Pushing the Limits of Radiology with Joint Modeling of Visual and Textual Information

Sonit Singh[1,2]
Department of Computing, Macquarie University[1]
DATA61, CSIRO[2]
Sydney, Australia
sonit.singh@hdr.mq.edu.au

Abstract

Recently, there has been increasing interest in the intersection of computer vision and natural language processing. Researchers have studied several interesting tasks, including generating text descriptions from images and videos and language embedding of images. More recent work has further extended the scope of this area to combine videos and language, learning to solve non-visual tasks using visual cues, visual question answering, and visual dialog. Despite a large body of research on the intersection of vision-language technology, its adaption to the medical domain is not fully explored. To address this research gap, we aim to develop machine learning models that can reason jointly on medical images and clinical text for advanced search, retrieval, annotation and description of medical images.

1 Introduction

Integrating information from various modalities is deeply rooted in human lives. Humans combine vision, language, speech and touch to acquire knowledge about the world and comprehend the world (Hall and McKevitt, 1995). *Vision* and *Language* are the most common ways of expressing our knowledge about the world. Both *Computer Vision* (CV) and *Natural Language Processing* (NLP) demonstrated successful results on various general purpose tasks such as image classification, object detection, semantic segmentation, and machine translation. Although research at the intersection of CV and NLP is gaining pace, its applications to healthcare are still under-explored. The success of Artificial Intelligence (AI) technologies in general purpose tasks is mainly attributed to publicly available large-scale datasets, enhanced compute power due to rise of Graphics Processing Units (GPUs), and due to advancements in Machine Learning (ML) algorithms and its various architectures. One of the biggest hurdles in deploying ML (especially Deep Learning) models in healthcare is a lack of annotated data. Although it is easy to get annotated data for general purpose tasks by crowdsourcing, it is almost impossible for medical data because of limited expertise, privacy and ethical issues. On the positive side, a lot of medical data in the form of medical images and accompanying text reports is stored in hospitals' Picture Archival and Communication Systems (PACS). For instance, Beth Israel Deaconnes Medical Center (Harvard) generates approximately 20 terabytes of image data and one terabyte of text data per year (Mastanduno, 2017). Also, the drive toward structured reporting in radiology definitely enhance NLP accuracy (Cai et al., 2016). *Interpreting* medical images and *summarising* them in natural text is a challenging, complex and tedious task. Various research studies show that the general rate of missed radiological findings can be as much as 30% (Berlin, 2001; Berlin and Hendrix, 1998). These errors are mainly due to limited expertise, increasing patient volumes, the subjectivity of human perception, fatigue, and inability to locate critical and subtle findings (Sohani, 2013). Based on a recent estimate one billion radiology examinations are performed worldwide annually. This equates to about 40 million radiologist errors per annum (Brady, 2017). In order to reduce these errors, there is a need to develop automated clinical decision support systems (CDSS) (Eickhoff et al., 2017) that can interpret medical images and generate written reports to augment radiologist's work.

Our research aims to develop machine learning

Proceedings of ACL 2018, Student Research Workshop, pages 28–36
Melbourne, Australia, July 15 - 20, 2018.

models that reason jointly on medical images and clinical text for advanced search, retrieval, annotation and description of medical images. Specifically, we aim to automatically generate description of medical images, to develop medical visual question answering system and to develop medical dialog agents that interact with patients to answer their queries based on their medical data.

2 Background

Deep Neural Networks (DNNs) are a special class of machine learning algorithms that learn in multiple levels, corresponding to different levels of abstraction. In this section, we provide an overview of two of the most common DNNs namely Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Also, we provide a profile of architectures and various successful applications in CV and NLP.

