

# Neural Captioning for the ImageCLEF 2017 Medical Image Challenges

David Lyndon, Ashnil Kumar, and Jinman Kim

School of Information Technologies, University of Sydney

**Abstract.** Manual image annotation is a major bottleneck in the processing of medical images and the accuracy of these reports varies depending on the clinician’s expertise. Automating some or all of the processes would have enormous impact in terms of efficiency, cost and accuracy. Previous approaches to automatically generating captions from images have relied on hand-crafted pipelines of feature extraction and techniques such as templating and nearest neighbour sentence retrieval to assemble likely sentences. Recent deep learning-based approaches to general image captioning use fully differentiable models to learn how to generate captions directly from images. In this paper, we address the challenge of end-to-end medical image captioning by pairing an image-encoding convolutional neural network (CNN) with a language-generating recurrent neural network (RNN). Our method is an adaptation of the NICv2 model that has shown state-of-the-art results in general image captioning. Using only data provided in the training dataset, we were able to attain a BLEU score of 0.0982 on the ImageCLEF 2017 Caption Prediction Challenge and an average F1 score of 0.0958 on the Concept Detection Challenge.

**Keywords:** Deep Learning, Image Captioning, LSTM, CNN, RNN

## 1 Introduction

Generating a textual summary of the insights gleaned from a medical image is a routine, yet nonetheless time-consuming task requiring much human effort on the part of highly trained clinicians. Prior efforts to automate this task relied on hand-crafted pipelines, employing manually designed feature extraction and techniques such as templating and sentence retrieval to assemble likely sentences [9, 13, 14]. Recent deep learning-based approaches to general image captioning, however, use fully differentiable models to learn how to generate captions directly from images. In general the advantages of such fully learnable models is that any part of the model can adapt in a manner most useful for the problem at hand, whereas a hand-designed system is constrained by the assumptions made during feature extraction, concept detection, and sentence generation.

In this paper we describe the submission of the University of Sydney’s Biomedical Engineering & Technology (BMET) group to the caption prediction and

concept detection task of the ImageCLEF 2017 caption challenge [4, 7]. This submission employs a fully differentiable model, pairing an image-encoding CNN with a language-generating RNN to generate captions for images from a range of modalities.

## 2 Background

Image captioning, whereby the contents of an image are automatically described in natural language, is a challenging task in machine learning, requiring methods from both image and natural language processing. Many early approaches to this problem involved complex systems comprising of visual feature extractors and rule based methods for sentence generation. Li et al. [11] utilise image feature similarity measures to locate likely n-grams from a large corpus of image and text, then use a simple sentence template and local search to generate a caption. Yao et al. [22] extract image features such as SIFT and edges to match images to a concept database, then apply a graph-based ontology to these concepts to produce readable sentences. Ordonez et al. [12] use image features and a ranking-based approach to locate likely sentences in an extremely large database of images and text. Such methods require a great deal of hand-crafted optimisation and produce systems which are brittle and limited to specialised domains.

Recently, deep learning-based encoder-decoder frameworks for machine translation [16] have been adapted and applied to the problem of image captioning. By replacing the language-encoding Long Short-Term Memory (LSTM) [6] RNN with an image-encoding CNN, the model is able to learn to generate captions directly from images. The entire model is completely differentiable so errors are propagated to the different components proportional to their contribution to the error, allowing them to adapt appropriately. While there were several precursors that replaced various components of existing image to caption frameworks with trainable RNNs or CNNs, Vinyals et al. [19] proposed the first end-to-end neural network based approach to captioning with their “Show and Tell” (also called Neural Image Captioning (NIC)) model. An updated method, NICv2 [20], won the Microsoft Common Objects in Context (MSCOCO) challenge in 2015. Qualitative analysis has shown that neural captioning methods are preferred in comparison with conventional nearest-neighbour sentence lookup approaches [3].

