Medical Report Generation from Jointly Learned Medical Image and Text Representations

Modestas Filipavicius, Marilena Oita, Fabien Pernot

2019-12-31

Novartis NIBR Text Mining Services, Basel, Switzerland

1 Abstract

This Novartis internship at NIBR Text Mining Services explored multimodal deep learning applications for medical images and text, in particular, automatic report generation from chest X-ray images, used daily in hospitals diagnose chest diseases. Reading the images and writing reports requires considerable training and time-investment. We speculate that to match the physician-level disease understanding, the representations learned in unsupervised manner for images and text should be jointly embedded into the same vector space. As such, we could perform cross modal retrieval, i. e., ask the model to generate a paragraph with Findings and Impressions sections for a given image.

To learn the necessary skillset and get the first-hand experience for training a multimodal encoder-decoder architecture and its computational resource requirements, we first re-implemented several automatic image captioning models based on Microsoft COCO dataset. Next, we attempted to address the issues specific for medical paragraph generation, namely, generating several rather than a single sentence and associating each word with the relevant image patch.

We focused on an image-encoder text-decoder architectural variant called Hierarchical LSTM Co-Attention model by Jing et al. (2018), and implemented this closed-source paper in Python and PyTorch. Although, our implementation could not verify their results, we are encouraged to have successfully applied the earlier encoder-decoder captioning (with attention) to Open-I dataset. We hope this exploratory study will encourage future research into generating representations with multimodal learning in pharmaceutical setting.

Contents

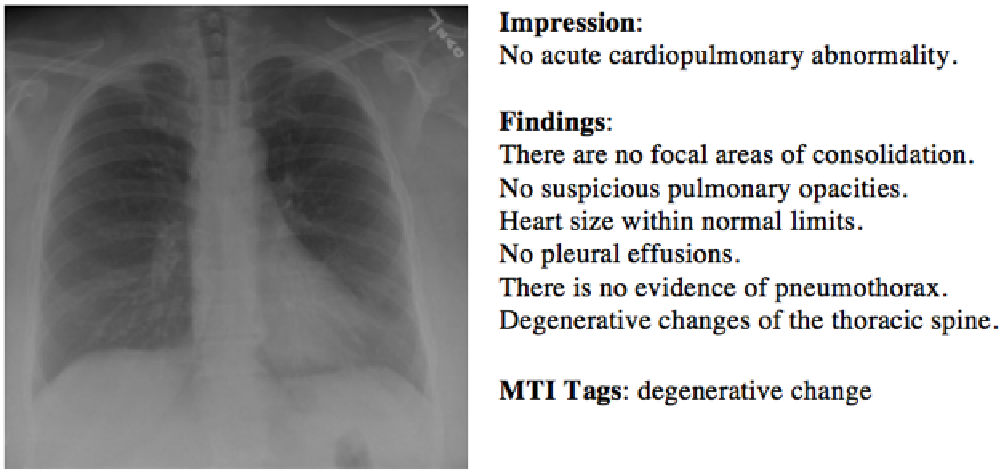
2 Introduction

Medical images such as chest radiograms, histopathology, retina and skin images are often used to diagnose and treat patients. Highly trained physicians are tasked with understanding and interpreting such images. They write a narrative text report in which they describe the state the organs examined, and state any anomalies detected. Additionally, a set of tags for the suspected diagnosis is provided.

Since writing such reports is time consuming and requires specialized training, it is not surprising that areas with low quality healthcare are most affected by time and financial burdens. An automatic report generation of medical images is one obvious approach to alleviate the problems above. Our task here is to create a model which, upon inputting any biomedical image, generates appropriate sentences describing the image objects, their relations and the overall clinical impact of the particular image.

Inevitably, such system must address the following challenges. First, the reports have several heterogeneous categories. For instance, report for a chest X-ray contains ”impression” section which is a single sentence (equivalent to regular image captioning problem), ”findings” section is a paragraph, and ”tags” which are a list of keywords (see **Figure 1**). This can be solved in a multitask setting by treating ”findings” generation as a hierarchical text generation problem which utilizes the tag-image embeddings (tags obtained from ”tagging” as a multi-label task). Second, correct location and classification of abnormal regions, and generating narrations for them is at the core of our task, the solution to it necessitates the use of a powerful joint image-text embedding strategy. Third, providing visual evidence for a particular generated sentence or a word is a desirable feature for a production-grade medical system. Attention mechanisms are used to provide such evidence.

In this exploratory study we will first overview the current state of the art in multimodal image-text learning, and the early methods for generating reports for chest x-rays. Second, we will experiment with the latest models and assess their suitability for medical report generation.



Impression:
No acute cardiopulmonary abnormality.

Findings:
There are no focal areas of consolidation.
No suspicious pulmonary opacities.
Heart size within normal limits.
No pleural effusions.
There is no evidence of pneumothorax.
Degenerative changes of the thoracic spine.

MTI Tags: degenerative change

Figure 1: Example of an annotated chest X-ray image, taken from [45]

Our contributions to Novartis TMS are:

- Tested traditional encoder-decoder image captioning models, assessed the training procedure and computational infrastructure requirements

- Adapted three popular image captioning models (at single sentence level) to chest x-ray captioning. These are used as baselines to evaluate the hierarchical LSTM co-attention model.

- Attempted an open source implementation of Jing et al. (2018) paper [21]

3 Literature Overview

In this section we will overview the seminal recent publications in the fields of representation and multimodal learning, as well as recent progress in image captioning task such as the use of encoder-decoder architectures and self-attention.

3.1 Importance of Representation Learning

Good representation of the data is crucial for downstream predictive tasks. In classical probabilistic learning framework such representation captures the underlying posterior distribution of explanatory factors, while in deep learning framework the input data is passed through a network of nonlinear functions which yields more abstract and thus useful representations [7]. According to [7] features of useful representations include smoothness, sparsity, expressivity, hierarchical nature of explanatory factors, and many others.

Typically deep representation learning is done with Encoder-Decoder architecture, while **autoencoder** [16] and sequence-to-sequence [47] being the most common models. The encoder is simply a function that maps an **input space** to a **latent space**, and the decoder is another function that maps the latent space to a **target space**. One can design an encoder-decoder system using any neural network components, such as CNN or RNN, to encode the complex input into a compressed latent space representation, and decode the representation to a target output. The latent representation, also known as **embedding**, located in the network layers between the encoder and decoder is the learned representation for a given input.

Autoencoders transform a complex input into a compressed representation, which can be translated (decoded) into a reconstruction of input, while minimizing the reconstruction loss. Such compressed representation captures, or encodes, the essential data features while ignoring non-salient features. Sequence-to-sequence, is also used for Neural Machine Translation tasks [47].

3.2 Language Representations - Word2Vec and BERT

Word embeddings are motivated by the limitations of traditional representations for words, such as a one-hot encoding or bag of words, which are high dimensional and inefficient, since their encodings capture none of the similarity or correlation information between words in the source text. For example, in a corpus composed of three words ”I”, ”eat”, ”pizza” the Euclidean distance between each encoding [1 0 0], [0 1 0], [0 0 1] is exactly 1, and therefore carries no useful distance information.

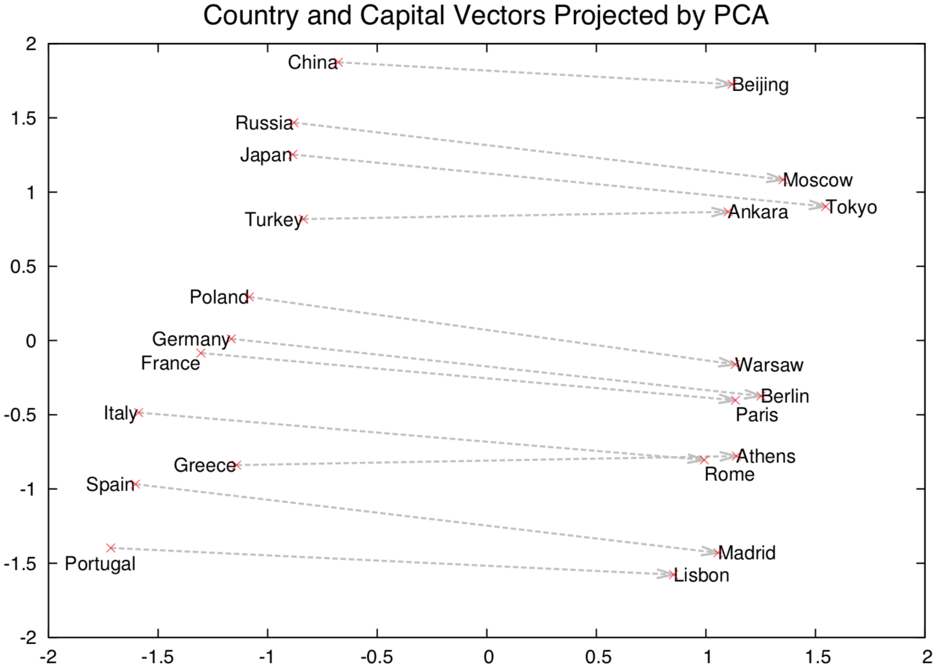


Figure 2: 2D PCA projection of 100 dimensional skip-gram vectors of country-capital pairs. Geographically and geopolitically close countries and their capitals cluster together. Also, the concept of a country capital can be expressed as an angle. Adapted from [36]

Conversely, the main idea behind **Word2Vec**, a popular word embedding is that a word’s meaning is largely captured by its context - neighboring words [36]. W2v embedding models this contextual information by taking a word w_t and predicting its context with a **skip-gram** model, which aims to maximize the mean log likelihood:

$$1/T \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t),$$

where c is context window size. The model creates a lower-dimensional space such that words that appear in similar contexts will also be close together in this new space, see Figure 2. Dimensionality for each word in the corpus is typically set to 200 or 300.