2.1 Convolutional Neural Networks

In the past, problems such as image classification and object detection were approached using a traditional CV pipeline where hand-crafted features were first extracted, followed by learning algorithms (Srinivas et al., 2016). The performance of these systems highly depends upon the quality of the extracted features and the ability of the learning algorithms (Fu and Rui, 2017). As CV progressed, extracting these complex features became a tedious task, giving rise to algorithms that can learn directly from the raw data without the need for hand-crafted feature engineering. The major breakthrough happened in 2012 when object classification on ImageNet (Russakovsky et al., 2015) improved vastly from top-5 error of 25% in 2011 to 16% in 2012. This was the result of shift from hand-engineered features to learned deep features (Felsberg, 2017). AlexNet was the first deep learning model that won the ILSVRC championship in 2012 by drastically reducing the top-5 error rate on the ImageNet Challege compared to the previous shallow networks. Since AlexNet, a series of CNN models have been proposed that advanced state-of-the-art such as VGG-16 (Simonyan and Zisserman, 2014), GoogleNet (Szegedy et al., 2015), and Residual Networks (ResNet) (He et al., 2016). All these models differ in terms of various structural decompositions which led them to have better learning ability and high predictive performance.

2.2 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are special networks that process sequential or temporal data including language, speech, and handwriting. Learning sequential data requires memory of previous states and a feedback mechanism. RNNs form an internal state of the network where connections between units form a cycle, which allows it to exhibit dynamic temporal behavior (Lee et al., 2017). Simple RNNs suffer from the vanishing or exploding gradients problem when trained with gradient based techniques. To overcome these challenges, Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Cho et al., 2014) were introduced which are able to learn very deep RNNs and can successfully remember sequences having duration of varying lengths.

2.3 Joint Image and Language Modeling

Due to the success of deep learning techniques in individual domains of AI including vision, speech and language, researchers are aiming at problems at the intersection of vision, language, knowledge representation and common-sense reasoning. An ultimate goal of CV is to have comprehensive visual understanding that involves not only naming the classes of objects present in a scene, but also describe their attributes and recognize relationship between objects (Krishna et al., 2017). Much progress has been made towards this goal, including object classification (Krizhevsky et al., 2012), object detection and localisation (Girshick et al., 2014), and object and instance segmentation (He et al., 2017a). On the other hand, the overall goal of NLP is to understand, draw inferences from, summarise, translate and generate accurate natural text and language. State-of-the-art results on various NLP tasks, including Part-of-Speech tagging (Collobert et al., 2011), Parsing (Dyer et al., 2015; Vinyals et al., 2015), Named Entity Recognition (Collobert et al., 2011), Semantic Role Labeling (Zhou and Xu, 2015; He et al., 2017b), and machine translation (Sutskever et al., 2014; Wu et al., 2016) are pushing towards that goal. Early work combining vision with language includes image annotation, where the task is to assign labels to an image. However, image annotation only associates isolated words with the image content and ignores the relationships between objects and their relation to the world. To generate

a coherent interpretation of a scene and describe it in a natural way, the task of image captioning emerged within the language-vision community, together with large-scale captioning datasets including Flickr30k [1] and MSCOCO [2]. Captioning involves generating a textual description that verbalizes the most salient aspects (objects, attributes, scene properties) of the image by analyzing it. In order to tackle more complex tasks that combine vision and language and to develop high-level reasoning, Visual Question Answering (VQA) (Stanislaw et al., 2015) was proposed which is equivalent to *Visual Turing Test*. In VQA, the goal is to predict the answer correctly after reasoning over the image and a question in natural text (Teney et al., 2017). To further extend this task, Visual Dialog (Das et al., 2017a,b) was proposed that requires an AI agent to hold meaningful dialog with humans in natural language about visual content. Apart from this, research is moving towards linking language to actions in the real world, also known as language grounding (Chen and Mooney, 2011), which finds applications in human-robot interaction, robotic navigation and manipulation. Although there has been language-vision research for these general purpose tasks, its progress has been underutilised in healthcare.

