

# Keyword Generation for Biomedical Image Retrieval with Recurrent Neural Networks

FHDO Biomedical Computer Science Group (BCSG)

Obioma Pelka<sup>1,2</sup> and Christoph M. Friedrich<sup>1</sup>

<sup>1</sup> Department of Computer Science  
University of Applied Sciences and Arts Dortmund (FHDO)  
Emil-Figge-Strasse 42, 44227 Dortmund, Germany  
obioma.pelka@fh-dortmund.de and christoph.friedrich@fh-dortmund.de  
<http://www.inf.fh-dortmund.de>

<sup>2</sup> Duisburg-Essen University School of Medicine  
Hufelandstrasse 55, 45147 Essen, Germany

**Abstract.** This paper presents the modeling approaches performed by the FHDO Biomedical Computer Science Group (BCSG) for the caption prediction task at ImageCLEF 2017. The goal of the caption prediction task is to recreate original image captions by detecting the interplay of present visible elements. A large-scale collection of 164,614 biomedical images, represented as imageID - caption pairs, extracted from open access biomedical journal articles (PubMed Central) was distributed for training. The aim of this presented work is the generation of image keywords, which can be substituted as text representation for classifications tasks and image retrieval purposes. Compound figure delimiters were detected and removed as estimated 40% of figures in PubMed Central are compound figures. Text preprocessing such as removal of stopwords, special characters and Porter stemming were applied before training the models. The images are visually represented using a Convolutional Neural Network (CNN) and the Long Short-Term Memory (LSTM) based Recurrent Neural Network (RNN) Show-and-Tell model is adopted for image caption generation. To improve model performance, a second training phase is initiated where parameters are fine-tuned using the pre-trained deep learning networks Inception-v3 and Inception-ResNet-v2. Ten runs representing the different model setups were submitted for evaluation.

**Keywords:** biomedical image retrieval, keyword generation, computer vision, convolutional neural networks, long short-term memory, recurrent neural network

## 1 Introduction

This paper describes the modeling methods and experiments performed by the FHDO Biomedical Computer Science Group (BCSG) at the ImageCLEF 2017

[8] Caption Prediction Task. The caption prediction task, which aims to recreate original image captions by detecting the interplay of present visible elements [5], is addressed in this paper. The focus of this presented work is more on the automated generation of keywords for biomedical and medical images and not caption prediction. Several approaches [3, 9, 10, 12] have shown that combining visual image representation with text obtains better image classification performance. However, for some image classification tasks, such as ImageCLEF2009 Medical Annotation Task [4], corresponding text representations are not available. These keywords can be substituted as text representations and combined with visual representations to obtain multi-modal image representations. These multi-modal image representations can be further adopted for image retrieval purposes.

The remaining of this paper is organized as follows: Section 2 explains the methodology adopted. The image keyword generation setups, submitted runs and results are displayed and discussed in section 3. Finally, conclusions are drawn in section 4.

## 2 Methodology

### 2.1 Dataset

All figures distributed in the ImageCLEF 2017 Caption Prediction Task originate from biomedical literature published in PubMed Central. The training set contains 164,614 image - caption pairs. An additional validation set of 10,000 biomedical image - caption pairs were distributed for evaluation purposes in the development stage. For the official evaluation, computed using BLEU scores [11], a test set of 10,000 biomedical images was distributed. For keyword generation tasks, BLEU score is not suited as an evaluation metric. The order of words and length of captions have significant effects on the calculated scores. Further information is detailed in [5].

### 2.2 Data Preprocessing

Focusing on image keyword generation, certain contents in biomedical figure captions are undesirable and were omitted. The preprocessing steps done before model training were:

**Compound Figure Delimiter:** Estimated 40% of biomedical figures in PubMed Central are compound figures [6]. These captions most likely address the subfigures using delimiters. Such delimiters were detected and removed. An excerpt of delimiters removed is listed in [12].

**English Stopwords:** Using the NLTK Stopword corpus, present stopwords in the captions were omitted. This corpus contains 2,400 stopwords for 11 languages [2].

**Special Characters:** Special characters such as symbols, punctuations, metrics, etc. were removed.

**Single Digits:** Single digits, words which consist of just numbers, were removed.

**Word Stemming:** To reduce complexity, the captions are stemmed using Porter Stemming [13]. For evaluation comparison, not all models were trained using stemmed captions. An overview of model setup is listed in Table 1. The Snowball stemming method [14] is used for the official evaluation by the task organizers.