The main application of such word embeddings has been in the use of **transfer learning**, where embeddings are first learned using extremely large sources of unlabeled general text (from web-crawls, Wikipedia dumps, etc), and then used in supervised learning with recurrent neural networks which accepts the pre-trained embeddings as inputs.

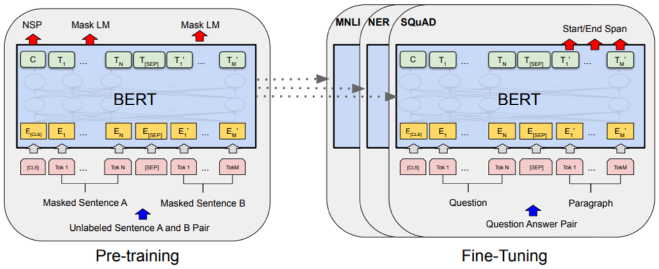


Figure 3: Training BERT (left) consists of reading a sequence of w2v word embeddings at once, passing them through a Transformer encoder and predicting the contexts of each word (masked out 15% of sequence words). Such pretrained model can be fine-tuned to succeed in MNLI, NER, SQuAD tasks (right) [9]

.

Recently, pretrained high-capacity language models such as **ELMo** [39] and **BERT** [9] have become increasingly important in NLP. They are optimised to either predict the next word in a sequence or some masked word anywhere in a given sequence ("Obama was born in [mask] in the year 1967"). For example, BERT reads a large sequence of words all at once (typically it's done one-by-one) to learn all the words' contexts at once, see **Figure 3**.

It utilizes a transformer architecture [50] for the encoder with w2v word embeddings as inputs, and an additional classifier layer on top of encoder to predict masked out 15% of words in the sequence. Finally, we get the soft-max probabilities for each word at each sequence position.

Such models with millions of parameters capture huge amounts of linguistic knowledge to facilitate downstream tasks. This knowledge is usually accessed either by conditioning on latent context representations produced by the original model or by using the original model weights to initialize a task-specific model which is then further fine-tuned. This type of knowledge transfer is crucial for current state-of-the-art results on a wide range of tasks.

Language models pretrained on generic corpora (such as Wikipedia) can be applied to a more specialized domain by transfer learning[40]. For example, recently Beam et al., [6] published a comprehensive set of embeddings for medical concepts, *cui2vec*, by combining extremely large sources of multimodal healthcare data. For a review of recent trends in deep learning NLP see [59].

3.3 Image Representations

Convolutional neural networks (**CNN**) such as LeNet [29], AlexNet [28], VGGNet [46], GoogleNet [48], and ResNet [15] are trained on ImageNet dataset for classification, detection and other vision tasks, for which they outperform human subjects.

CNNs can be easily integrated into multimodal learning models (see 3.4) and trained jointly with other modalities like text. However, they take significant time and computational resources to achieve human-level accuracy. Thus, pre-trained versions of CNNs are used instead, namely, by taking the penultimate CNN layer weights (for classification task, before SoftMax function) which are good image representations and inputting them to the multimodal model.

However, popular datasets like ImageNet contain images of generic everyday life scenes which are very different from medical images. To apply CNN models in medical imaging domain, they are fine-tuned. For example, Inception CNN model was fine-tuned by [10] to classify skin lesions into malignant or benign, achieving results close to the ones predicted by dermatologists. This proved that CNNs, pre-trained on ImageNet, can be successfully fine-tuned for medical imaging tasks, despite the differences between general and medical images. In another seminal work, [41] encoded chest X-ray images with a pretrained DenseNet-121 architecture to predict 14 types of thoracic diseases.

Practically, **penultimate layers** of various CNN-based architectures are used. For example, VGGNet, a 4096-dimensional embedding [46], is often chosen as the input into image-text joint embedding procedure. Also, the full-sized input image can be encoded with the final convolutional layer of Resnet-101 [15]. Facebook's Pythia uses fc6 and fc7 layers from Detectron (based on ResNeXt [56]).

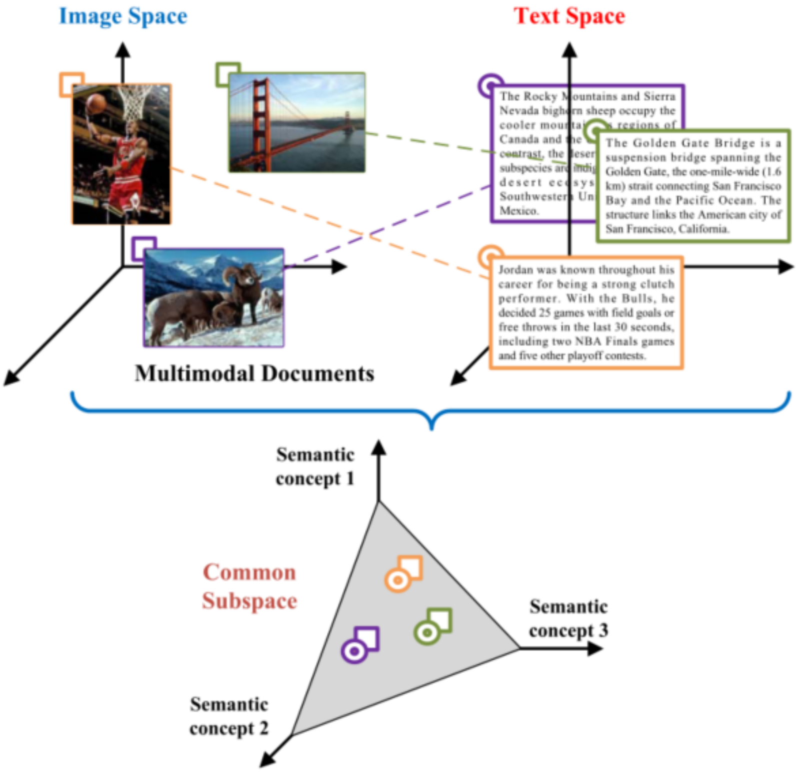
3.4 Multimodal Learning

What is multimodal learning? The objects or concepts in our world can be represented by different signal classes, often recorded by a different instrument, such as sound, text, image, video, graph, etc. Therefore, a "modality" here refers to a particular way of acquiring information about the object, and representation learning from several classes or modalities of signals is known as **multimodal learning**.

Why it's useful to combine two modalities? In multimodal learning setting we aim to learn a shared representation from different modalities for the same phenomenon. Consequently, more information is learned about the phenomenon than if only one mode was considered. Using these multimodal representations has paved ground for more accurate models compared to unimodal models. For a broader review of multimodal benefits and current challenges see [5] and [14].

As seen in **Figure 4**, the feature vectors from text and image modalities are originally located in unequal subspaces, that is the correlations between the two modalities are highly nonlinear, thus proving hard to learn. Therefore, the vector representations associated with similar semantics would be completely different, a problem known as heterogeneity gap. Consequently, subsequent learning algorithms will not make accurate predictions. Multimodal representation learning aims to project the heterogeneous data of different modalities into a shared vector subspace, where the multimodal data with similar semantics will be represented by similar vectors. Effectively, we are lowering the distribution gap in a joint semantic space while keeping the modality specific semantics intact [42].

In practice, integrating individual feature representations into a multimodal one is achieved by three common frameworks with distinct architectures: **joint representation, coordinated representation, and encoder-**



The Rocky Mountains and Sierra
Nevada bighorn sheep occupy the
cooler mou ta s regions of
Canada and the
contrast, the deser
subspecies are indig
desert ecosys
Southwestern Un
Mexico.

The Golden Gate Bridge is a suspension bridge spanning the Golden Gate, the one-mile-wide (1.6 km) strait connecting San Francisco Bay and the Pacific Ocean. The structure links the American city of San Francisco, California.

Jordan was known throughout his career for being a strong clutch performer. With the Bulls, he decided 25 games with field goals or free throws in the last 30 seconds, including two NBA Finals games and five other playoff contests.

Figure 4: Multimodal learning allows to learn a common subspace in which heterogeneous data of two different modalities (circles and squares) is projected into a common vector space. As a result, the data with similar semantics will be represented by similar vectors. Adapted from [42]

decoder, Figure 5. Joint representations are projected to the same space using all of the modalities as input, in the simplest case, by concatenating the two feature vectors. Meanwhile the coordinated representations exist in their own space, but are coordinated through a similarity (e.g. Euclidean/cosine distance) or structure constraint (e.g. partial order). Finally, encoder-decoder architecture maps source modality to a latent (representation) vector, then decoder generates a new sample of target modality. Such architecture, first used in neural machine translation, has proved a leading method for image captioning task.