3 Related Work

Developing clinical decision support has long been a major research focus in medical image processing. In recent years, deep learning models have outperformed conventional machine learning approaches in tasks such as dermatologist level classification of skin lesions (Esteva et al., 2017), detection of liver lesions (Ben-Cohen et al., 2016), detection of pathological-image findings (Zhang et al., 2017a), automated detection of pneumonia from chest X-rays (Rajpurkar et al., 2017), and segmentation of brain MRI (Milletari et al., 2016). Although there are many publicly available datasets for general purpose tasks (Ferraro et al., 2015), there are few publicly available datasets in the medical domain. Recently, (Wang et al., 2017) introduced a large-scale Chest X-ray dataset named *ChestX-ray8* that is publicly available. The dataset consists of $112,120$ frontal-view chest X-rays images of $30,805$ patients. The labels of the images are automatically assigned by applying NLP techniques to the paired radiology reports. (Zhang et al., 2017b) proposed MD-Net, that can read pathology bladder cancer images, can generate diagnostic reports, retrieve images by symptom descriptions, and provide justification of the decision process by highlighting image regions using an attention mechanism. Moreover, (Shin et al., 2016) proposed CNN-RNN model that can efficiently detect a disease in medical image, find the context (*e.g.* location and severity of affected organ) and also correlate the salient regions of the image with Medical Subject Headings (MeSH) terms. They work on Open-i (U.S. NLM), a publicly available dataset that consists of 3955 radiology reports from the Indiana Network for Patient Care, and $7,470$ associated chest X-rays from the hospitals' PACS. Evaluation in terms of BLEU score (Papineni et al., 2002) demonstrates that the model is able to locate diseases and able to generate Medical Subject Headings (MeSH) terms with high precision.

In addition, ImageCLEF challenges[3] have been leading advances in the medical field by promoting evaluation of technologies for annotation, indexing and retrieval of textual data and medical images (Ionescu et al., 2017). Motivated by the need for automated image understanding methods in the healthcare domain, ImageCLEF organized its first *concept detection* and *caption prediction* tasks in 2017 (Eickhoff et al., 2017). The ImageCLEFcaption challenge consists of two sub tasks including Concept detection and Caption prediction. The concept detection task consists of identifying the UMLS Concept Unique Identifiers (CUIs). Majority of the submissions consider concept detection as a *multi-label classification* task. As both of these tasks are inter-related, there has been work where first concepts in the medical images are identified and then captions are generated based on the predicted concepts. (Abacha et al., 2017) consider CUIs in the training set as the labels to be assigned. Two methods namely CNN based approach and the Binary Relevance via Decision Trees (BR-DT) were used. In (Hasan et al., 2017), an encoder-decoder based framework is used where image features are extracted using CNN and RNN-based architecture with attention mechanism is used to translate the image features to relevant captions. In (Rahman et al., 2017), a

[1] http://shannon.cs.illinois.edu/DenotationGraph/

[2] http://cocodataset.org/

[3] http://www.imageclef.org/

Content Based Image Retrieval (CBIR) based approach is used where first images in both training and validation sets are indexed by extracting several low-level color, texture and edge-related visual features. The similarity search is then used to find the closest matching image in the train (or validation set) for each each query (test) image for caption prediction. For similarity matching, each feature is concatenated to form a combined feature vector and Euclidean distance is used for k-Nearest Neighbor image similarity.

In the ImageCLEF challenge, submissions varried in their usage of external resources. For instance (Hasan et al., 2017) do semantic preprocessing of captions using MetaMap and UMLS meta-thesaurus. Pre-training CNN models on PubMed Central images helped in boosting effectiveness compared to training on general purpose ImageNet dataset. Although these challenges provide labeled medical images for modality classification and concept predictions, the datasets are still much smaller (thousands of images) than the ImageNet dataset (Russakovsky et al., 2015) which contains 1.2 million natural images. Moreover, there are issues with ImageCLEFcaption dataset as the UMLS concepts are extracted using probabilistic process which introduces errors. The analysis of dataset showed that some of the images had no concepts attached.

