Recurrent Attention Networks for Medical Concept Prediction

Sam Maksoud, Arnold Wiliem, and Brian Lovell

School of Information Technology and Electrical Engineering, The University of Queensland,
Brisbane QLD, Australia

Abstract. This paper presents the working notes for the CRADLE group's participation in the ImageCLEF2019 medical competition. Our group focused on the concept detection task which challenged participants to approximate the mapping from radiology images to concept labels. Traditionally, such a task is often modelled as an image tagging or image retrieval problem. However, we empirically discovered that many concept labels had weak visual connotations; hence, image features alone are insufficient for this task. To this end, we utilize a recurrent neural network architecture which enables our model to capture the relational dependencies among concepts in a label set to supplement visual grounding when their association to image features is weak or unclear. We also exploit soft attention and visual gating mechanisms to enable our network to dynamically regulate "where" and "when" to extract visual data for concept generation.

1 Introduction

In 2019, ImageCLEF [8] hosted its 3rd edition of the medical image captioning task. The participants of this task were challenged to develop a method for generating Concept Unique Identifiers (CUI) to describe the contents of a radiology image [13]. In contrast to natural language captions, CUIs parse out standardized concept terms from the medical texts. Resolving captions into key concepts alleviates the constraint of modelling the syntactic structures of free text. Removing the language modelling component results in a task akin to image tagging i.e. identifying the presence of a label (CUI) by its most distinguishable visual features.

However, a considerable number of CUI terms in the supplied subset of the ROCO dataset [14] have no obvious association to visual features. This is due to the fact that the CUIs were extracted automatically from the figure captions using only natural language processing tools; there was no constraint for the CUIs to be associated with visual features. Consequently, concepts with weak visual connotations such as "study , "rehab" and "supplement" are abundant throughout the ROCO dataset [14]; it is unreasonable to assume that a model can learn general visual features to reliably identify such concepts. While it is true that in isolation, accurately identifying these non-visual concepts is unlikely or impossible, their relevance to an image can be indirectly estimated by modelling relational dependencies to other CUIs in the set of concept labels. This is because all CUIs in a set of concepts are derived from a common source: the original figure caption.

Under these conditions, our group concluded it would be best to model the problem as an image to sequence translation task; emphasizing the need to map an image to a set of concepts, rather than mapping individual CUIs directly to image features. Thus, we design our model as a recurrent neural network (RNN) given their unrivalled performance in capturing the long term dependencies in sequential data [11]. Our proposed RNN is conditioned on features from both the image and CUI labels. We utilize a soft attention mechanism [18] which dynamically attends to different regions of an image in order to select the most distinguishable visual features for each CUI. In situations where a CUI has weak visual connotations, a visual feature gating mechanism [18] allows the model to focus on textual features as they are likely to provide greater discriminatory power in such contexts.

2 Dataset Challenges

In order to design an appropriate model for the task, our group carried out an extensive investigation of the supplied subset of the ROCO dataset [14]. During this investigation we identified several challenges that would complicate the task of mapping text to visual features. These challenges pertain to incidences of redundant, inconsistent, and/or nonsensical assignment of CUIs to an image. We describe these challenges and how they influenced our approach to this task in detail below.

Table 1. Top 10 most frequent concepts in the training data.

RANK	CUI	FREQUENCY	CONCEPT
1	C0040395	6033	tomogr
2	C0034579	6002	pantomogr
3	C0043299	5830	x-ray procedure
4	C0441633	5283	diagnostic scanning
5	C1548003	5045	radiograph
6	C1962945	5044	radiogr
7	C0817096	4794	thoracics
8	C0772294	4372	alesion
9	C0040405	3113	x-ray computer assisted tomography
10	C0009924	2771	materials/contrast media

First and foremost, we identified that a majority of concepts redundantly describe generic radiology images. In Table. 1 we list the top 10 most frequent concepts in the training dataset. Eight out of the top 10 concepts (all but "alesion" and "thoracics") could arguably describe most of the radiology images in the dataset. The ROCO dataset [14] exclusively contains radiological images; a concept such as "radiograph" would be appropriate for all of the images. However, since the umbrella concept of "radiograph" can be expressed using a variety of different CUIs, we are forced to find arbitrary features to distinguish these cognate identifiers. The CUIs C1548003 and C1962945 describe "radiograph" as a diagnostic procedure and a diagnostic service ID respectively. While distinguishing these different types of "radiograph" is trivial in natural language contexts, identifying discriminating visual features is an extremely dubious pursuit. As such, any model tasked with learning the haphazard distribution of these semantically interchangeable (and often universal) concepts in the ROCO dataset [14] is expected to have limited generalizability.

This property of the dataset has implications on the F1 score used to evaluate this task. The F1 score will penalize models for misidentifying the arbitrary instances or absences of these CUIs in the test data. This is because of the inherent stochasticity of the ROCO dataset [14]; where unobservable variations in source figure captions determine the CUIs assigned to a sample.

