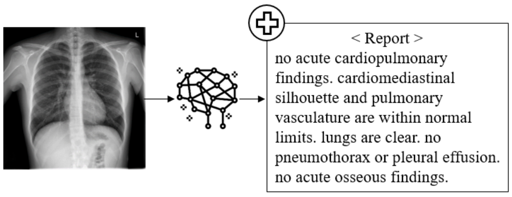
Abstract

Medical image captioning can reduce the workload of physicians and save time and expense by automatically generating reports. However, current datasets are small and limited, creating additional challenges for researchers. In this study, we propose a feature difference and tag information combined long short-term memory (LSTM) model for chest x-ray report generation. A feature vector extracted from the image conveys visual information, but its ability to describe the image is limited. Other image captioning studies exhibited improved performance by exploiting feature differences, so the proposed model also utilizes them. First, we propose a difference and tag (DiTag) model containing the difference between the patient and normal images. Then, we propose a multi-difference and tag (mDiTag) model that also contains information about low-level differences, such as contrast, texture, and localized area. Evaluation of the proposed models demonstrates that the mDiTag model provides more information to generate captions and outperforms all other models.

1 Introduction

Image captioning is a research area that generates text describing natural images, representing a convergence of computer vision and natural language processing. There are several existing methods for image captioning. One way involves filling up templates with detected objects or properties (Li et al., 2011; Yang et al., 2011), but this has limitations about diversity. Especially, sentences describing abnormal findings in medical images are relatively diverse and rare. Another involves retrieving the captions of images that are similar to the query image and selecting relevant phrases from those captions to generate new captions (Gupta et al., 2012; Kuznetsova et al., 2014). However, this method does not generalize well when applied to unfamiliar images.

To overcome the weaknesses of current methods, we adopted the encoder-decoder architecture with an attention mechanism. The encoder encodes an image into a feature vector, and the decoder decodes the feature vector into text. The encoder-decoder is one of the neural networks successfully used in other recent image captioning studies (Vinyals et al., 2015; Xu et al., 2015; Karpathy and Fei-Fei, 2015; You et al., 2016; Zhou et al., 2017; Anderson et al., 2018).



< Report >
no acute cardiopulmonary findings. cardiomediastinal silhouette and pulmonary vasculature are within normal limits. lungs are clear. no pneumothorax or pleural effusion. no acute osseous findings.

Figure 1: An example of a medical image captioning system that generates a report given a chest x-ray image.

Medical image captioning is the field of generating medical reports that describe medical images, as shown in Figure 1. The first challenge in medical image captioning is the lack of quality in training sets. Researchers have difficulty accessing chest x-ray datasets, which slows the development of related technologies. There are publicly available datasets that have images and reports: IU X-RAY, PEIR GROSS, and ICLEF-CAPTION (Kougia et al., 2019). Using only these datasets, state-of-the-art caption generation models do not generate medical reports correctly. Recently, MIMIC-CXR (Johnson et al., 2019), the largest dataset with images, reports, and labels, is released. The second challenge is that there are too many normal descriptions in the dataset, which creates a skewed dataset that poses problems for

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 95–102
July 5 - July 10, 2020. ©2020 Association for Computational Linguistics

supervised learning. Besides, some types of significant abnormal findings appear too rarely in the dataset to appropriately train the model.

In this study, we propose a model that can identify and focus on abnormal findings more specifically and precisely, similar to the way that physicians would typically read, interpret, and write chest x-ray reports. Since physicians look for the differences between the normal group and the disease group, we also focus on image feature differences. Therefore, the proposed model sets the criteria based on a normal x-ray image and creates a feature difference vector that explains the difference between a normal x-ray image and a patient's x-ray image. This feature difference vector is a subtraction of visual feature vectors extracted from the two images. To improve the model, we also exploit tag information obtained from the medical report. Tags provide important information about the images and also convey meaningful semantics to the decoder. Several previous studies (Jhamtani and Berg-Kirkpatrick, 2018; Tan et al., 2019; Forbes et al., 2019) show methods that leverage feature vectors of images to account for differences between two images.

Next, since physicians obtain information not only from the overall image but also from the localized lesion areas, we consider that each convolutional level would also convey meaningful details such as contrast, texture, and localized area. Therefore, another proposed model fully exploits information contained in each layer. Previous studies (Darlow et al., 2018; Bau et al., 2017; Zhou et al., 2018) analyze and interpret convolutional neural networks (CNNs) utilizing feature vectors extracted from lower convolutional layers.