Learning image context from the corresponding clinical text and generating textual reports very similar to radiologists has not yet been achieved. With recent advancements in machine learning (specially deep learning), it is not hard to imagine an opportunity to aid radiologists by developing multimodal clinical decision support systems.

4 Proposed Research

We identify research gaps in the intersection of medical imaging, computer vision and natural language processing as listed in in the following research questions. Our work will address some of these gaps.

How to automatically generate a radiology report for a given medical image?

In medical imaging, the accurate diagnosis or assessment of a disease depends on both image acquisition and image interpretation. While image acquisition has improved substantially due to faster rates and increased resolution of the acquisition devices, image interpretation is still performed by a radiologist, where the radiologist has only a few minutes with an imaging study to describe the findings in the form of radiology report. Such reporting is a time-consuming task and often represents a bottleneck in the clinical diagnosis pipeline (Ionescu et al., 2017). We will develop machine learning models that automatically generate radiology reports by interpreting medical images in order to augment the radiology practice.

How to develop a question answering system that can reason over medical images?

There has been growing interest in AI to support clinical decision making and in improving clinical work-flow by better patient engagement. Automated systems that can interpret complex medical images and provide findings in natural language text can significantly enhance the productivity of hospitals, reduce burden on radiologists and provide a "second opinion", leading to reduced errors in radiology practice. VQA (Stanislaw et al., 2015) has been successful on generic images, but it has not been explored in the medical domain. We will develop machine learning models that combine NLP and CV techniques to answer clinically relevant questions based on medical images. Subsequently, clinical visual dialog systems could be developed based on the models for medical VQA. The dialog agent will respond to patient's queries in an interactive manner based on medical images, clinical text reports and past history of the patient.

How to annotate medical images from the accompanied radiology reports in a weakly supervised manner?

A large volume of medical imaging data and text is accumulated in hosptitals' PACS. To harness this data for advancing healthcare is challenging. Manual annotation of medical data is almost impossible due to the complex nature of medical images, requirement of domain expertise, privacy, ethics and healthcare data regulations. The processing of clinical text is challenging due to combinations of ad-hoc formatting, eliding words which can be inferred from context, and liberal use of parenthetical expressions, jargon and acronyms to increase the information density. We will explore NLP techniques to annotate medical images from the accompanying radiology reports.

How to highlight the relevant area in a medical image based on the features extracted from radiology reports?

Although machine learning, especially deep learning, models have been successful in various domains, they are often treated as black boxes. While this might not be a problem in other more deterministic domains such as image annotation (where the end user can objectively validate the tags assigned to the images), in health care, not only the quantitative algorithmic performance is important, but also the reason why the algorithms works is relevant. In fact, model interpretability is crucial for convincing medical professionals of the validity of actions recommended by predictive systems. We will develop models using CV, NLP and attention mechanisms which highlight the relevant area in a medical image based on the feature extracted from the radiology reports.

How to train machine learning models when data is small or classes are imbalanced?

Obtaining datasets in the medical imaging domain that are as comprehensively annotated as ImageNet remains a challenge. When sufficient data is not available, transfer learning or fine tuning are the ways to proceed. In transfer learning, CNN models pre-trained from natural image dataset or from a different medical domain are used for a new medical task at hand. On the other hand, in fine-tuning, when a medium sized dataset does exist for the task at hand, one suggested scheme is to use a pre-trained CNN as initialisation of the network, following which further supervised training is conducted, of selected network layers, using the new data for the task at hand. In this task, we will explore the effectiveness of transfer learning and fine-tuning in the medical domain.

How to incorporate the temporal nature of diseases in machine learning models?