**Vocabulary Size:** Using the ImageCLEF 2017 Caption Prediction Task Training Set, three vocabularies were generated with different minimum word occurrence cutoffs:

- Vocab01: 21,191 Words {Cutoff  $\geq$  4; Word Stemming}
- Vocab02: 14,295 Words {Cutoff  $\geq$  4; Word Stemming}
- Vocab03: 18,629 Words {Cutoff  $\geq$  7; No Word Stemming}

### 2.3 Image Keyword Generator

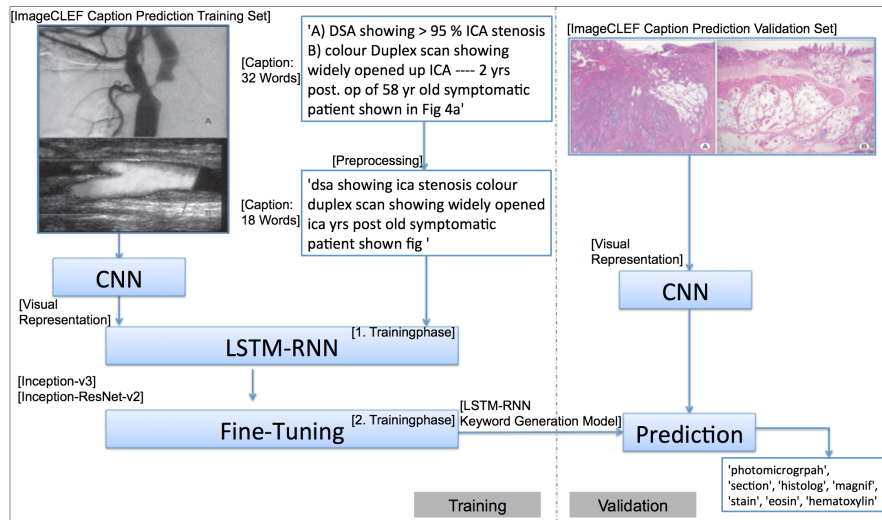
For keyword generation, a combination of encoding and decoding using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) [7] based Recurrent Neural Networks (RNN) [1] is adopted. This approach, also known as Show-And-Tell model was proposed in [17] and further improved in [18].

The CNN is used as an image encoder, to produce rich visual representations of the images, by pre-training it for an image classification task. The LSTM-RNN utilized as caption decoder generates the image keywords, using the CNN last hidden layer as input [17]. The parameters for the image keyword generation model are:

1. Minibatch size = [1. Trainingphase = 32; 2. Trainingphase = 4]
2. Vocabulary size = 23,000
3. Initial learning rate = 2
4. Model optimizer = stochastic gradient descent
5. Learning rate decay factor = 0.5
6. Number of epochs per decay = 8
7. Inception learning rate = 0.0005
8. Inception model initialization = Inception-v3
9. LSTM embedding size = 512
10. LSTM units number = 512
11. LSTM initializer scale = 0.08
12. LSTM dropout keep probability = 0.7

In the first training phase, the LSTM is trained using a corpus of paired image and captions generated from the biomedical figures in the ImageCLEF 2017 Caption Prediction Task Training Set. No further dataset was used for training.

Several models were further trained in the second training phase. In the second phase, parameters of the image submodel and LSTM are fine-tuned using the deep learning networks Inception-v3 [16] and Inception-ResNet-v2 [15]. Figure 1 shows the keyword generation model training setup.



**Fig. 1.** Overview of Long Short-Term Memory based Recurrent Neural Network Model applied for biomedical image keyword generation.

## 2.4 Model Setup

**Table 1.** Model training setups applied for image keyword generation

Run ID	PorterStemming	Special Characters Stopwords	Compound figure delimiters	Cutoff Words	Vocabulary Size	1. Trainingphase Number of Epochs	2. Trainingphase Number of Epochs	Deep Learning Network	BLEU Score Validation Set
R01	✓	✓	✓	4	Vocab01	25	4	Inception ResNet-v2	0.0686
R02	✓	✓	✓	4	Vocab01	27	✗	✗	0.0670
R03	✓	✓	✓	4	Vocab01	25	4	Inception-v3	0.0674
R04	✓	✓	✓	7	Vocab02	25	4	Inception ResNet-v2	0.0336
R05	✓	✓	✓	7	Vocab02	25	4	Inception-v3	0.0323
R06	✓	✓	✓	7	Vocab02	27	✗	✗	0.0579
R10	✗	✓	✓	7	Vocab03	27	✗	✗	0.0918
R11	✗	✓	✓	7	Vocab03	25	4	Inception ResNet-v2	0.0661
R12	✗	✓	✓	7	Vocab03	25	4	Inception-v3	0.0656
R16	✗	✓	✓	7	Vocab03	39	7	Inception-v3	0.0678