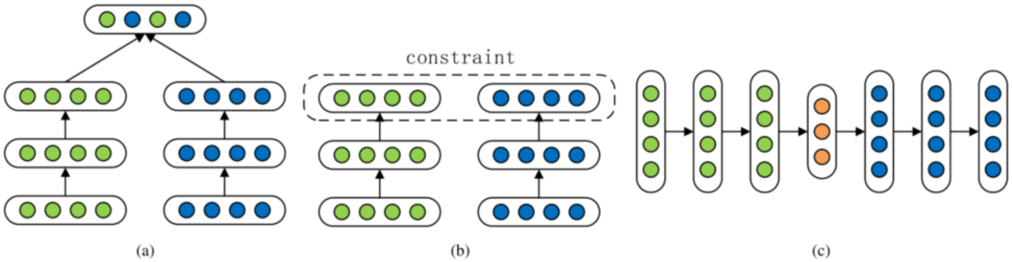


Figure 5: Approaches to learn deep multimodal representations. (a) Joint approach – project individual representations (notice several layers) from two modalities to the shared semantic space so that they can be fused. (b) Coordinated approach permits individual modalities to remain in their own space, but they are coordinated through a similarity (cosine distance) or structure constraints. (c) Encoder-Decoder maps source modality to a latent representation, from which the decoder generates a new sample of target modality. Such framework translates one modality into another while keeping their semantics consistent.

3.5 Image Captioning - Multimodal Learning Task with Encoder-Decoder Architecture

Deep learning has revolutionised the computer vision field. Image classification, where the task is to assign one label to an image, was the first success story in deep learning. Then followed object detection: identifying and labeling multiple salient regions of an image. Most recently, the image captioning task expanded the complexity of the label space from a fixed set of categories (1,000 possible labels in ImageNet dataset) to sequence of words able to express significantly richer concepts [25]. Now it is multimodal learning's turn to uproot computer vision!

A big part of multimodal machine learning is concerned with **translating** (mapping) from one modality to another. Given an entity in one modality the task is to generate the same entity in a different modality. For example, given an image we might want to generate a sentence describing it or given a textual description generate an image matching it. To solve it, we not only need to fully understand the visual scene and to identify its salient parts, but also to produce grammatically correct and comprehensive yet concise sentences describing it.

Since 2014 a steady wave of image captioning models based on encoder-decoder framework has flooded the conference halls. This upsurge was enabled by the success of such frameworks in **neural machine translation** task. Indeed, image captioning with encoder-decoder can be seen as machine translation of an image into a description.

[25] first proposed **multimodal neural language models** - models of natural language that can be conditioned on other modalities, such as images. An image-text multimodal neural language model can be used to retrieve images given complex sentence queries, retrieve phrase descriptions given image queries, or generate text conditioned on images. In their seminal work [26] demonstrated that we can jointly learn word representations and image features by training a CNN-based model with image and text data. However, the joint embedding is done via two separate pathways, and even though they can generate text, their approach is highly tuned for the ranking task by finding the best encoder. Thus unseen objects cannot be recognized.

Karpathy et al. [24] introduced a model of **bidirectional retrieval** of images and sentences. Unlike previous works, they do not map images or sentences into a common space. Instead, their model works at a finer scale and embeds fragments of images and fragments of sentences into a common space. The sentence fragments are represented as dependency tree relations that are based on the dependence tree of the sentence, and the image fragments are represented by a CNN. First, objects in the image are detected using Region Convolutional Neural Network (RCNN). The top 19 detected locations and the entire image are used as image fragments. Each image fragment is embedded using a CNN which takes the image inside a given bounding box and returns the embedding. Finally, they suggest a similarity score for any image-sentence pair.

The next big success came with an **encoder-decoder** model by Vinyals et al. [52], the first end-to-end image captioning model. As such, the image is shown to the RNN at the beginning, unlike in [26], where the model sees the image at each time step of the output word sequence. Both above methods, represent images as single feature vectors from the top layer of pre-trained CNN.

3.6 Image Captioning with Attention Models

The above methods compressed an entire image into a static representation and did not generate good captions when multiple objects or complex scenes were present. **Attention** to the rescue! Attention is a way of obtaining a weighted sum of the vector representations of a layer in a neural network model [4]. Attention layer allows for important features to come forward as needed, and elucidates what the model ”sees”. It is used in diverse tasks ranging from machine translation, language modeling to image captioning, and object recognition. Apart from substantial performance benefit, attention also provides interpretability to neural models, which are usually criticized for being black-box function approximators.

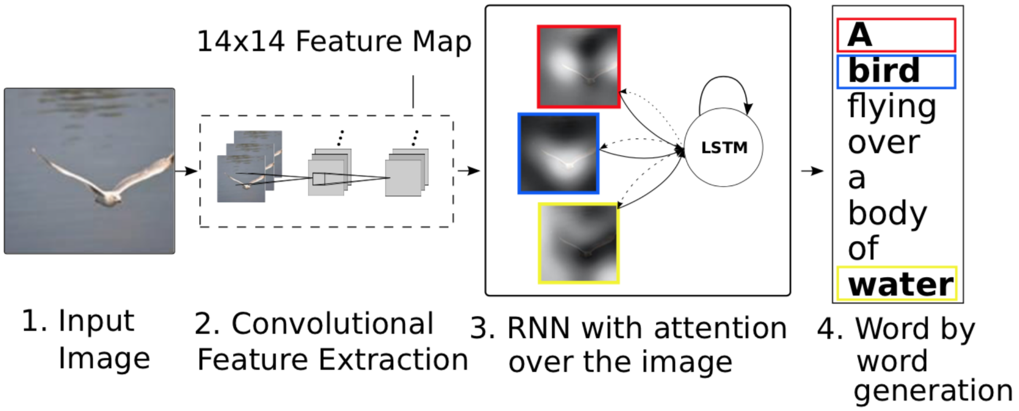

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Figure 6: CNN encoder - LSTM attention decoder architecture. From an ”early” CNN layer several D-dimensional vectors are extracted, and correspond to regions in the image. LSTM generates one word per time-step, conditioned on previous hidden state, generated words and context vector for that word. The image areas responsible for generating a particular word (three colored boxes) are colored with a stronger shade of white. Adapted from [57].

Xu et al., [57] first introduced the sequence-to-sequence model with **spatial attention** for the image captioning task. They conditioned the LSTM decoder on different parts of the input image during each decoding step (1 word generated per step), thus producing a distribution over image regions for each word (**Figure 6**). Compared to [52], here lower-level features from CNN are used, as opposed to the penultimate CNN layer, in this way preserving correspondence to the 2-D image portions. This allows the decoder to selectively focus on certain parts of an image by selecting a subset of all the feature vectors. LSTM generates one word at every time-step conditioned on a **context vector** (which captures the dynamic representation of the relevant part of the image input at a current step), the previous hidden state and the previously generated words. Thus we can learn which locations to focus on for producing the next word. There are two attention mechanisms compared: stochastic *Hard* and deterministic *Soft* attention.

Such model forces visual attention to be active for every generated word, except for short conjunctions like ”and” or ”but”, and other words that may seem visual can often be predicted reliably just from the language model e.g., ”sign” after ”behind a red stop” or ”phone” following ”talking on a cell”. [33] propose a novel adaptive attention model with a visual sentinel. At each time step, the model decides to which image regions to attend or to the visual sentinel in order to extract meaningful information for sequential word generation.

3.7 Paragraph-level Image Captioning

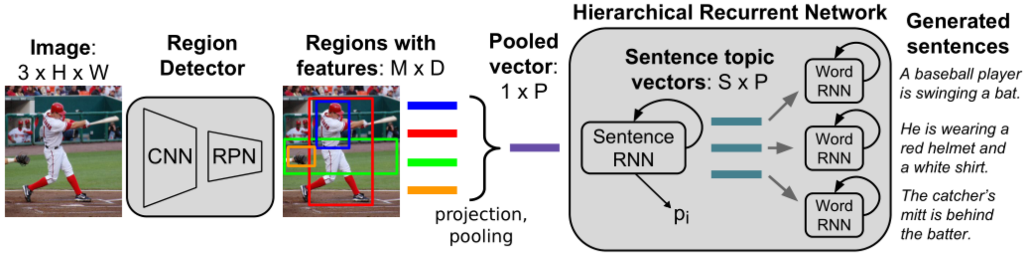


Image: 3 x H x W
Region Detector
CNN
RPN
Regions with features: M x D
Pooled vector: 1 x P
projection, pooling
Hierarchical Recurrent Network
Sentence topic vectors: S x P
Sentence RNN
p_i
Word RNN
Word RNN
Word RNN
Generated sentences
A baseball player is swinging a bat.
He is wearing a red helmet and a white shirt.
The catcher's mitt is behind the batter.