Diseases evolve and change over time in a non-deterministic manner. The existing deep learning models assume static vector-based inputs, which do not take time factor into consideration. In order to understand the temporal nature of healthcare data, we need to develop deep learning models whose parameters gets incrementally updated with time. Considering that the time factor is important in all kinds of healthcare problems, training a time-sensitive machine learning model is critical for a better understanding of the patient condition and providing timely clinical decision support. We will work towards exploring ways of how to incorporate temporal information in the machine learning models to have temporal reasoning. This will help in understanding the progressive nature of diseases and to alert medical staff about the changing conditions of patients at right time.

How to increase the number of features to improve performance and robustness of CDSS?

Due to rise of Electronic Health Records (or EHR), hospitals store data in various forms including patient's medical history, demographics, progress notes, medications, vital signs, immunizations, laboratory data, genetics and genomics data, and radiology reports. Combining two or more modalities allows integration of the strengths of individual modalities. We will work towards combining various data sources in healthcare so that better decisions can be made, in turn resulting in achieving the overall goal of precision medicine.

How to develop bi-directional models for medical indexing and retrieval?

With the widespread use of EHR and PACS technology in hospitals, the size of medical data is growing rapidly, which in turn demands effective and efficient retrieval systems. Clinical and radiology practices heavily rely on processing stored medical data providing aid in decision making and increasing productivity. Existing medical retrieval systems have limitations in terms of the *semantic gap* (between the low level visual information captured by imaging devices and the high level semantics perceived by humans) (Qayyum et al., 2017). We will develop bi-directional multimodal machine learning models that perform retrieval based on both textual and visual content. The proposed approach can retrieve medical images either based on the textual query as an input or by providing sample query images. In addition, the developed model can also align images and text in large medical data collections.

5 Expermental Framework

5.1 Datasets

The proposed research work has approval from Macquarie University Human Research Ethics Committee to use medical data from Macquarie University Hospital. We will also use datasets

that are publicly available such as ChestX-Ray8, Open-i [4], and ImageCLEF [5] challenge datasets. These datasets comprise of medical images and their accompanied text in the form of disease labels or caption, mined from open source biomedical literature and image collections.

5.2 Evaluation Metrics

For medical captioning task, we will use standard image captioning metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). For VQA in the medical domain, we will use *accuracy* for multiple-choice questions, but to measure how much a predicted answer differs from ground truth based on differences in their semantic meaning, *Wu-Palmer Similarity* (Wu and Palmer, 1994) will be used. For the Visual-Dialog task in the medical domain, an algorithm has to return candidate answers for a given medical image, dialog history, question, and a list of candidate answers. We will use two standard retrieval metrics namely, *recall@k* and *mean reciprocal rank* (MRR) (Das et al., 2017a). In the task of medical retrieval system, the evaluation task is to measure how effectively an algorithm is able to produce search results to satisfy the user's query in the form of sample image or complex textual query. For this task, standard information retrieval metrics such as Precision, Recall, and F-score will be used.

5.3 Baseline Methods

There are three main approaches to generate image captions: (1) Using templates that rely on detectors and map the output to linguistic structures; (2) Using language models that yield more expressive captions overcoming the limitations of template based approach; and, (3) Caption retrieval and recombination that involves retrieving captions based on training data instead of generating new captions. We will work on CNN-RNN framework and caption retrieval approaches. The model proposed by Hasan et al. (2017) was ranked first in the caption prediction task in the ImageCLEF challenge, which was based on deep learning approach using language models. Apart from this, deep learning methods have demonstrated successful results in general purpose image captioning, therefore the first baseline method is to incorporate an encoder-decoder based architecture. Specifically, initial image features will be extracted using a CNN model, namely VGG-19, which is pre-trained on the ImageNet dataset and is fine-tuned on the given ImageCLEF training dataset to extract the image features from a lower convolution layer such that the decoder can focus on the salient aspects of the image via an attention mechanism. Second, text features will be extracted and pre-processed. Two reserved words namely *start* and *end* are appended to indicate the start and end of the captions. While training, the output of the last hidden layer of the CNN model (Encoder) is given to the first time step of the LSTM (decoder). We set $x1$ = start and the desired label, $y1$ = first word of the caption. Similarly, we set the all the remaining words and finally the last target label y^T = end token. The model will be trained with an adaptive learning rate optimization algorithm, and dropout as a regularization mechanism. The model hyper-parameters are tuned based on the BLEU score on the validation set. Once the model is trained, captions are generated on the test images by predicting one word at every time step based on the context vector, the previous hidden state, and the previously generated words.