Several model setups were evaluated and those selected for creating submission runs are listed in Table 1. Columns 2 - 4 display applied preprocessing methods. The columns 5 and 6 shows the minimum word occurrence cutoff and vocabulary size, as described in subsection 2.2, respectively. The number of epochs for the first and second training phase are listed in columns 7 and 8, respectively. Column 9 shows deep learning networks adopted for parameter fine-tuning.

### 3 Submitted Runs

Based on the model training setups listed in Table 1, ten runs were submitted for evaluation. Image keywords were generated for 10,000 biomedical images distributed in the ImageCLEF 2017 Caption Prediction Task Test Set. These runs contain several ensembles of the model setups:

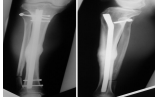
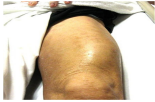
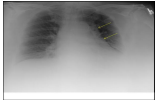
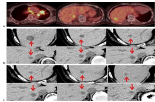
- **PRED\_Sub01:** Combination of keywords from models R10, R11, and R12
- **PRED\_Sub02:** Combination of keywords from models R01, R02, and R03
- **PRED\_Sub03:** Predicted keywords from model R12
- **PRED\_Sub04:** Predicted keywords from model R11
- **PRED\_Sub05:** Combination of keywords from models R04, R05, and R06
- **PRED\_Sub06:** Predicted keywords from model R03
- **PRED\_Sub07:** Predicted keywords from model R01
- **PRED\_Sub08:** Concatenation of keywords from models R01 and R04
- **PRED\_Sub09:** Concatenation of keywords from models R03 and R05
- **PRED\_Sub10:** Predicted keywords from model R16

For better understanding the difference between predicted keywords *Combination* and *Concatenation* is explained as follows:

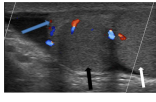
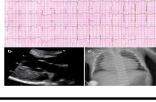
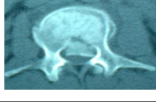
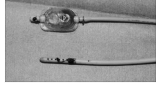
**Combination = OR:** The keyword generator models were not always able to predict the caption of a given image. Some results were `<UNK>` representing an empty string. In such cases, the predicted keywords of three models are combined. Taking PRED\_Sub01 for example: when model setup *R01* returns an empty string, the final results is substituted with the predicted keywords from model setup *R02*. In the case where *R02* returns an empty string as well, the predicted keywords of *R03* is taken as the final caption. When all three models predict `<UNK>`, the final result is 'unknown'. This process is highlighted in Table 2. All three models have the same preprocessing steps and vocabulary sizes but differ in the second training phase.

**Concatenation = AND:** The predicted keywords of two models are simply concatenated. Both models were trained using the same preprocessing methods, first and second training phase. The minimum cutoff for word occurrence is different. Multiple keywords are removed. An example using submission run PRED\_Sub08 is shown in Table 3.

**Table 2.** The combination of predicted captions as done for submission run PRED\_Sub01. The images shown were hand picked from the ImageCLEF 2017 Caption Prediction Task Validation Set.

Image	Setup R11	Setup R12	Setup R10	Final Caption
	'anteroposterior', 'left', 'knee', 'radiograph'	'preoperative', 'radiograph'	'anteroposterior', 'view'	'anteroposterior', 'left', 'radio- graph', 'knee'
	<UNK>	'clinical', 'photo- graph', 'patient'	'swelling', 'leg'	'clinical', 'photo- graph', 'patient'
	<UNK>	<UNK>	'chest, 'ray'	'chest, 'ray'
	<UNK>	<UNK>	<UNK>	'unknown'

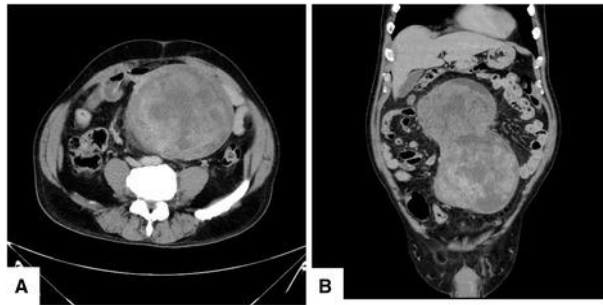
**Table 3.** The concatenation of predicted captions as done for submission run PRED\_Sub08. The images shown were hand picked from the ImageCLEF 2017 Caption Prediction Task Validation Set.