Figure 7: Krause 2017 implementation uses a region detector, comprised of CNN and region proposal network, to detect regions of interest and encode them one-by-one to $d = 4096$ vectors. These vectors are max-pooled into a single vector. Decoder is a hierarchical sentence RNN, which for each sentence/topic generates words with word RNN. Adapted from [27]

.

Above image captioning models have a key limitation – describing images with a single high-level sentence. **Dense captioning** model [23] solves this problem by feeding convolutional image features through a so-called localization layer which proposes a variable number of regions of interest (with Faster R-CNN). Next, each region of interest gets described with an LSTM-generated caption. However, each caption is independent from one another and the generated sentences are incoherent.

This problem is addressed by Krause et al (2016) [27] where the input image is decomposed by detecting objects and other regions of interest, then aggregate features across these regions to produce a pooled representation richly expressing the image semantics (**Figure 7**).This feature vector is taken as input by a hierarchical recurrent neural network composed of two levels: a sentence RNN and a word RNN. The sentence RNN receives the image features, decides how many sentences to generate in the resulting paragraph, and produces an input topic vector for each sentence. Given this topic vector, the word RNN generates the words of a single sentence.

3.8 Methods for constructing image-text joint embedding space

Let X and Y denote the collections of training images and sentences, each encoded according to their own feature vector representation. We want to map the image and sentence vectors (which may have different dimensions initially) to a joint space of common dimension. If the embeddings are L2 normalized then we can use inner product over embedding space to measure similarity/distance between two vectors in such space by Euclidean distance [53]. Benefits of joint embedding: for document retrieval tasks with the learned representations, only a limited amount of supervision is needed to yield results comparable to those of fully-supervised methods (Hsu 2018). Task evaluation is done by image-sentence retrieval.

3.9 Visual Question Answering

Despite the above-mentioned benchmark-crushing results, there is a strong evidence that the image captions alone do not capture an informative image representation. For example, [2] demonstrated that human subjects asked to answer a question about an image only after seeing image's captions were significantly less accurate compared to answering the same questions but this time while observing the actual image. Authors, in their seminal work, proposed instead the VQA - a new multimodal learning task. Microsoft's researchers launched the first **VQA Challenge**, which still runs yearly since 2015. Most importantly, they created a new MSCOCO-based [31] VQA dataset (200k images) supplemented with synthetic clipart, featuring 600k questions and 8mln answers. Task performed best with an LSTM and CNN architecture supplemented with captions. Dataset was recently balanced [13] to include complementary images such that every question in our balanced dataset is associated with not just a single image, but rather a pair of similar images that result in two different answers to the question.

Since then, various attention tricks were borrowed from neural machine translation field to help with the problem of ”where to look” in the training images. [58] proposed a stacked attention model which queries the image for multiple times to infer the answer progressively. Meanwhile [34] exploit a hierarchical question-image co-attention strategy to attend to both related regions in the image and crucial words in the question. Attention mechanism can find the question-related regions in the image, which accounts for the answer to some extent. But the attended regions still don't explicitly exhibit what the system learns from the image and it is also not explained why these regions should be attended to.

The so-called *Bottom-Up Top-Down* model is the winner of 2017 VQA challenge [1; 49]. *Bottom-up* attention gives bounded region boxes around the most salient image objects (obtained through a Faster R-CNN framework), each region is represented by a pooled convolutional feature vector. Such approach is supported by the recent success of regional proposal based (**R-CNN**) object detection algorithms [11; 12; 43]. The method uses a fixed threshold on object detection, and the number of features K is therefore adaptive to the contents of the image. The question text is then used to compute the task-specific *top-down* attention for each object in the image (with ResNet-101). It is worth mentioning, that bottom-up attention model is pretrained by initializing Faster R-CNN with ResNet-101 pretrained on ImageNet classification task. Multi-modal fusion of features is achieved by an entry-wise product, followed by a multi-label classifier with a sigmoid activation function to predict the candidate answer scores.

Next year's VQA winner *Pythia* [20] improved model's top-down attention by combining visual and textual feature vectors via element-wise multiplication, instead of concatenating them. Similarly, previous batch normalization's shortcomings due to mini-batch dependence in RNN architectures were ameliorated by normalizing the weight within a layer [44].

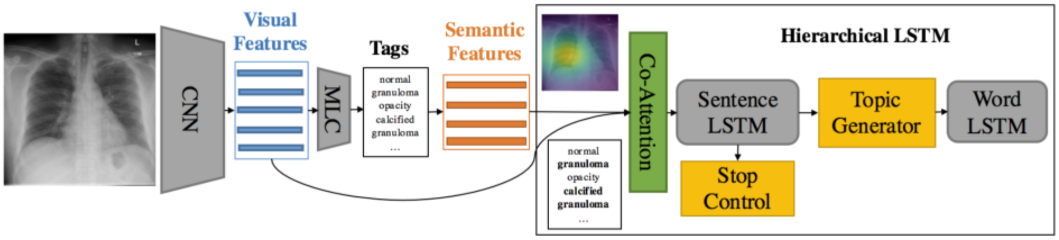


Figure 8: Hierarchical LSTM Co-Attention architecture as used by Jing et al (2018). Decoding begins with VGG-19 CNN visual features for image patches, fed directly into a multi-label classifier to predict disease tags, represented as word2vec 512 dim embedding (semantic feature) vectors. Those are fed together with visual features to generate context vector which pays attention to both of these features. Decoder takes in the context vector, and produces sentences in a hierarchical manner: context vector is fed into a sentence LSTM, which unrolls for a few steps to produce a 512-dim topic vector at each step. Meanwhile, word LSTM produces words for each topic/sentence vector.

3.10 Medical Image Captioning and Report Generation

Within the last 5 years, tens of medical image datasets have been made publicly available. Leveraging medical images and accompanying free text reports to improve disease state representations is an emerging field.

Up until recently, most of text trained with X-ray images was structured or semi-structured (templates, tags). For example, [45] proposed an convolutional encoder and recurrent network decoder framework that jointly trained from chest X-ray and doctor's report to predict simple tags relating to disease categories, abnormality locations and severities.

[21] improved X-ray image annotation and report text generation baselines in a multi-task learning framework, which includes a co-attention and hierarchical LSTMs. Model generates paragraph captions using a hierarchical LSTM, but unlike Krause et al. (2017), uses a co-attention network to generate topics.

[55] proposed a text-image embedding network to automatically generate X-ray reports (based on ChestX-ray dataset [54]) in an end-to-end trainable CNN-RNN architecture. Meaningful report words and image regions were

highlighted via multi-level attention models.

The recently released **MIMIC-CXR** [22] was used establish benchmarks in both supervised and unsupervised text-image embeddings [17], and more recently, a domain-aware automatic chest X-ray radiology report generation system [32] which first predicts what medical topics will be discussed in the report, then conditionally generates sentences corresponding to these topics. At the core, the system uses hierarchical convolutional-RNN, and is trained by a reinforcement learning model with the Clinically Coherent Reward policy, considering both readability and clinical accuracy, as assessed by the proposed Clinically Coherent Reward.

4 Datasets

4.1 Proof of Concept Datasets

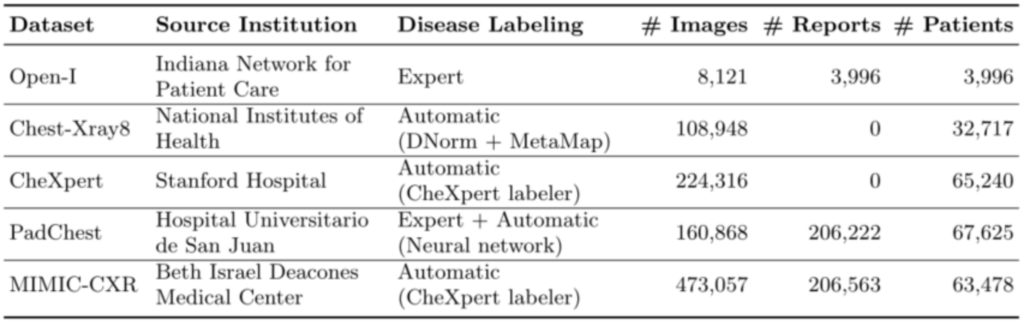
We envision building and training the first iterations of multimodal models on a series of well-established image-text datasets. **Microsoft COCO** [31] consists of 318k images with 2.5mln labeled instances of 91 object categories. Additionally, each object instance is segmented, and every image has five human-written single-sentence captions. Human subjects often disagree on the "correct" answer, so 10 people had to answer them. Machine generated answer accuracy metric is: $min(1, \#$ humans that provided that answer$)/3)$, so 100 % accuracy is reached if at least 3 humans provided that exact answer.