6 Conclusion

We argue the need for language and vision research in the medical domain by showing its successful applications on general purpose tasks. We identify various research directions in the medical imaging applications that have not been fully explored, and can be solved by combining vision and language processing. This research aims to develop machine learning models that jointly reason over medical images and accompanying clinical text in radiology. The proposed research is fruitful in advancing healthcare by building various clinical decision support systems to augment radiologist's work.

Acknowledgements I would like to thank my supervisors, Dr. Sarvnaz Karimi, Dr. Kevin Ho-Shon and Dr. Len Hamey, for their feedback that greatly improved the paper. This research is supported by Macquarie University Research Training Program scholarship. Also thankful to Google for providing travel grant to attend the conference.

[4] https://openi.nlm.nih.gov/
[5] http://www.imageclef.org/

References

Asma Ben Abacha, Alba G. Seco de Herrera, Soumya Gayen, Dina Demner-Fushman, and Sameer Antani. 2017. NLM at ImageCLEF 2017 Caption Task. In *CLEF2017 Working Notes*, CEUR Workshop Proceedings, Dublin, Ireland.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *European Conference on Computer Vision*, Amsterdam, The Netherlands.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, United States.

Avi Ben-Cohen, Idit Diamant, Eyal Klang, Michal Amitai, and Hayit Greenspan. 2016. Fully convolutional network for liver segmentation and lesions detection. In *Deep Learning and Data Labeling for Medical Applications*, pages 77–85, Cham. Springer International Publishing.

Leonard Berlin. 2001. Defending the "missed" radiographic diagnosis. *American Journal of Roentgenology*, 176(2):863–867.

Leonard Berlin and Ronald W. Hendrix. 1998. Perceptual errors and negligence. *American Journal of Roentgenology*, 170(4):863–867.

Adrian P. Brady. 2017. Error and discrepancy in radiology: inevitable or avoidable? *Insights into Imaging*, 8(1):171–182.

Tianrun Cai, Andreas A. Giannopoulos, Sheng Yu, Tatiana Kelil, Beth Ripley, Kanako K. Kumamaru, Frank J. Rybicki, and Dimitrios Mitsouras. 2016. Natural language processing technologies in radiology research and clinical applications. *RadioGraphics*, 36(1):176–191. PMID: 26761536.

David L. Chen and Raymond J. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pages 859–865, San Francisco, California.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1080–1089, Hawaii, United States.

Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979, Venice, Italy.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack Long Short-Term Memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343, Beijing, China.

Carsten Eickhoff, Immanuel Schwall, Alba Garcia Seco de Herrera, and Henning Müller. 2017. Overview of ImageCLEFcaption 2017 - Image Caption Prediction and Concept Detection for Biomedical Images. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118.

Michael Felsberg. 2017. Five years after the deep learning revolution in computer vision: State of the art methods for online image and video analysis. *Linkoping: Linkoping University Press*, pages 1–13.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. 2015. A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 207–213, Lisbon, Portugal.

Jianlong Fu and Yong Rui. 2017. Advances in deep learning approaches for image tagging. *APSIPA Transactions on Signal and Information Processing*, 6:e11.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, Washington, DC, United States.

Peter Hall and Paul McKevitt. 1995. Integrating vision processing and natural language processing with a clinical application. In *Proceedings 1995 2nd New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, pages 373–376, Dunedin, New Zealand.