Image	Setup R01 Caption	Setup R04 Caption	Final Caption
	'ultrasound', 'imag'	'doppler'	'ultrasound', 'imag', 'doppler'
	'ecg'	<UNK>	'ecg'
	<UNK>	'scan', 'comput' 'tomographi', 'pelvi'	'scan', 'comput', 'tomographi', 'pelvi'
	<UNK>	<UNK>	<UNK>

### 3.1 Results

The evaluation metrics achieved with the submitted runs for the ImageCLEF 2017 Caption Prediction Task is listed in Table 5. For a single biomedical image, the predicted keywords and corresponding BLEU scores are shown in Table 4. The image was hand picked from the ImageCLEF 2017 Caption Prediction Task Validation Set. For better comparison, the ground truth caption is shown below the image.

**Table 4.** Predicted keywords and achieved BLEU scores of single biomedical figure hand picked from ImageCLEF 2017 Caption Prediction Task Validation Set



'(A) Abdominal computed tomography demonstrated a 25 x 11 cm, heterogeneous, lobulated mass in the abdominal cavity. (B) Colonal view demonstrated lobulated mass.'

ID	Predicted Keywords	BLEU
Sub01	'abdomen' 'tomography' 'computed' 'scan' 'abdominal'	0.0240
Sub02	'abdomen' 'tomographi' 'comput' 'scan' 'abdomin'	0.0240
Sub03	'kidney' 'right' 'mass' 'cystic' 'large' 'showing' 'abdomen' 'tomography' 'computed' 'scan' 'abdominal'	0.2608
Sub04	'abdomen' 'tomography' 'computed' 'scan' 'abdominal'	0.0240
Sub05	'kidnei' 'right' 'lesion' 'cystic' 'larg' 'show' 'pelvi' 'abdomen' 'scan' 'tomographi' 'comput'	0.2193
Sub06	'abdomen' 'tomographi' 'comput' 'scan' 'abdomin'	0.0240
Sub07	'abdomen' 'imag' 'axial' 'scan' 'tomographi' 'comput' 'abdomin'	0.0822
Sub08	'abdomen' 'imag' 'axial' 'scan' 'tomographi' 'comput' 'abdomin' 'enhanc' 'contrast' 'pelvi'	0.2016
Sub09	'kidnei' 'right' 'lesion' 'cystic' 'larg' 'show' 'pelvi' 'abdomen' 'scan' 'tomographi' 'comput' 'abdomin'	<b>0.2827</b>
Sub10	abdomen' 'tomography' 'computed' 'scan' 'abdominal'	0.0240

The first and second columns of Table 5 list the mean BLEU [11] score obtained on the ImageCLEF 2017 Caption Prediction Task Test and Validation Set

respectively. Both datasets contain 10,000 biomedical figures. The third column displays the precision score obtained on the validation set.

**Table 5.** Evaluation metrics obtained on the official test and validation set for all submitted runs. Each set contains 10,000 biomedical figures.

Run ID	Test Set	Validation Set	Validation Set	Validation Set
	BLEU Score	BLEU Score	Precision	Recall
PRED_BCSG_Sub01	0.0624	0.0772	0.1782	0.1245
PRED_BCSG_Sub02	0.0411	0.0687	0.2270	0.1582
PRED_BCSG_Sub03	0.0527	0.0656	0.1769	0.1253
PRED_BCSG_Sub04	0.0537	0.0661	0.1782	0.1245
PRED_BCSG_Sub05	0.0200	0.0428	0.1310	0.0828
PRED_BCSG_Sub06	0.0365	0.0674	0.2281	0.1582
PRED_BCSG_Sub07	0.0375	0.0686	0.2279	0.1612
PRED_BCSG_Sub08	0.0675	<b>0.1111</b>	<b>0.2495</b>	<b>0.1888</b>
PRED_BCSG_Sub09	<b>0.0749</b>	0.1086	0.2431	0.1861
PRED_BCSG_Sub10	0.0326	0.0678	0.2108	0.1418

### 3.2 Discussion

Analyzing Table 5, it can be seen that submitting keywords instead of captions for evaluation on the ImageCLEF 2017 Caption Prediction Task Test Set achieved low BLEU scores. The best score was attained on the test set with submission run PRED\_Sub09. This is a concatenation of predicted keywords using model setup R03 and R05. Both models parameters were fine-tuned using the deep learning network Inception-v3 and were trained with different vocabulary sizes.