The following section describes the organization of the dataset, and section 3 introduces the baseline and our proposed models. Section 4 provides the experimental settings and results with analysis, and draws some conclusions in Section 5.

2 Dataset

This study uses IU X-RAY, which consists of a series of image-text-tag triplets. This dataset is anonymous and is from the Open Access Biomedical Image Search Engine (OpenI) [1] (Demner-Fushman et al., 2016).

The 7,470 chest x-ray images have two views: posteroanterior (PA) and lateral. The baseline model uses all images, but the proposed model uses only 3,821 images, which are PA views. The report corresponding to each image has four sections: comparison, indication, findings, and impression. The output of the model is a concatenation of the findings and the impression section (Jing et al., 2018). The findings section describes observations in each area of the body, and the most crucial impression section explains the problem and then provides a diagnosis. The output excludes the comparison and indication sections, which contain patient information and symptoms.

One or more tags are automatically extracted from each report using the Medical Text Indexer (MTI)[2] program (Jing et al., 2018). MTI produces index recommendations based on Medical Subject Heading (MeSH)[3] terms. There are a total of 210 unique tags, with an average of 2 tags per image. Without the normal tag, there is an average of 25 images per tag. Class imbalance arises because 1,502 images contain normal tag, so we randomly sample 75 images for a better balance between tags. The tags still have a class imbalance because the scope is too broad, making the term rare.

The prepared datasets are 3,821 image-text-tag triplets, all PA view images. After adjusting the number of images with the normal tag, we use random selection to get 1,911, 238, and 245 triplets for the training, validation, and test sets.

3 Models

3.1 Baseline Model

Among the recent models, the basis is the Jing (2018) model[4]. Our baseline model is similar to this model, which includes a CNN-RNN (encoder-decoder) with an attention mechanism. The Jing (2018) model's encoder part utilizes VGG-19 (Simonyan and Zisserman, 2014) for the visual feature extractor, multi-label classification (MLC) for tag classification, and decoder part uses Hierarchical LSTM (Hochreiter and Schmidhuber,

[1] https://openi.nlm.nih.gov/

[2] https://ii.nlm.nih.gov/MTI/

[3] https://www.nlm.nih.gov/mesh/meshhome.html

[4] Reference code available at https://github.com/ZexinYan/Medical-Report-Generation

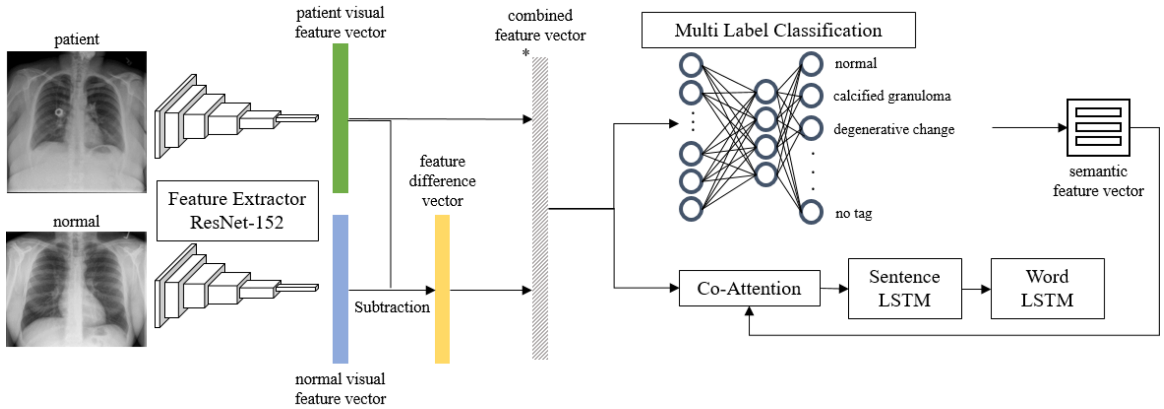


Figure 2: Two difference and tag (DiTag) model structures. The DiTag model uses only feature difference vector and sends it to MLC and co-attention. The combined DiTag (cDiTag) model uses a combined feature vector (*), which is a concatenation of patient visual feature vector and feature difference vector.