Sadid A. Hasan, Yuan Ling, Joey Liu, Rithesh Sreenivasan, Shreya Anand, Tilak Raj Arora, Vivek Datla, Kathy Lee, Ashequl Qadir, Christine Swisher, and Oladimeji Farri. 2017. PRNA at ImageCLEF 2017 Caption Prediction and Concept Detection tasks. In *CLEF2017 Working Notes*, CEUR Workshop Proceedings, Dublin, Ireland.

Kaiming He, Georgia Gkioxari, Piotr Dollr, and Ross Girshick. 2017a. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision*, pages 2980–2988, Venice, Italy.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, Nevada, United States.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017b. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Bogdan Ionescu, Henning Müller, Mauricio Villegas, Helbert Arenas, Giulia Boato, Duc-Tien Dang-Nguyen, Yashin Dicente Cid, Carsten Eickhoff, Alba G. Seco de Herrera, Cathal Gurrin, Bayzidul Islam, Vassili Kovalev, Vitali Liauchuk, Josiane Mothe, Luca Piras, Michael Riegler, and Immanuel Schwall. 2017. Overview of ImageCLEF 2017: Information extraction from images. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 315–337, Cham. Springer International Publishing.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA.

JG Lee, S Jun, YW Cho, H Lee, GB Kim, JB Seo, and N Kim. 2017. Deep learning in medical imaging: General overview. *Korean Journal of Radiology*, 18(4):570–584.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *42nd Annual Meeting of the Association for Computational Linguistics*, volume Text Summarization Branches Out, pages 1–8, Barcelona, Spain.

Mike Mastanduno. 2017. Survey of deep learning in radiology. [Online; posted 19-January-2017].

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *CoRR*, abs/1606.04797.

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*, Philadelphia, Pennsylvania, United States.

Adnan Qayyum, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid. 2017. Medical image retrieval using deep convolutional neural network. *Neurocomputing*, 266:8 – 20.

Mahmudur Rahman, Terrance Lagree, and Martina Taylor. 2017. A cross-modal concept detection and caption prediction approach in ImageCLEFcaption track of ImageCLEF 2017. In *CLEF2017 Working Notes*, CEUR Workshop Proceedings, Dublin, Ireland.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. 2017. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *CoRR*, abs/1711.05225.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M. Summers. 2016. Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2497–2506.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

C. A Sohani. 2013. A difficult challenge for radiology. *The Indian Journal of Radiology and Imaging*, 23(1):110–112.

Suraj Srinivas, Ravi Kiran Sarvadevabhatla, Konda Reddy Mopuri, Nikita Prabhu, Srinivas S. S. Kruthiventi, and R. Venkatesh Babu. 2016. A taxonomy of deep convolutional neural nets for computer vision. *Frontiers in Robotics and AI*, 2:36.

Antol Stanislaw, Agrawal Aishwarya, Lu Jiasen, Mitchell Margaret, Batra Dhruv, Zitnick C. Lawrence, and Parikh Devi. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision*, pages 2425–2433, Santiago, Chile.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, United States.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, Boston, Massachusetts, United States.

Damien Teney, Qi Wu, and Anton van den Hengel. 2017. Visual Question Answering: A tutorial. *IEEE Signal Processing Magazine*, 34(6):63–75.

NIH U.S. NLM. Open-i: An open access biomedical search engine. `https://openi.nlm.nih.gov/`.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, Boston, Massachusetts, United States.

Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2773–2781, Cambridge, MA, United States.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. 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*, pages 3462–3471, Hawaii, United States.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico.

Zizhao Zhang, Pingjun Chen, Manish Sapkota, and Lin Yang. 2017a. TandemNet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In *Medical Image Computing and Computer-Assisted Intervention MICCAI 2017*, pages 320–328, Quebec City, Quebec, Canada.

Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. 2017b. MDNet: A semantically and visually interpretable medical image diagnosis network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3549–3557, Hawaii, United States.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1127–1137, Beijing, China.