The first specialized dataset for **Visual uestion answering** task, VQA [2; 13], is derived from MSCOCO, VQA 2.0 contains 200k images, 1.1mln questions and 11mln answers. **GQA** is a recent VQA dataset [18] that focuses on real-world compositional reasoning. It contains 113k images and 2M questions which are answered from scene graphs, serving as a form of structured semantic representation, thus describing objects' attributes and physical inter-relations.them comes with an underlying structured representation of their semantics

Taking a further step towards "strong AI" is **OK-VQA** dataset [35], providing a benchmark for knowledge-based VQA, where the answer is not explicitly present in the image. For instance, upon asking which political entity is depicted in a blue flag with 12 yellow stars, the model would answer "the European Union".

4.2 Medical Image Datasets

It is a challenge to find text labeled medical images, for example the recent headline breaking binary classifiers such as **CheXNet** were done on images labeled for one particular disease, and only in binary fashion (lesion present/absent). See **Figure 9** for full list of datasets and their limitations.



Dataset	Source Institution	Disease Labeling	# Images	# Reports	# Patients
Open-I	Indiana Network for Patient Care	Expert	8,121	3,996	3,996
Chest-Xray8	National Institutes of Health	Automatic (DNorm + MetaMap)	108,948	0	32,717
CheXpert	Stanford Hospital	Automatic (CheXpert labeler)	224,316	0	65,240
PadChest	Hospital Universitario de San Juan	Expert + Automatic (Neural network)	160,868	206,222	67,625
MIMIC-CXR	Beth Israel Deacones Medical Center	Automatic (CheXpert labeler)	473,057	206,563	63,478

Figure 9: Available chest X-ray datasets and their characteristics, taken from [32]. Only Open-I and MIMIC-CXR datasets are useful for us, since they have real doctor reports.

There are only two datasets that contain radiologist's text (see Figure 9. For our proof-of-concept work, we chose the smaller, more established **IU X-ray** [19], which contains 3,826 radiology reports associated with 7,430 X-rays. We used 80:10:10 training:validation:test split.

In terms of further pre-processing, punctuation and numbers removed with NLTK library. Corpus was 1,800 unique words, and we limited our vocabulary to 1,286 words after filtering for words that occurred at least 3 times in the training set, resulting in 1,268 words. punctuation and numbers, which resulted to 1820 unique words. Average paragraph length is 30 words, in 5 sentences.

Report consists of the **impressions** and **findings** sections, which for our purposes are concatenated together as a long paragraph, since impression can be viewed as a conclusion or topic sentence of the report.

Findings section posed as the most important component, ought to cover contents of various aspects such as heart size, lung opacity, bone structure; any abnormality appearing at lungs, aortic and hilum; and potential diseases such as effusion, pneumothorax and consolidation. And, in terms of content ordering, the narrative of findings section usually follows a presumptive order, e.g. heart size, mediastinum contour followed by lung opacity, remarkable abnormalities followed by mild or potential abnormalities.

Eventually, we aim to expand our model to **MIMIC-CXR** [22] - a new publicly available dataset of chest radiographs with structured labels, containing 370k images from 220k radiographic studies, each marked by the CheXpert labeler with one of 14 categories. The image-accompanying report contains "findings" and "impression" sections, which describe, respectively, the image patterns as seen by the clinician, and clinical interpretation of those patterns. Medical Text Indexer (MTI) [37] is used to extract tags from the "findings" and "impression" sections of the reports, since they're not present by default. We will use MTI labeler and, since it doesn't handle negations, MetaMap [3] to detect tags with negation and discarded them.

4.3 Computing Resources

JupyterHub on Novartis HPC. PythonDS environment module has pytorch 1.2.0, torchvision 0.4.0 and keras-gpu 2.2.4

5 Project Goals

5.1 Test current Image Captioning Models

In order to get a feel for the data, and the computational complexity for image-text embedding, we will attempt to replicate a study with a generic dataset (MSCOCO) first, and later, apply it to medical image dataset. MSCOCO image captioning is performed by the following three models:

- Encoder-decoder model from "Show and Tell" paper [52], see section 3.5. Model implementation adapted from github user sgrvinod's repository.

- Encoder-decoder model with attention from "Show, Attend and Tell" by [57], see section 3.6. Implementation adapted from github user sgrvinod's repository

- Bottom-Up and Top-Down Attention system from [1], see section 3.9. Implementation adapted from github user poojahira's repository

5.2 Implement Chest X-ray Captioning Baselines

In the second step, we will adapt the Encoder-decoder model from "Show and Tell" paper [52] and encoder-decoder model with attention from "Show, Attend and Tell" by [57] to UI dataset (see section 9) from the famous MSCOCO dataset [31].

We will evaluate Pairwise and Triplet Ranking Loss functions. Both of these objectives focus on updating the weights only when the distance $s()$ between generated sentences and true images (and vice versa) is less than some margin m.

5.3 Implement Hierarchical LSTM Co-Attention

We followed the closed source paper by Jing et al (2017) [21], which achieved state of the art results on IU dataset at the time of publications (see Figure 8). The primary objective was to write an open source implementation of this study, as the authors have kindly refused to release the source code, and attempt to replicate their results. The Hierarchical LSTM Co-Attention model is based on Krause et al. [27], except that a number of assumptions had to be made to adapt the model from MSCOCO to IU X-ray dataset, most importantly that a co-attention network is used to generate topics, as opposed to the attention-free region proposal network approach.

For further details see the full implementation at `./utils/models.py`

6 Model of Interest: Hierarchical LSTM Co-Attention (HLCA)

We will describe in detail the HLCA model that we attempted to implement in PyTorch and apply to IU and MIMIC-CXR datasets.

6.1 HLCA model at a glance

A sample X-ray diagnostic report is shown in **Figure 1**, and consists of long paragraphs and list of medical term/disease tags. The decoding process begins when VGG-19 CNN is used to learn visual features for image patches, and feed them directly into a multi-label classifier to predicts medical disease MTI tags (600 possible tags), represented as word2vec 512 dim embedding vectors (see Figure 8). These word embedding vectors are the **semantic features** for the image, and are fed together with the "raw" VGG-199 CNN **visual features** to generate context vector which pays attention to both visual and semantic features.

Decoder takes in the context vector as its input, and outputs generated sentences in a hierarchical manner: context vector is inputted into a sentence LSTM, which unrolls for a few steps and produces a 512-dim topic vector at each step. A topic vector represents the semantics of a sentence to be generated. Meanwhile, word LSTM generates a sequence of words for each topic/sentence vector.

6.2 MTI Tag Prediction with Multi-label Classifier

For an image I, a total of N D-dimensional features are extracted $\{\mathbf{v}_n\}_{n=1}^N \in R^D$ from the last convolutional layer of VGG-CNN. Next these features are fed into MLC network to generate a distribution of over all $L = 600$ tags:

$$\mathbf{p}_{\mathbf{l},pred}(\mathbf{l}_i = 1|\{\mathbf{v}_n\}_{n=1}^N) \propto \exp(MLC_i(\{\mathbf{v}_n\}_{n=1}^N)) \tag{1}$$

where $\mathbf{l} \in R^L$ is a tag vector, identity $\mathbf{l}_i = 1, 0$ denotes presence of the i-th tag respectively, and MLC$_i$ means the i-th output of the MLC network. Then, M most probable tags ($M = 5$ in our experiments) are embedded with $E = 512$-dimensional word2vec as the semantic features: $\{\mathbf{a}_m\}_{m=1}^M \in R^E$.

6.3 Attention for visual and semantic features

With every "unrollling" of sentence LSTM at some time step s, the context vector $\mathbf{ctx}^{(s)} \in R^C$ is generated by a co-attention network $f_{co_{att}}(\{\mathbf{v}_n\}_{n=1}^N, \{\mathbf{a}_m\}_{m=1}^M, \mathbf{h}_{sent}^{(s-1)})$, where $\mathbf{h}_{sent}^{(s-1)} \in R^H$ is the sentence LSTM hidden state at time step $s-1$. The co-attention network $f_{co_{att}}$ uses a single layer fully connected net to compute the soft visual attentions and soft semantic attentions over input image features and tags:

$$\alpha_{\mathbf{v},n} \propto \exp(\mathbf{W}_{\mathbf{v}_{att}} \tanh(\mathbf{W}_{\mathbf{v}}\mathbf{v}_n + \mathbf{W}_{\mathbf{v},\mathbf{h}}\mathbf{h}_{sent}^{(s-1)}))$$

$$\alpha_{\mathbf{a},m} \propto \exp(\mathbf{W}_{\mathbf{a}_{att}} \tanh(\mathbf{W}_{\mathbf{a}}\mathbf{a}_m + \mathbf{W}_{\mathbf{a},\mathbf{h}}\mathbf{h}_{sent}^{(s-1)}),$$

where $\mathbf{W_v}$, $\mathbf{W_{v,h}}$, $\mathbf{W}_{\mathbf{v}_{att}}$ and $\mathbf{W_a}$, $\mathbf{W_{a,h}}$, $\mathbf{W}_{\mathbf{a}_{att}}$ are learned parameters for visual and semantic attention neural net, respectively.