1997) with a co-attention mechanism. The only difference between the Jing (2018) model and our baseline model is that we use ResNet-152 (He et al., 2016) instead of VGG-19 to extract the visual feature vector. MLC uses the visual feature vector to predict one or more tags and generates semantic feature vectors that are word embedding of the predicted tags. To obtain an embedding vector of each tag, we train an embedding layer from the training data. Hierarchical LSTM combines sentence LSTM with co-attention and word LSTM. Sentence LSTM creates a topic vector and a stop vector by independently attending to the visual feature vector and semantic feature vector using co-attention. The word LSTM concatenates the topic vector and previous word embedding for a new embedding as input to generate words. The way to get a word embedding vector is the same as the tag, but the embedding matrix is different.

The overall loss is the sum of tag loss, stop loss, and word loss. First, tag loss $\mathrm{L}_{\mathrm{tag}}$ is a cross-entropy loss between predicted tag distributions by MLC and the normalized real tag distributions. Second, stop loss L_{stop} is a cross-entropy loss between predicted stop distributions by Sentence LSTM and ground truth distributions. The stop loss is binary cross-entropy, and the class is stop or continue. Third, word loss L_{word} is a cross-entropy loss between predicted word distribution by Word LSTM and real word distribution. λ_{tag}, λ_{stop}, λ_{word} scale all the losses. The report consists of S sentences, with each sentence having W_s words. Total loss for the baseline model is:

$$L_{base} = \lambda_{tag} L_{tag} + \lambda_{stop} \sum_{s=1}^{S} L^{s}{}_{stop} + \lambda_{word} \sum_{s=1}^{S} \sum_{w=1}^{W_s} L^{s,w}{}_{word} \qquad (1)$$

3.2 Difference and Tag Model

The weakness of our baseline model is that it mainly generates general content (such as "the heart is normal in size" and "the lungs are clear") and does not correctly describe the aspects of the patient image associated with the disease. The model does not adequately capture the differences between the images because the chest x-ray images are similar. Also, when clinicians diagnose patients, they look for the differences between the patient group and the normal group.

Therefore, the first goal of this study was to provide the model with more information about these differences. Our difference and tag (DiTag) model creates a feature difference vector that contains the differences between the patient image and the normal image. The feature difference vector is the result of subtracting the visual feature vector of the normal image from the visual feature vector of the patient image extracted through ResNet-152. The visual feature vector is a global average pooling of feature map produced by the last convolution layer.

We experimented with this feature difference vector using two model structures, as shown in Figure 2. The first structure, the DiTag model, passes the feature difference vector directly to the MLC and the co-attention and does not use the combined feature vector. Co-attention allows the model to attend to the feature difference vector $\{d_n\}_{n=1}^{N}$ and the semantic feature vector $\{t_m\}_{m=1}^{M}$ independently to create a context vector, which is then passed to the sentence LSTM to generate topic vector and stop vector, as shown in Figure 3. The co-attention is only associated with the sentence LSTM, not the word LSTM. The co-attention

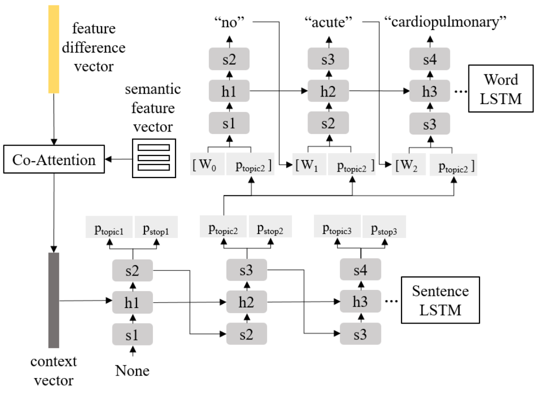


Figure 3: An example of generating a second sentence. For each sentence LSTM step, the co-attention creates a context vector, and the sentence LSTM outputs a topic vector and a stop vector. The word LSTM generates words based on the topic vector and embedding of the previous word.

computes attention score α independently to create a feature difference context vector d^s and a semantic context vector t^s at time step s:

$$d^s = \sum_{n=1}^{N} \alpha_{d,n} d_n \, , \; t^s = \sum_{m=1}^{M} \alpha_{t,m} t_m \quad (2)$$

Concatenate these context vectors, then use a fully connected layer W to obtain the final context vector c^S at time step s:

$$c^s = W[d^s; t^s] \quad (3)$$

A topic vector contains context information by combining the current hidden state of the sentence LSTM and the context vector of the current step. A stop vector decides to stop or continue generating the topic vector and words by combining the previous and current hidden state of sentence LSTM to calculate the probability of stopping. Figure 3 also shows how the word LSTM works.