There has been limited work in adapting such methods to the medical domain, despite the large volume of image and text data found in PACS. Schlegl et al. [13] present the first such work that leveraged text reports to improve classification accuracy of CNN applied to Optical Coherence Tomography (OCT) images. Mahmood et al. [17] present a method that uses hand-coded topic extraction, hand-coded image features and a SVM-based correlation system. Shin et al. [14] document efforts to mine an extremely large database of images and text extracted from the PACS of the National Institutes of Health Clinical Center (approximately 216 thousand images) using latent Dirichlet allocation (LDA) to extract topics from the raw text and then correlate these topics to image features. Kisilev et al. [9] proposes an SVM-based approach to highlight regions of

interest (ROIs) and generate template-based captions for the Digital Database for Screening Mammography (DDSM). This is extended using a multi-task loss CNN in a later work [8].

To the best of our knowledge, only one published work exists for applying neural image captioning to a medical dataset [15]. In this work the authors employ an architecture similar to Vinyals et al. [19] to generate an array of keywords for a radiological dataset.

### 3 Method

Unless otherwise specified, the same method was applied for both the caption prediction and concept detection tasks. The set of concepts assigned to an image in the concept detection task is considered to be a caption where each concept label is a word in the sentence. In both cases only the supplied training dataset was used to train the models.

#### 3.1 Preprocessing

In order to simplify the task, each image in the training set was preprocessed in accordance with the task’s evaluation preprocessing specifications. This involved converting the caption to lower case, removing all punctuation (some captions contained multiple sentences, however, after this step each caption became a single sentence), removing stopwords using the NLTK [1] English stopword list and finally applying stemming using NLTK’s Snowball stemmer. No preprocessing was applied to the ‘sentences’ for the concept detection task. After this preprocessing the count of each unique word in the training corpus was taken. Words that appeared less than 4 times were discarded and this resulted in a dictionary of 25237 distinct words. For the RNN framework described below, two reserved words indicating the start and end of sentences are added to the dictionary and used to prepend and append each sentence.

The images are first resized to 324x324px, and a 299x299px crop is then selected. During training this is a random crop, but during evaluation a central crop is used. We apply image augmentation during training to regularise the model [10]. This augmentation consists of distorting the image, first by randomly flipping it horizontally then randomly adjusting the brightness, saturation, hue and contrast. The random cropping and distortion are performed each time an image is passed into the model and means that it is extremely rare that exactly the same image is seen twice.

A validation set was provided by the organisers of the task and it entirely reserved for validation. No part of it was used for training, and specifically we did use it to build the dictionary of unique words.

## 3.2 Model

Our method extends Vinyals et. al’s NICv2 model [20]<sup>1</sup>. The NICv2 model consists of two different types of neural networks paired together to form an image-to-language, encoder-decoder pair. A CNN, specifically the InceptionV3 [18] architecture, is used as an image encoder. InceptionV3 is one of the most accurate architecture for general image classification according to the ImageNet [2] benchmark, but is significantly more computationally efficient than alternatives such as Residual Networks [5]. We utilised a RNN based on LSTM units as the language decoder as per the original paper, however, we doubled the number of units from 512 to 1024 as this showed improved results in our experiments.

An image is first preprocessed as described above and then fed to the input of the CNN. The logits of the CNN are passed into a single layer fully-connected neural network which functions as an image embedding layer. This image embedding then becomes the initial state of the LSTM network. As per [20] the embedding was passed only at the initial state and is not used subsequently. At each state subsequent to the initial state, then LSTM’s output is passed to a word embedding layer and then to a softmax layer. At each time step the output of the softmax is the probability of each word in the dictionary. For two of the caption prediction experiments (PRED2 & PRED4) we modified the baseline language model to use a 3-layer LSTM model with a single dropout layer on the output. Increasing the number of LSTM layers improves the ability of the language model to represent complex sentences and long term dependencies. Industrial neural machine translation models have been demonstrated to use decoder layers with up to 8 layers [21].