On the validation set, the best score was obtained with submission run PRED\_Sub08, which is the concatenation of predicted keywords using models R01 and R04. This run is similar to PRED\_Sub09 with the exception of parameter fine-tuning with Inception-ResNet-v2. The BLEU scores achieved on the validation set are similar to those of the test set. Captions of biomedical figures mostly consist of multiple sentences and can not be accurately predicted using few keywords, as word order and caption length have effects on the calculated scores.

The precision score is one of the adequate metrics for image keyword generation. The best precision score was obtained using submission run PRED\_Sub08. With more extensive text preprocessing steps, higher precision scores can be expected.

The removal of compound figure delimiters, stop words, single numbers and special characters led to compact and precise captions. However, captions contain



several adjectives, pronouns, adverbs etc. which do not necessarily describe the semantic content, characteristics or modality of the images. The reduction of captions to contain just nouns is one preprocessing steps that should be evaluated with the aim of modeling an accurate image keyword generator.

In comparison to the Show-And-Tell Model, a smaller number of epochs was used to train the image keyword generator. The Show-And-Tell Model has 387 epochs in the first training phase and 1,160 epochs in the fine-tuning phase [17, 18]. Potentially, training the image keyword generator with a larger number of epochs could improve the results.

## 4 Conclusion

An approach for image keyword generation was presented. Using image - caption pairs of 164,614 biomedical figures, distributed for training at the ImageCLEF Caption Prediction Task, long short-term memory based Recurrent Neural Network models were trained. All compound figure delimiters, stop words, special characters and single numbers were removed from captions before training. For comparison, some models were trained with stemmed captions and different vocabulary sizes. These vocabularies were obtained by using different minimum word occurrence cutoffs. The BLEU and precision scores were applied as evaluation metrics. With the aim of further model accuracy improvement, the reduction of captions to just nouns before training the models should be evaluated. To increase keyword prediction ability, the models should be trained and fine-tuned with a higher number of epochs, as proposed in the Show-And-Tell model. These automatically generated keywords can be substituted as text representation for classification tasks and image retrieval purposes will be researched and evaluated in future work.

## References

1. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5(2), 157–166 (1994)
2. Bird, S., Klein, E., Loper, E.: *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc. (2009)
3. Codella, N., Connell, J., Pankanti, S., Merler, M., Smith, J.R.: Automated medical image modality recognition by fusion of visual and text information. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI Conference Proceedings 2014*, pp. 487–495. Springer (2014)
4. Dimitrovski, I., Kocev, D., Loskovska, S., Džeroski, S.: Imageclef 2009 medical image annotation task: PCTs for hierarchical multi-label classification. In: *Multilingual Information Access Evaluation II. Multimedia Experiments*, pp. 231–238. Springer (2009)
5. 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)

6. García Seco de Herrera, A., Müller, H., Bromuri, S.: Overview of the ImageCLEF 2015 medical classification task. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. (2015)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* 9(8), 1735–1780 (1997)
8. Ionescu, B., Müller, 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 (September 11-14 2017)
9. Kalpathy-Cramer, J., de Herrera, A.G.S., Demner-Fushman, D., Antani, S.K., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems - an overview of the medical image retrieval task at imageclef 2004-2013. *Comp. Med. Imag. and Graph.* 39, 55–61 (2015)
10. Koitka, S., Friedrich, C.M.: Traditional feature engineering and deep learning approaches at medical classification task of imageclef 2016. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS. org, Évora, Portugal (September 5-8 2016). pp. 304–317 (2016)
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
12. Pelka, O., Friedrich, C.M.: FHDO biomedical computer science group at medical classification task of imageclef 2015. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. (2015)
13. Porter, M.F.: An algorithm for suffix stripping. *Program* 14(3), 130–137 (1980)
14. Porter, M.F.: Snowball: A language for stemming algorithms (2001)
15. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. pp. 4278–4284 (2017)
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)
17. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3156–3164 (2015)
18. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(4), 652–663 (2017)