The second structure is the combined DiTag (cDiTag) model, which sends the combined feature vector that represents the concatenation of the feature difference vector and the patient visual feature vector to the MLC and the co-attention. Co-attention is the same as DiTag model, except that it attends to the combined feature vector rather than the feature difference vector. The overall loss of both structures is the same as the baseline model.

3.3 Multi-Difference and Tag Model

Physicians provide diagnoses using information obtained not only from the overall image but also from localized lesion areas. Therefore, the second goal of this study was to offer lower-level differences to the model, such as the contrast, texture, and localized area. The DiTag model extracts the visual feature vector from the last convolutional layer of ResNet-152, while the mDiTag model further extracts additional visual feature vectors from three lower convolutional layers. Using four visual features from the patient images and four from the normal images, we experimented with the three model structures to compare the effects of model components, as shown in Figure 4.

The mDiTag(-) model subtracts the normal visual feature vector from the patient visual feature vector obtained in each layer to generate four feature difference vectors and then sends all four vectors to the co-attention. The model excludes the MLC, and co-attention attends to the four feature difference vectors and creates a context vector and sends it to the LSTM. Total loss for the mDiTag(-) model is:

$$L_{DiTag} = \lambda_{stop} \sum_{s=1}^{S} {L^s}_{stop} + \lambda_{word} \sum_{s=1}^{S} \sum_{w=1}^{W} {L^{s,w}}_{word} \quad (4)$$

The mDiTag(+) model obtains new visual feature vectors by sending the visual feature vectors of each layer into four different MLCs, one for each layer. The co-attention is identical to that of the mDiTag(-) model. The total loss is the sum of the four tag losses, each occurring in four layers, stop loss and word loss. The model is backpropagated based on the previous four tag losses and then backpropagated based on the overall loss.

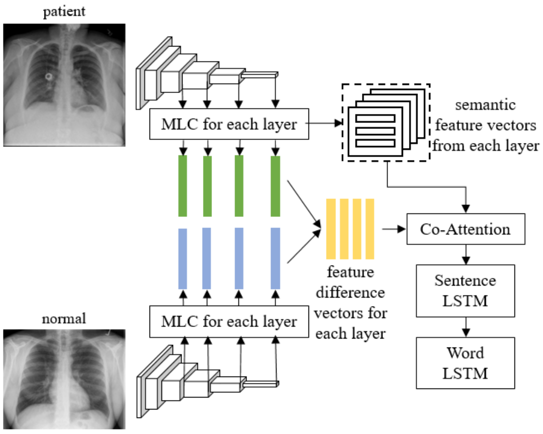


Figure 4: Three mDiTag model structures. The mDiTag(-) model excludes MLC and semantic feature vectors. The mDiTag(+) model excludes only the semantic feature vectors. The whole structure is mDiTag(s) model.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
Baseline model	0.2738	0.1585	0.1045	0.0682	0.2099	0.1226
DiTag model	0.3015	0.1795	0.1204	0.0811	0.2438	0.1939
cDiTag model	0.2501	0.1413	0.0913	0.0597	0.2177	0.0903
mDiTag(-) model	**0.3293**	**0.1985**	**0.1354**	**0.0945**	**0.2731**	**0.1944**
mDiTag(+) model	0.3227	0.1919	0.1271	0.0852	0.2575	0.1829
mDiTag(s) model	0.2086	0.1225	0.0795	0.0566	0.1719	0.1252

Table 1: Metric Evaluation for all models. The DiTag model utilizes feature difference vector, the cDiTag model uses combined feature vector, and the mDiTag models use multiple feature difference vectors. The mDiTag(-) model excludes MLC and semantic feature vectors, the mDiTag(+) model excludes semantic feature vectors, and the mDiTag(s) model uses all. The best model for all metric scores is the mDiTag(-) model.