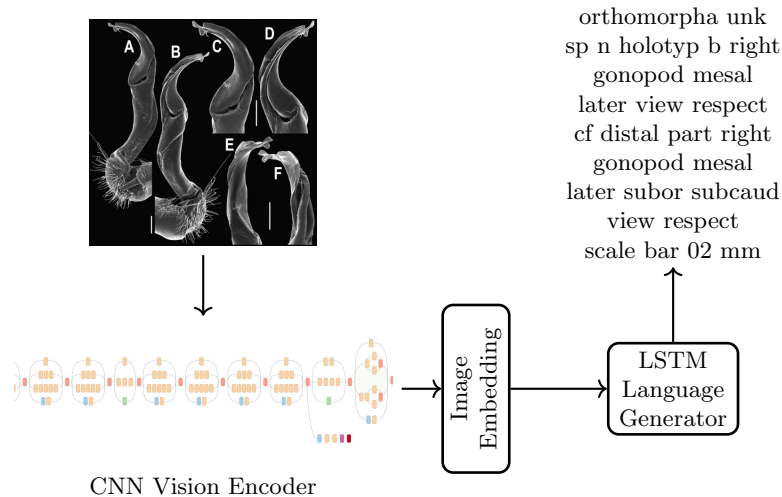
In all our models we used 1024 units for both the image and word embedding layers. The CNN is initialised using weights from a model trained on the ImageNet dataset, while the weights for the LSTM are initialised from a random normal distribution with values between -0.08 and 0.08 as per [16]. For the final caption prediction experiment (PRED4) we attempted to domain transfer the updated CNN weights from the DET2 caption detection model. This was attempted to avoid corruption of CNNs during end-to-end training (discussed in Sect. 4).

## 3.3 Training

The loss optimised during training is the summed cross entropy of the output of the softmax compared to the one-hot encoding of the next word in the ground truth sentence. This loss was minimised with standard Stochastic Gradient Descent (SGD) using an initial learning rate of 1.0 and a decay procedure that reduced the learning rate by half every 8 epochs (there were 164541 examples in the training set and a batch size of 16, so each epoch contains 10284 mini-batches). Gradients were clipped to 5.0 for all experiments.

---

<sup>1</sup> As per the Tensorflow-Slim implementation available at: <https://github.com/tensorflow/models/tree/master/im2txt>



**Fig. 1.** Schematic of the Neural Image Captioning architecture with a validation image and the actual generated caption.

### 3.4 Inference

As suggested by Vinyals et al. [19] we use Beam Search to generate sentences at inference time. This avoids the non-trivial issue that greedily selecting the most probable word at each time-step may result in a sentence which is itself of low probability. Ideally we would search the entire space for the most probable sentence, however, this would have an exponential computational cost associated with it as a forward pass through the entire model must be made for each node of the search tree. Therefore some search procedure is required in order to find the most probable sentence given limited computational resources. Our best results were achieved with a beam size of 3 and maximum caption length of 50.

### 3.5 Post Processing

The sentence output of the concept detection task was converted to an ordered set of concept labels.

### 3.6 Details of Submitted Runs

Tables 1 & 2 detail the specifics of the 3 runs submitted to the concept detection challenge and the 4 runs submitted the caption prediction challenge.

Concept Detection	
Run	Description
DET1	CNN weights frozen. LSTM and embedding layers trained for 958069 minibatches (approx. 90 epochs)
DET2	DET1 trained end-to-end for an additional 2658763 minibatches (approx. 350 epochs total). Model suffered from destructive gradient issue discussed below.
DET3	Naive merge of DET1 & DET2, using a set union of each model’s predicted labels

**Table 1.** Details of Concept Detection submitted runs.

Caption Prediction	
Run	Description
PRED1	CNN weights frozen. LSTM and embedding layers trained for 1499176 minibatches (approx. 145 epochs)
PRED2	CNN weights frozen. 3-layer LSTM with a single dropout layer. Trained for 998981 minibatches approx. 97 epochs)
PRED3	Naive Merge of PRED1 and PRED2 based on most non-known words in sentence
PRED4	CNN Domain transfer from fine tuned detection task model DET3. 3-layer LSTM with a single dropout layer. CNN weights then frozen and LSTM and embedding layers trained for 437805 minibatches (approx. 42 epochs)

**Table 2.** Details of Caption Prediction submitted runs.

## 4 Results

Tables 3 & 4 detail the training, validation and test results of the various runs. Please note that due to the high cost of inference, the training scores are estimated based on a random sample of 10000 images from the training set.