From $\alpha_{\mathbf{v},n}$ and $\alpha_{\mathbf{a},m}$, we can calculate the visual and semantic context vectors:

$$\mathbf{v}_{att}^{(s)} = \sum_{n=1}^{N} \alpha_{\mathbf{v},n}\mathbf{v}_n, \quad \mathbf{a}_{att}^{(s)} = \sum_{m=1}^{M} \alpha_{\mathbf{a},m}\mathbf{a}_m.$$

Finally, visual and semantic context vectors are combined into a joint context vector by:

$$\mathbf{ctx}^{(s)} = \mathbf{W}_{fc}[\mathbf{v}_{att}^{(s)}; \mathbf{a}_{att}^{(s)}], \tag{2}$$

where the two vectors are first concatenated, then passed through a fully-connected layer. A more straightforward alternative to using another layer would be a simple concatenation operation.

6.4 Sentence LSTM

This is a single-layer LSTM that takes the joint context vector $\mathbf{ctx} \in R^C$ as its input, and generates topic vector $\mathbf{t} \in R^K$ for word LSTM through topic generator:

$$\mathbf{t}^{(s)} = \tanh(\mathbf{W}_{\mathbf{t},\mathbf{h}_{sent}}\mathbf{h}_{sent}^{(s)} + \mathbf{W}_{\mathbf{t},\mathbf{ctx}}\mathbf{ctx}^{(s)}) \tag{3}$$

where $\mathbf{W}_{\mathbf{t},\mathbf{h}_{sent}}$ and $\mathbf{W}_{\mathbf{t},\mathbf{ctx}}$ are weight parameters.

In order to stop generating sentences, the RNN takes in the previous and current hidden state $\mathbf{h}_{sent}^{(s-1)}$, $\mathbf{h}_{sent}^{(s)}$ as input and produces a distribution over $\{STOP{=}1,\ CONTINUE{=}0\}$:

$$p(STOP|\mathbf{h}_{sent}^{(s-1)}, \mathbf{h}_{sent}^{(s)}) \propto \exp\{\mathbf{W}_{stop}\tanh(\mathbf{W}_{stop,s-1}\mathbf{h}_{sent}^{(s-1)} + \mathbf{W}_{stop,s}\mathbf{h}_{sent}^{(s)})\} \tag{4}$$

where $\mathbf{W}_{stop}$, $\mathbf{W}_{stop,s-1}$ and $\mathbf{W}_{stop,s}$ are parameter matrices. If $p(STOP|\mathbf{h}_{sent}^{(s-1)}, \mathbf{h}_{sent}^{(s)})$ is greater than 0.5, then the sentence LSTM will stop producing new topic vectors and the word LSTM will also stop producing words.

6.5 Word LSTM

Word LSTM is identical to that of Krause et al (2017) [27], and takes in the topic vector $\mathbf{t}$ produced by the sentence LSTM and the special $START$ signal as inputs. The hidden state $\mathbf{h}_{word} \in R^H$ of the word LSTM is directly used to predict the distribution over words:

$$p(word|\mathbf{h}_{word}) \propto \exp(\mathbf{W}_{out}\mathbf{h}_{word}) \tag{5}$$

6.6 Loss Function for Sentence and Word Generation

Loss function used by [21] Given a training image I, true tag vector $\mathbf{l}$ and paragraph $\mathbf{w}$ (with s sentences), our model first performs multi-label classification on I and produces a distribution $\mathbf{p}_{\mathbf{l},pred}$ over all tags. Note that $\mathbf{l}$ is a binary vector which encodes the presence and absence of tags. Ground-truth tag distribution by normalizing $\mathbf{l}$: $\mathbf{p}_{\mathbf{l}} = \mathbf{l}/||\mathbf{l}||_1$. The training loss of this step is a cross-entropy loss ℓ_{tag} between $\mathbf{p}_{\mathbf{l}}$ and $\mathbf{p}_{\mathbf{l},pred}$.

Next, the sentence LSTM is unrolled for S steps to produce topic vectors and also distributions over $\{STOP, CONTINUE\}$: p^s_{stop}. Finally, the S topic vectors are fed into the word LSTM to generate words $\mathbf{w}_{s,t}$. The training loss of caption generation is the combination of two cross-entropy losses: ℓ_{sent} over stop distributions p^s_{stop} and ℓ_{word} over word distributions $p_{s,t}$. Combining the pieces together, we obtain the overall training loss:

$$\ell(I,\mathbf{l},\mathbf{w}) = \lambda_{tag}\ell_{tag} + \lambda_{sent}\sum_{s=1}^{S}\ell_{sent}(p^s_{stop}, I\{s = S\}) + \lambda_{word}\sum_{s=1}^{S}\sum_{t=1}^{T_s}\ell_{word}(p_{s,t}, w_{s,t}) \tag{6}$$

Trained model, the joint image and text embedding, can be found at `/report_v4_models/<model_name>/train_best_loss.pth.tar`

6.7 Metrics

BLEU [38] - used for translation; precision and recall are approximated by modified n-gram precision (fraction of n-grams in the candidate text which are present in any of the reference texts) and best match length.

ROUGE [30] is based only on recall, and is mostly used for summary evaluation. ROUGE-n: This is based on n-grams. For example, ROUGE-1 counts recall based on matching unigrams, and so on. For any n, we count the total number of n-grams across all the reference summaries, and find out how many of them are present in the candidate summary. This fraction is the required metric value.

METEOR [8] is another metric for machine translation evaluation, and it claims to have better correlation with human judgement. Similar to BLEU, but it reduced BLEU's dependency on average length value across the corpus, by replacing precision and recall calculations with weighted F-score.

CIDER [51] is a recent automatic consensus metric of image description quality, measuring the similarity of a generated sentence against a set of ground truth sentences written by humans. It shows high agreement with consensus as assessed by humans. Using sentence similarity, the notions of grammaticality, saliency, precision and recall are inherently captured.

7 Results

7.1 Performance on NLP Metrics

In this section we compare the results from the original to our implementation of Hierarchical LSTM Co-Attenion model by [21] (see **Table 1**). Training took for 350 epochs, each epoch lasting around 15 minutes, with NVIDIA Tesla K80.

Unfortunately, we could not replicate the performance as seen in [21] Table 1, as the Train, Val and Test scores are significantly lower than the original Test scores. Multi-label classifier and overall model training performance is shown in **Figure 10** and **Figure 11**. We can see that the model failed at building a multi-label tag classifier. However, the project time ran up before we could fix the bug. Training implementation is located at `./train.py`

On the upside, the performance of our attempt at Enc-Dec and Enc-Dec+Att models was only marginally lower compared to the baselines reported in [21].

Also, we independently trained Encoder-Decoder and Encoder-Decoder with Attention architectures with IU X-ray dataset (see * in Table 1).

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDER
Train	0.286	0.275	0.215	0.176	0.187	0.369	0.403
Val	0.240	0.182	0.118	0.077	0.143	0.256	0.172
Test	0.213	0.190	0.123	0.081	0.158	0.224	0.120
Jing 2018 [21]	0.517	0.386	0.306	0.247	0.217	0.447	0.327
Enc-Dec* [52]	0.298	0.207	0.111	0.088	0.151	0.249	0.114
Enc-Dec + Att* [57]	0.356	0.243	0.165	0.120	0.159	0.323	0.301

Table 1: Results for paragraph generation on the IU X-Ray dataset. We compare our implementation (Train, Val, Test sets) to the original hierarchical LSTM study [21] test results, shown here as ”Jing 2018”. Also, we independently trained Encoder-Decoder and Encoder-Decoder with Attention architectures with IU X-ray dataset (two last rows, *). BLEU-n denotes the BLEU score n-grams. We could not replicate the performance as seen in [21].

7.2 Sentence Generation by Hierarchical LSTM Co-Attention

Sample paragraphs generated by of training are shown in **Figure 12** and **Figure 13** (see notebook `./explore_generated_text.ipynb` for more generated captions). The first image is negative, while the second has some abnormalities and is also a lateral x-ray.

Observe, that at this training stage the model overfits to the most popular (negative) sentences: ”no acute cardiopulmonary abnormality”, ”the heart is normal in size”, ”the lungs are clear”, ”no focal consolidation pneumothorax or pleural effusion identified”.

The model is clearly overfitting, as corroborated by the low generated sentence diversity **Figure 15**. Test set consisted of 600 images, and almost at least half of the paragraphs were unique, they consisted of only 14 different sentences. These sentences not surprisingly are also the most frequently seen sentences in the training set.

We should invest more time fixing the tag classifier and in grid/random parameter search, especially for the learning rate and regularization.