In the supplied subset of the ROCO dataset [14], we observe recurring patterns of these semantically similar CUIs. For example C0043299 and C1962945 occur frequently as a pair but they also regularly occur as a tripartite alongside C1548003. An RNN architecture enables our model to exploit the statistical co-occurrence of these concepts when modelling probability distributions for a set of CUIs [7, 11]. To achieve a competitive F1 score, a model must not only learn "what" visual features best represent a CUI, but also "when" that CUI is most likely to occur in a given set of labels. Since all CUIs in a set of labels are derived from the same figure caption, modelling their interdependencies will ensure our model is more robust to the unobservable variations in the original figure caption. This enables our model to more reliably predict "when" a label is assigned to an image based on the learned co-occurrence statistics with previously generated concepts.

Another challenge encountered in this task is the assignment of nonsensical CUIs by the quickUMLS system [16] used to create the ROCO dataset [14]. The quickUMLS system utilizes the CPMerge algorithm for dictionary mapping [16]. CPMerge uses character trigrams as features and maps terms to a dictionary based on overlapping features [16]. This method introduces a significant source of error resulting in random and nonsensical CUIs being extracted from a medical figure caption. Table. 3 showcases examples of when trigram feature matching resulted in nonsensical or redundant CUIs being assigned to an image. This presents a major obstacle for multi-modal retrieval as minor changes in descriptive syntax results in significant and erratic variations of the CUIs extracted from the figure caption. For example, one could rephrase the last sentence for ROCO CLEF ID 25756 in Table. 3 as "*Visualization of the proximal ACL is poor, suggesting an ACL rupture*". The arbitrary decision to remove the word "*substance*" would

Table 2. This table compares CUIs produced by MetaMap [1] compared to the quickUMLS [16] for the following medical text: "Intensity modulated radiotherapy (IMRT) planning axial CT post-contrast showing a residual post-operative cystic nodal metastasis from papillary carcinoma (arrow). The patient underwent a further neck dissection to remove the node before IMRT was performed." (source caption for ROCO CLEF ID 42724).

quickUMLS	MetaMap
C0034619 (radiother)	C1512814 (Intensity modulated radiotherapy)
C1522449 (radiation therapy)	C0032074 (Planning)
C2939420 (mets)	C0205131 (Axial)
C0027530 (collum)	C3888140 (CT)
C0012737 (dissection procedure)	C1609982 (Residual)
C0226964 (papillae linguales)	C0032790 (Postoperative Period)
C0935624 (capillaris)	C0205207 (Cystic)
C0746922 (noded)	C0443268 (Nodal)
C1328685 (metastat)	C0027627 (Metastasis)
C0227296 (papilla)	C0007133 (Carcinoma, Papillary)
C0027627 (spreading of cancer)	–
C0006901 (smallest blood vessel)	–

no longer produce the erroneous CUI describing 11-Deoxycortisol (C0075414) based on its common name *"Reichsteins Substance S*".

To satisfy our scientific curiosity, we compared CUIs extracted using quickUMLS to those extracted by MetaMap [1]; as MetaMap is a commonly used alternative for automatic concept extraction. In the particular instance shown in Table. 2, the CUIs produced by MetaMap are undoubtedly higher quality than those produced by quickUMLS. This is likely due to the fact that MetaMap does not use trigram character matching [1] and so it accurately captures C0007133 (Papillary Carcinoma) instead of C0226964 (Papilla of tongue). In the original paper describing quickUMLS [16], the authors claim the quickUMLS system could outperform MetaMap in certain tasks. However, an important caveat of this claim is that they use SpaCy models to pre-process texts instead of the MetaMaps inbuilt pre-processing tools [16]). SpaCy pre-processing models are trained on a general text corpus whereas MetaMap utilizes the SPECIALIST lexicon [1]. Medical terms are highly featured in the SPECIALIST lexicon [3]; the lexems are likely to be more representative of those seen in radiology figure captions. Thus, substituting MetaMap's pre-processing tools with SpaCy's may not accurately reflect the performance of the end-to-end MetaMap system.

Furthermore, we empirically discovered that certain semantic types were more prone to erroneous assignment; the majority of nonsensical CUIs encountered were chemical names and abbreviations. In light of this issue, it may be worthwhile to investigate semantic types more prone to error and identify those which have the strongest visual connotations. This could assist our multi-modal retrieval models in determining how to weigh the importance of visual features and CUI relational dependencies based on the

Table 3. This table shows examples of erroneous CUI assignment. To retrieve the original caption, each image was used to query the original figure caption pair using the Openi image search engine [5].