The mDiTag(s) model is similar to the mDiTag(+) model, but MLC obtains a new visual feature vector and a semantic feature vector. The model sends four feature difference vectors and four semantic feature vectors to the decoder. Co-attention attends to the four feature difference vectors and four semantic feature vectors to create a context vector, and then sends it to the LSTM. The loss function and backpropagation method of this model is the same as that of the mDiTag(+) model. There are four tag losses in each intermediate convolutional layer of mDiTag(+) and mDiTag(s) model. Total loss for these models is:

$$L_{mDiTag} = \lambda_{tag\_1} L_{tag\_1} + \lambda_{tag\_2} L_{tag\_2} + \lambda_{tag\_3} L_{tag\_3} + \lambda_{tag\_4} L_{tag\_4} + \lambda_{stop} \sum_{s=1}^{S} L^{s}{}_{stop} + \lambda_{word} \sum_{s=1}^{S} \sum_{w=1}^{W} L^{s,w}{}_{word} \tag{5}$$

4 Experimental Settings and Results

4.1 Experimental Settings

All model experiments use the same parameters and hyperparameters. For MLC, the number of classes corresponding is 210, the number of classes to predict is 10, and the generated semantic feature vector dimension is 512. In the decoder, the Sentence LSTM is 1 layer, the Word LSTM is 1 layer, the hidden vector dimension is 512, the maximum number of sentences generated is 6, and the maximum number of words created is 30. The learning rate starts from $1e-4$ and is optimized by Adam optimizer. Total epoch is 1,000 but tested with a model of minimum loss. It took four days to train with a 1080Ti GPU with 11G Memory.

4.2 Metric Evaluation

Table 1 provides information on the performance of the models evaluated for the test dataset. We use BLEU score (Papineni et al., 2002), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015) for the metrics. The DiTag model has higher metric scores than the baseline model, and for cDiTag model, only the ROUGE-L score increases. Since the DiTag model structure is more suitable, mDiTag model structures also only utilizes the feature difference vector.

Next, based on all metric scores, the best model is the mDiTag(-) model. When the model includes MLC, the metric score reduces. Since there are two tags per image on average, when predicting 10 tags, there are wrong tag information. Also, the

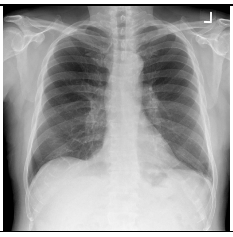
Model	generation result
Baseline Model	no acute cardiopulmonary abnormality the heart is normal in size the heart and lungs have in the interval
mDiTag(-) Model	<num> no acute cardiopulmonary abnormality <num> chronic changes consistent with emphysema the heart is normal in size the lungs are clear no pleural effusion or pneumothorax is seen
mDiTag(+) Model	no acute cardiopulmonary abnormality the heart is normal in size the lungs are clear there is no focal air space opacity to suggest a pneumonia
Ground Truth Report	left base atelectasis lungs otherwise clear there is minimal opacity in the left lung base representing atelectasis the lungs are otherwise clear heart size is normal no <unk>
Image	

Table 2: The first example of the models' outputs with corresponding ground truth report, and image.

significant class imbalance makes MLC challenging to train. Further, when the model uses the semantic feature vector, metric scores reduce. The semantic feature vector is word embedding of the top 10 tags predicted by MLC. However, the semantic feature vector provides incorrect information because of the wrong tags among the 10 predicted tags.

4.3 Analysis of Model Output

Table 2 and Table 3 show examples of the models' output. To make the model outputs easier to see, we eliminate the repeated sentences in the table. The mDiTag(-) model generates more detailed reports than the other models. There are some abnormal findings in the images and ground truth reports in Table 2 and Table 3. The baseline model only explains about the normal findings, while the mDiTag(-) model produces some disease-related sentences, but is not accurate. The outputs show that exploiting multiple feature differences allows the model to generate a relatively diverse explanation of the patient's disease. However, the output still produces general description and does not present enough information about specific features of the disease. As expected, there are incorrect disease descriptions because the tag prediction is not accurate. In addition, as there are too many types of abnormal findings, the terms become too rare to train the model adequately. The components of the text generation part should be modified to resolve the issue of the repeated sentence. Another limitation of this paper is the lack of human evaluation.

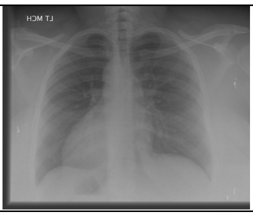
Model	generation result
Baseline Model	no acute cardiopulmonary abnormality the heart is normal in size the lungs are clear
mDiTag(-) Model	<num> no acute cardiopulmonary abnormality <num> left midlung subsegmental atelectasis versus scar the heart is normal in size the mediastinum is unremarkable no pleural effusion or pneumothorax no acute bony abnormality
mDiTag(+) Model	no acute cardiopulmonary abnormality the heart is normal in size the lungs are clear no focal airspace consolidation or pleural <unk>
Ground Truth Report	low lung volumes no acute cardiopulmonary findings the cardiomediastinal silhouette is stable lung volumes remain low there is no pleural line to suggest pneumothorax or costophrenic blunting to suggest large pleural effusion bony structures are within normal <unk>
Image	

Table 3. The second example of the models' outputs with corresponding ground truth report, and image.