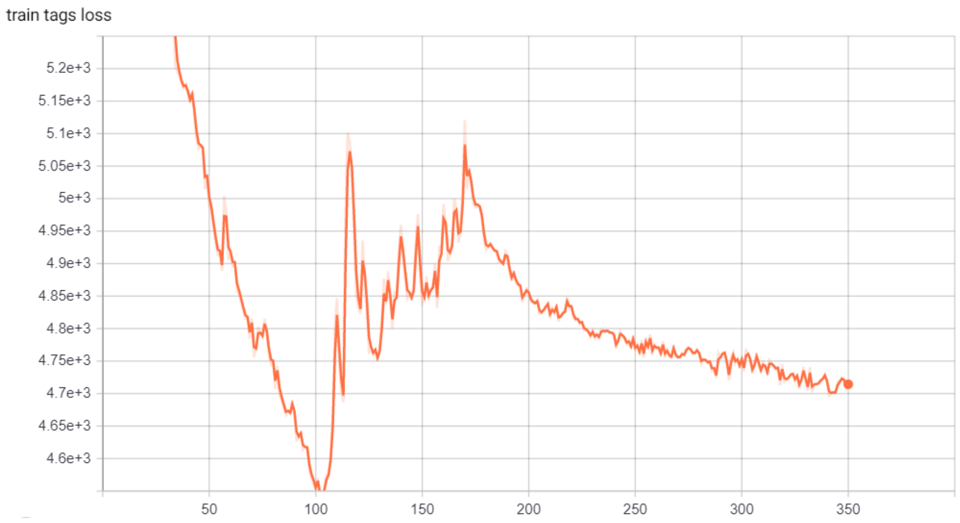


Figure 10: Tag classifier training loss for Hierarchical LSTM Co-Attenion model. Multi-label classifier training failed. For details of MLC implementation see section 6.2

.

7.3 Co-Attention Implementation

A joint image and semantic feature attention mechanism was implement as described in section 6.3. Since the model implementation has a bug, we cannot verify attention results, however we give an example of the visual attention responsible for diagnosing ”Pneumothorax” (shown in yellow) in **6.3**.

Code for attention mechanism is found at: `./caption_and_attend.py`.

7.4 Testing current Image captioning models

Successfully implemented and tested existing captioning models. See directory `./code/attend_tell` for [52] and [57] implementations. We had no problems replicating their results, although they were lower because we did not allocate the necessary compute time (see the last 2 rows of Table 1).

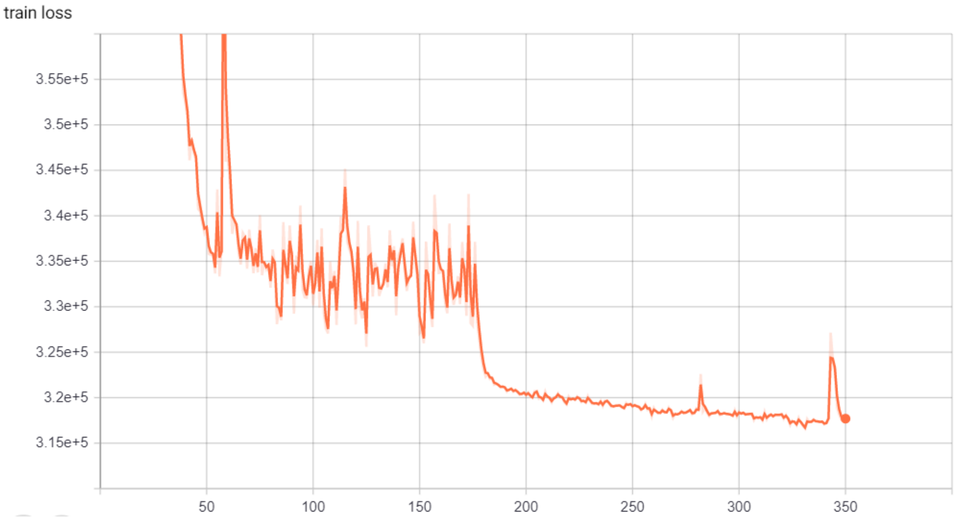


Figure 11: Training loss for Hierarchical LSTM Co-Attenion model.

8 Discussion and Future Directions

8.1 Report Generation from Medical Images

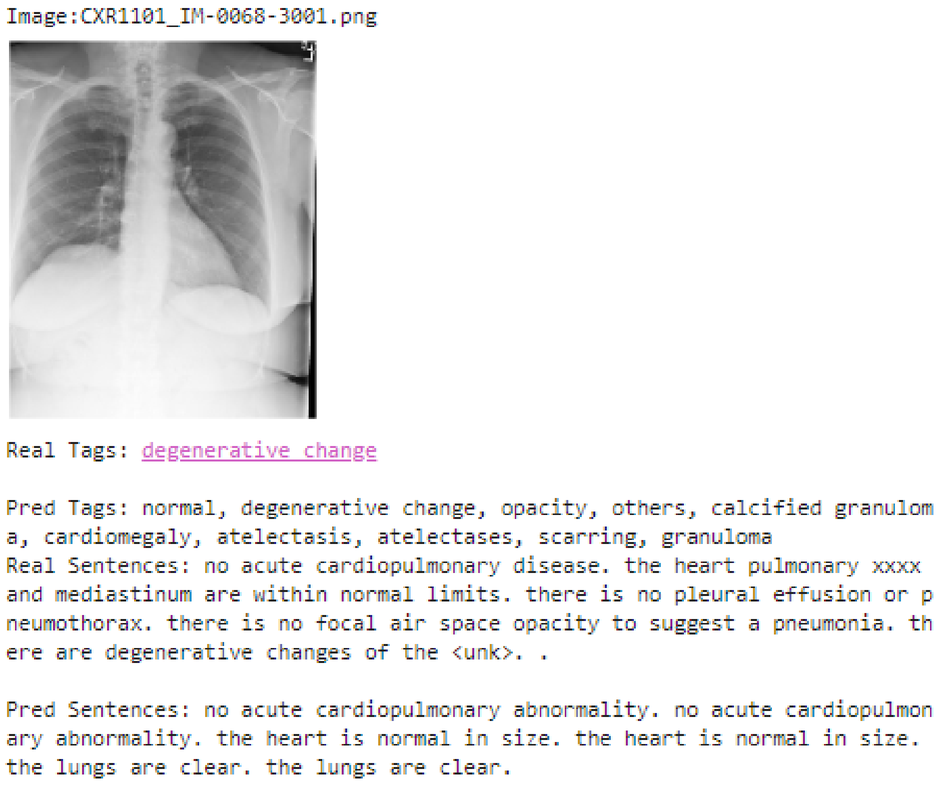
We were able to successfully replicate the encoder-decoder image captioning models [52][57] for MSCOCO dataset, which contains images from everyday life, as well as generating an ”impressions” caption for IU X-ray dataset [19].

In the absence of a reference implementation, we set out to verify the state of the art chest x-ray report generation results from Jing at al (2018) study [21]. To this end, we built a model and adapted it to IU dataset. Unfortunately, the model was not training correctly, as explained is section 7.2, mostly due to a bug in multi-laber classifier implementation.

8.2 MIMIC-CXR

Since the Hierarchical LSTM Co-Attention model could not be replicated successfully with IU dataset, we have not tried adapting it for an order of magnitude larger MIMIC-CXR dataset (see [22], Table 1).

Since this dataset lacks MTI tags, the Medical Text Indexer (MTI) [37] was used to extract tags from the ”findings” and ”impression” sections of the reports, since they’re not present by default. Next, we used the MTI



```
Image:CXR1101_IM-0068-3001.png

Real Tags: degenerative change

Pred Tags: normal, degenerative change, opacity, others, calcified granulom
a, cardiomegaly, atelectasis, atelectases, scarring, granuloma
Real Sentences: no acute cardiopulmonary disease. the heart pulmonary xxxx
and mediastinum are within normal limits. there is no pleural effusion or p
neumothorax. there is no focal air space opacity to suggest a pneumonia. th
ere are degenerative changes of the <unk>. .

Pred Sentences: no acute cardiopulmonary abnormality. no acute cardiopulmon
ary abnormality. the heart is normal in size. the heart is normal in size.
the lungs are clear. the lungs are clear.
```

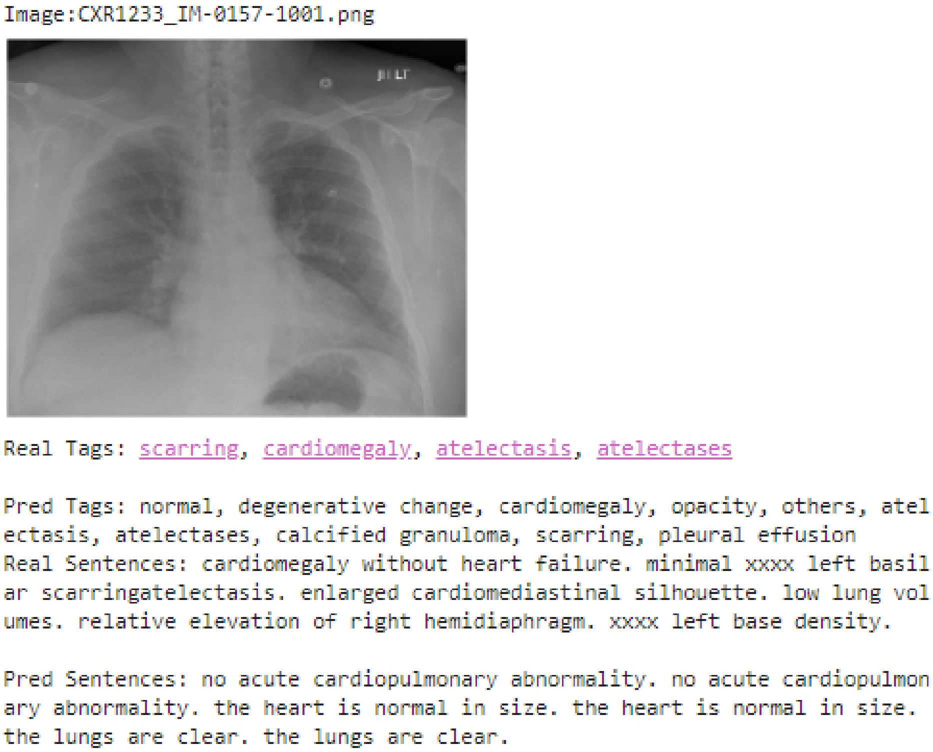
Figure 12: Sample caption 1

labeler and, since it doesn't handle negations, MetaMap [3] to detect tags with negation and discarded them.