We attempted a two-phase training procedure as suggested by Vinyals et al. [20] for DET2 and some unsubmitted experiments. In the first phase we froze the CNN weights and trained only the LSTM and embedding layers. Then, once the language model had begun to converge we trained the entire model end-to-end with a very small learning rate ( $1e - 5$ ). This is suggested by the Vinyals et al. as necessary as otherwise the CNN model will become corrupted and never recover. However, we found that despite training the LSTM for a very long time in the first phase and using a very small learning rate in the second phase we would very quickly corrupt the CNN as evidenced by a sharp increase in dead ReLUs and a large decrease in BLEU score. We found that BLEU scores would eventually return to those achieved in the first phase of training, however, the dead ReLUs did not revive. We believe that the underlying issue is that the degree of domain transfer required to go from general images to medical images is vastly greater than that required to go from one collection of general images to another (i.e. ImageNet to MSCOCO).

Concept Detection			
	Average F1 Score		
Run	Train (estimated*)	Validation	Test
DET1	0.1055	0.1088	0.0838
DET2	0.1119	0.1117	0.0880
DET3	0.0860	0.0952	0.0958

**Table 3.** Details of Concept Detection results. \*Training score estimated on a random sample of 10k training images.

Caption Prediction			
	Average BLEU Score		
Run	Train (estimated*)	Validation	Test
PRED1	0.1367	0.1315	0.0656
PRED2	0.1590	0.1533	0.0851
PRED3	0.1383	0.1734	0.0982
PRED4	0.1556	0.1489	0.0826

**Table 4.** Details of Caption Prediction results. \*Training score estimated on a random sample of 10k training images.

#### 4.1 Qualitative Analysis

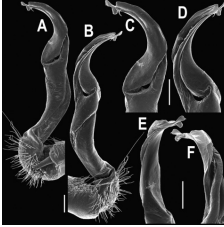

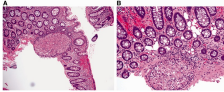
Table 5 shows some generated and predicted captions from the validation set. For the first image, the model has learned to generate the exact same caption as the ground truth, with the exception of the second word which is unknown. This exact caption, with only the second word different, appears alongside 21 images in the training set. The unknown word, ‘latiterga’, does not appear in the training set. In the second example, the model has correctly identified the modality, orientation and anatomy from the image and has generated a very similar sentence even though no such sentence exists in the training set. It has not, however, determined that this is a preoperative image. The third example demonstrates a very poor result. The model has not correctly determined that there are two subfigures, although it has correctly estimated the magnification of the left subfigure.

## 5 Conclusion

Based on the small variance between our training and validation scores we do not believe that the models were overfitting, however, the large variance between validation and test scores indicates that there was a large disparity between the training and validation data and the data in test set. Based on the non-overfitting analysis we could potentially train a much larger vision model for a longer time and improve the overall performance.

Additionally the fact that we could not successfully train the vision model without corrupting the network was a major limiting factor in our experiments.

Future work will investigate the potential of larger language models and devising a training regime that allows true end-to-end training for medical images.

	Actual	orthomorpha latiterra sp n holotyp b right gonopod mesal later view respect cf distal part right gonopod mesal later subor subcaud view respect scale bar 02 mm
	Predicted	orthomorpha unk sp n holotyp b right gonopod mesal later view respect cf distal part right gonopod mesal later subor subcaud view respect scale bar 02 mm
	Scores	<b>BLEU</b> : 0.9304 <b>BLEU1</b> : 0.9629 <b>BLEU2</b> : 0.92 <b>BLEU3</b> : 0.9231 <b>BLEU4</b> : 0.9167
	Actual	preoper later radiograph right knee
	Predicted	later radiograph right knee
	Scores	<b>BLEU</b> : 0.7788 <b>BLEU1</b> : 1.0 <b>BLEU2</b> : 1.0 <b>BLEU3</b> : 1.0 <b>BLEU4</b> : 1.0
	Actual	discret epithelioid granuloma crohn diseas discret epithelioid granuloma associ epitheli injuri stain 100x b discret epithelioid granuloma musculari mucosa stain 200x histiocyt epithelioid contain abund eosinophil cytoplasm
	Predicted	histolog examin resect specimen hematoxylin eosin stain magnif 100
	Scores	<b>BLEU</b> : 0.0781 <b>BLEU1</b> : 0.1111 <b>BLEU2</b> : 0.0 <b>BLEU3</b> : 0.0 <b>BLEU4</b> : 0.0

**Table 5.** Sample of predicted and actual captions with associated BLEU metrics.

## Acknowledgement

The authors are grateful to the NVIDIA Corporation for their donation of the Titan X GPU used in this research.