5 Conclusion

We propose models that exploit feature differences and tag information. As expected, the model that uses low-level convolutional features from the CNN model can convey low-level details, such as contrast, texture, and localized area. Some of our models outperform the conventional image captioning models in terms of BLEU score, ROUGE-L, and CIDEr. The mDiTag(-) model performs best according to every metric. Based on these experiments, we can conclude that the feature differences between images and semantic tags are crucial elements necessary for training. In the future, we will strengthen tags that contain semantic information to extract keywords for more accurate information, such as disease information, location, and size. Furthermore, improving the accuracy of multiple tag prediction is crucial to deliver semantic facts accurately. We are also considering obtaining more images from hospitals to reduce the proportion of abnormal images in the datasets.

Acknowledgments

The authors would like to thank the three anonymous reviewers and mentor Steven Wilson for their valuable feedback on this work. This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI18C0316).

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, pages 6077-6086. https://doi.org/10.1109/CVPR.2018.00636.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pages 3319-3327. http://doi.org/10.1109/CVPR.2017.354.

Luke Nicholas Darlow, and Amos Storkey. 2018. What Information Does a ResNet Compress? In *Proceedings of the International Conference on Learning Representations*.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310. https://doi.org/10.1093/jamia/ocv080.

Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. 2019. Neural Naturalist: Generating Fine-Grained Image Comparisons. 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)*. Association for Computational Linguistics, pages 708-717. http://doi.org/10.18653/v1/D19-1065.

Ankush Gupta, Yashaswi Verma, and C. V. Jawahar. 2012. Choosing linguistics over vision to describe images. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, pages 606-612.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pages 770-778. https://doi.org/10.1109/CVPR.2016.90.

Sepp Hochreiter, and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Harsh Jhamtani, and Taylor Berg-Kirkpatrick. 2018. Learning to Describe Differences Between Pairs of Similar Images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 4024-4034. https://doi.org/10.18653/v1/D18-1436.

Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the Long-Short Term Memory Model for Image Caption Generation. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, pages 2407-2415. https://doi.org/10.1109/ICCV.2015.277.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 2577–2586. http://doi.org/10.18653/v1/P18-1240.

Alistair E. W. Johnson, Tom J. Pollard, Seth Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR: A Large Publicly Available Database of Labeled Chest Radiographs. https://arxiv.org/abs/1901.07042.

Andrej Karpathy, and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE*, pages 3128-3137. https://doi.org/10.1109/CVPR.2015.7298932.

Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos. 2019. A Survey on Biomedical Image Captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*. Association for Computational Linguistics, pages 26–36. http://doi.org/10.18653/v1/W19-1803.

Polina Kuznetsova, Vicente Ordonez, Tamara L. Berg and Yejin Choi. 2014. TreeTalk: Composition and Compression of Trees for Image Descriptions. *Transactions of the Association for Computational Linguistics*, 2:351-362. https://doi.org/10.1162/tacl_a_00188.

Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing Simple Image Descriptions using Web-scale N-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pages 220-228. https://www.aclweb.org/anthology/W11-0326.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for

Computational Linguistics, pages 74–81. https://www.aclweb.org/anthology/W04-1013.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. In *Computing Research Repository*. http://arxiv.org/abs/1409.1556.

Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. 2019. Expressing Visual Relationships via Language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1873-1883. http://doi.org/10.18653/v1/P19-1182.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pages 4566-4575. https://doi.org/10.1109/CVPR.2015.7299087.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pages 3156-3164. https://doi.org/10.1109/CVPR.2015.7298935.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning (Volume 37)*. PMLR, pages 2048-2057.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-Guided Sentence Generation of Natural Images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 444–454.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning with Semantic Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE*, pages 4651-4659. https://doi.org/10.1109/CVPR.2016.503.

Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. 2018. Interpreting Deep Visual Representations via Network Dissection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 41(9):2131-2145. http://doi.org/10.1109/TPAMI.2018.2858759.

Luowei Zhou, Chenliang Xu, Parker Koch, and Jason J. Corso. 2017. Watch What You Just Said: Image Captioning with Text-Conditional Attention. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. Association for Computing Machinery, pages 305-313. https://doi.org/10.1145/3126686.3126717.