8.3 Addressing lack of factual correctness in generated metrics

Neural summarization models are able to generate summaries which have high overlap with human references. However, existing models are not optimized and do not guarantee for factual correctness [60]. In the report generation field we have observed that even metrics like CIDER, which are developed for image captioning tasks, will not prioritize factual correctness. For example, that a ground truth chest disease observation (by a doctor) is:

pneumothorax is seen, bilateral pleural effusions continue. Now imagine two competing models produce these observations:



```
Image:CXR1233_IM-0157-1001.png
```

```
Real Tags: scarring, cardiomegaly, atelectasis, atelectases

Pred Tags: normal, degenerative change, cardiomegaly, opacity, others, atel
ectasis, atelectases, calcified granuloma, scarring, pleural effusion
Real Sentences: cardiomegaly without heart failure. minimal xxxx left basil
ar scarringatelectasis. enlarged cardiomediastinal silhouette. low lung vol
umes. relative elevation of right hemidiaphragm. xxxx left base density.

Pred Sentences: no acute cardiopulmonary abnormality. no acute cardiopulmon
ary abnormality. the heart is normal in size. the heart is normal in size.
the lungs are clear. the lungs are clear.
```

Figure 13: Sample caption 2

A) ***no*** *pneumothorax is observed, bilateral pleural effusions continue*,

B) *pneumothorax is observed on radiograph, bilateral pleural effusions continue to be seen*

Although B) is factually correct, it overlaps less with the ground truth observation when CIDER or ROUGE metrics are considered. Such observations could invalidate most of the recent clinical image report generation studies that relied on traditional NLP metrics as objectives to be optimized during the training. Therefore, the future models will have to be trained with a new "factual correctness" objective function or a reinforcement learning policy.

8.4 Embeddings as a NIBR Service Initiative

Joint multimodal embeddings have proved to increase the accuracy of many medical prediction tasks. Such trained embeddings from various sources could be made available to Novartis researchers and increase the predictive

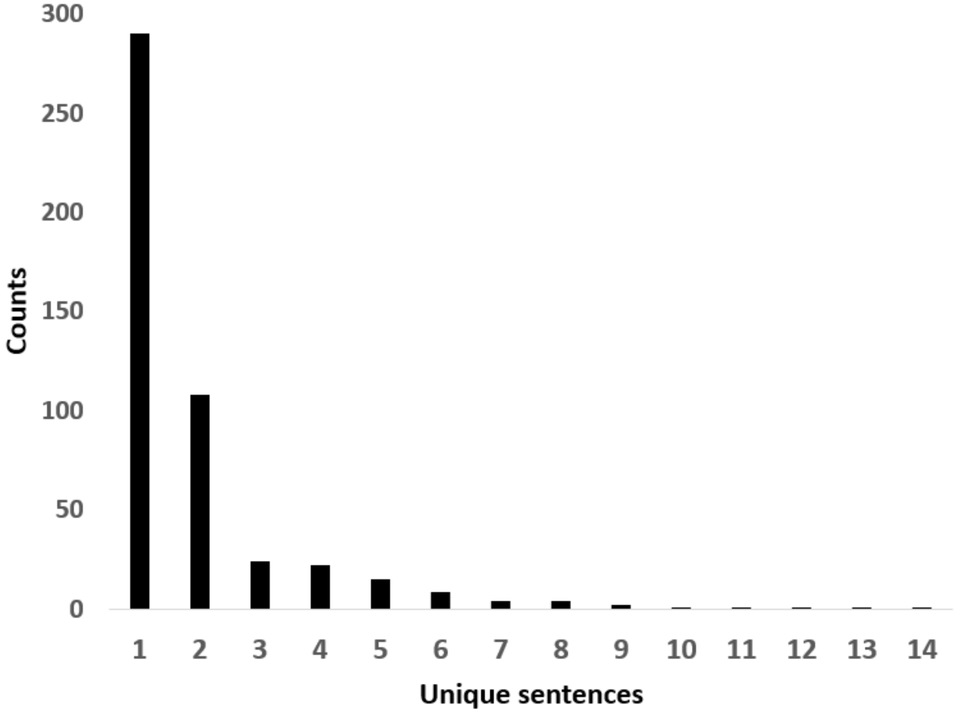


Figure 14: Sentence diversity

capabilities of their models. For example, text-annotated image datasets of cancer biopsies, retina, chest x-rays, tissue stains and tomography scans are abundant in academic and private databases. New initiatives such as Novartis's **Data42** are designed to integrate the patient data from multiple sources, therefore multimodal patient or disease level embeddings could produce useful patient and disease representations for downstream learning tasks.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

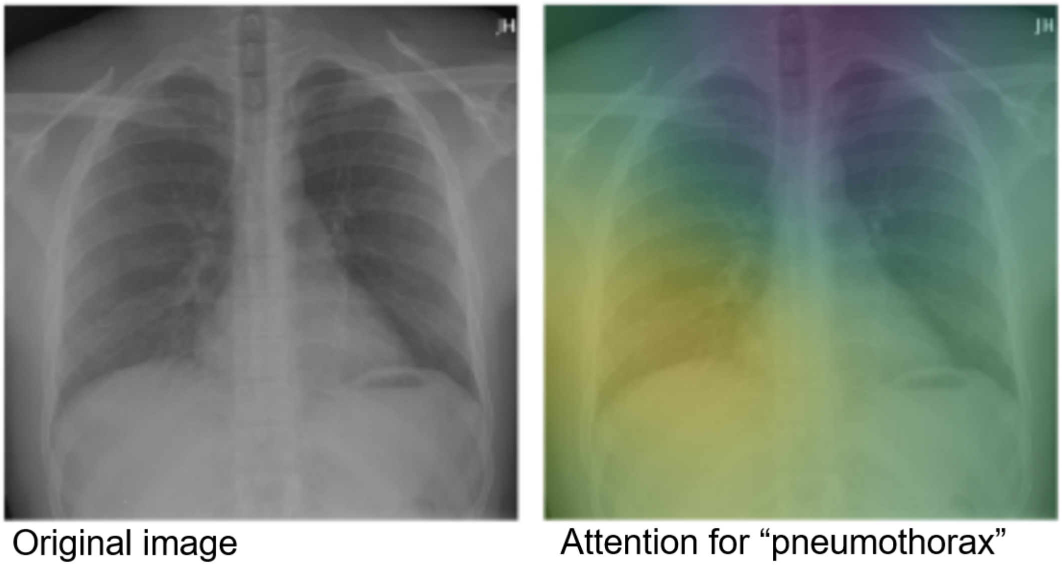


Figure 15: The visual attention responsible for diagnosing ”Pneumothorax” is shown in yellow

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[6] Andrew L Beam, Benjamin Kompa, Inbar Fried, Nathan P Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. Clinical concept embeddings

learned from massive sources of multimodal medical data. *arXiv preprint arXiv:1804.01486*, 2018.

[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[8] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[14] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[17] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615*, 2018.

[18] Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 2019.

[19] Open i. Indiana university - chest x-rays (png images).

[20] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[21] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

[22] Alistair EW Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[23] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

[24] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[25] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *International Conference on Machine Learning*, pages 595–603, 2014.

[26] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[27] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–325, 2017.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[32] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation, 2019.

[33] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.

[34] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[35] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3195–3204, 2019.

[36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[37] James Mork, Alan Aronson, and Dina Demner-Fushman. 12 years on–is the nlm medical text indexer still useful and relevant? *Journal of biomedical semantics*, 8(1):8, 2017.

[38] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[39] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[40] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[41] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[42] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM, 2010.

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[44] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks.

In *Advances in Neural Information Processing Systems*, pages 901–909, 2016.

[45] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506, 2016.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[47] I Sutskever, O Vinyals, and QV Le. Sequence to sequence learning with neural networks. *Advances in NIPS*, 2014.

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[49] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4223–4232, 2018.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[51] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[52] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[53] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[54] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[55] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 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*, pages 9049–9058, 2018.

[56] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[57] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[58] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[59] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[60] Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D. Manning, and Curtis P. Langlotz. Optimizing the factual correctness of a summary: A study of summarizing radiology reports, 2019.