## References

1. Bird, S.: NLTK: The natural language toolkit. In: Proceedings of the COLING/ACL on Interactive Presentation Sessions. pp. 69–72. COLING-ACL '06, Association for Computational Linguistics, Stroudsburg, PA, USA (2006)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (Jun 2009)
3. Devlin, J., Gupta, S., Girshick, R., Mitchell, M., Lawrence Zitnick, C.: Exploring nearest neighbor approaches for image captioning (May 2015)
4. Eickhoff, C., Schwall, I., García Seco de Herrera, A., Müller, H.: Overview of ImageCLEFcaption 2017 - image caption prediction and concept detection for biomedical images. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <<http://ceur-ws.org>>, Dublin, Ireland (September 11-14 2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (Dec 2015)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* 9(8), 1735–1780 (Nov 1997)
7. Ionescu, B., Mller, H., Villegas, M., Arenas, H., Boato, G., Dang-Nguyen, D.T., Dicente Cid, Y., Eickhoff, C., Garcia Seco de Herrera, A., Gurrin, C., Islam, B., Kovalev, V., Liauchuk, V., Mothe, J., Piras, L., Riegler, M., Schwall, I.: Overview of ImageCLEF 2017: Information extraction from images. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017. Lecture Notes in Computer Science, vol. 10456. Springer, Dublin, Ireland (2017)
8. Kisilev, P., Sason, E., Barkan, E., Hashoul, S.: Medical image description using multi-task-loss CNN. In: Deep Learning and Data Labeling for Medical Applications, pp. 121–129. Springer, Cham (Oct 2016)
9. Kisilev, P., Walach, E., Hashoul, S., Barkan, E., Ophir, B., Alpert, S.: Semantic description of medical image findings: structured learning approach. In: Proceedings of the British Machine Vision Conference 2015. pp. 171.1–171.11. British Machine Vision Association (2015)
10. Kumar, A., Kim, J., Lyndon, D., Fulham, M., Feng, D.: An ensemble of Fine-Tuned convolutional neural networks for medical image classification. *IEEE journal of biomedical and health informatics* 21(1), 31–40 (Jan 2017)
11. Li, S., Kulkarni, G., Berg, T.L., Berg, A.C., Choi, Y.: Composing simple image descriptions using web-scale n-grams. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning. pp. 220–228. CoNLL '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
12. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2Text: Describing images using 1 million captioned photographs. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems* 24, pp. 1143–1151. Curran Associates, Inc. (2011)
13. Schlegl, T., Waldstein, S.M., Vogl, W.D., Schmidt-Erfurth, U., Langs, G.: Predicting semantic descriptions from medical images with convolutional neural networks. In: *Information Processing in Medical Imaging*. pp. 437–448. Springer, Cham (Jun 2015)
14. Shin, H.C., Lu, L., Kim, L., Seff, A., Yao, J., Summers, R.M.: Interleaved text/image deep mining on a very large-scale radiology database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1090–1099 (2015)

15. Shin, H.C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R.M.: 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. pp. 2497–2506 (2016)
16. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 3104–3112. Curran Associates, Inc. (2014)
17. Syeda-Mahmood, T., Kumar, R., Compas, C.: Learning the correlation between images and disease labels using ambiguous learning. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 185–193. Springer, Cham (Oct 2015)
18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision (Dec 2015)
19. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator (Nov 2014)
20. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Trans. Pattern Anal. Mach. Intell. (Jul 2016)
21. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, ., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google’s neural machine translation system: Bridging the gap between human and machine translation (Sep 2016)
22. Yao, B.Z., Yang, X., Lin, L., Lee, M.W., Zhu, S.C.: I2t: Image parsing to text description. Proc. IEEE 98(8), 1485–1508 (Aug 2010)