ROCO CLEF ID	ORIGINAL CAPTION	ERRONEOUS CUIs	REASON FOR ERROR
24120	Tc99m pertechnetate thyroid scan did not show any tracer **concentration** by the thyroid gland	C0004268 (concentration) C0086045 (attention concentration) C3827302 (i can concentrate well)	The caption refers to the chemistry definition of concentration however is has also been mistakenly interpreted as a verb
25756	A twenty five year old female suffering from internal derangement of the left knee. The MRI report described ACL rupture due to poor visualization of the ACL **substance**.	C0075414 (Reichstein's Substance)	Poor understanding of sentence semantics has resulted in the word substance triggering the assignment of chemical Reichstein's Substance to the image.
22356	CECT abdomen showing the **lesion**	C0772294 (alesion)	Another example of chemical drug name Alesion (antihistamine) being mistakenly matched to a commonly used term (lesion).
24120	**Severe** Bilateral secretion and concentration **alterations**	C0076106 (Teration) C1306232 (Sever)	Poor trigram matching has produced CUIs for Teration (a type of Organothiophosphate) and Sever (verb) from alteration and severe respectively.

CUI's semantic type.

3 Proposed Methodology

To overcome the challenges described in Section. 2 we seek to construct a model that satisfies the following requirements:

1. It must be able to identify the most distinguishable visual characteristics for a CUI;
2. It must capture interdependences among CUIs in a set of labels and;
3. It must be able to regulate the weight of visual features based on the variable strength of a CUI's visual connotation.

To this end, the proposed methodology borrows many features from the works of Xu et al [18]. Although their architecture was originally designed for use in image captioning tasks, the dynamic soft attention mechanism, recurrent inductive bias of long short term memory networks (LSTM) [7] and deterministic visual gating mechanism can be exploited to satisfy requirements (1), (2) and (3) respectively. We describe our methodology in detail below.

Firstly, we resize all images to 244x244x3 pixels in order to exploit a VGG16 [15] convolutional neural network (CNN) pre-trained on ImageNet [6]. Although the distribution of images in the ROCO dataset [14] differs greatly to the ImageNet dataset; there is empirical evidence to suggest that ImageNet-trained CNNs produce state-of-the-art results when transfer learning techniques are applied to smaller datasets. [12]. Given that the ROCO dataset [14] is over 200x smaller than ImageNet, we exploit ImageNet-trained VGG16 models to benefit from this effect of transfer learning. Thus, we use the Keras [4] implementation of a VGG16 model with pre-trained ImageNet weights and extract the 14x14x512 vector from the *"block-4 max-pooling"* intermediary layer to represent the image features. We keep the weights of the CNN fixed during training to limit the number of trainable parameters; hence reducing the complexity of our model.

The image features are then passed into a recurrent network where each CUI is processed one at a time until maximum time T has passed. The unconstrained maximum number of CUIs in the training data is 72; however, we observe that we can reduce the number of time steps by 74% and retain 99% of the training data if we constrain the maximum number of CUIs to 19. Hence, to maximize efficiency, we exclude samples with CUIs greater than 19. We add "START" and "END" tokens respectfully to the beginning and end of each label set; NULL tokens are added to sets with fewer than 19 CUIs to attain a fixed length time sequence $T = 21$.

We pre-process each label set such that each CUI is represented by its unique index in the concept vocabulary $V = 5531$. To represent concept features, we train an embedding space $E \in \mathbb{R}^{V\text{x}d}$; where d is the concept vector dimensions. At the beginning of every time step, the CUI index at position t is used to retrieve its vector representation $X_t = 1\text{x}d$ from the embedding. Meanwhile, the attention mechanism takes the LSTM hidden state vector h_t to construct a probability distribution, A_t over the 14x14 spatial dimensions for the image; we multiply the feature vector by A_t and average the spatial dimensions to produce a visual context vector with $C_t = 1\text{x}512$ dimensions as per [18]. A visual sentinel learns to estimate a gating scalar $S \in [0, 1]$ from h_t to dynamically assign an attention weighting to C_t; this process is described in depth in [18]. As C_t and S_t are both produced as a function of h_t, the network learns "where" to look for discriminatory visual features and how important those visual features are in generating the CUI at time $t + 1$.

Once we multiply gating scalar S_t to context vector C_t, we concatenate the image features with CUI feature vector X_t along the last dimension to produce the $512 + d$ dimensional input to the LSTM network. A fully connected layer with relu activation reduces the D dimensional output of the LSTM network into a vector with d dimensions; residual connections to previous CUI are added by adding X_{t-1} to the output. We then multiply the resulting vector by $P = \mathbb{R}^{d\text{x}V}$ and apply a softmax function to construct a probability distribution over all the concepts in the vocabulary. At $t = 0$, the LSTM is initialized on global image feature vector G. To produce G, image features are averaged along their 14x14 spatial dimensions and pushed through a fully connected layer to create a vector with the same dimensions as the LSTM input i.e. $512 + d$.

The protocol described above represents the general framework for all 6 model variants used in this task. The learning rate $lr = 0.0001$ and batch size $n = 125$ were fixed across all variant training protocols and their performance was evaluated on the validation dataset after 20 epochs. We now describe each model variant in detail below.

3.1 Model A

Model A is the standard implementation of our model. We set the dimensions D and d to 1024 and 512 respectively. The loss at each time step is calculated as the cross entropy between the estimated probability distribution and the ground truth concept label. In addition to using cross entropy loss, we use an alpha regularizing strategy described in [10] to regulate the outputs of the attention mechanism. When no constraints are placed on an attention network, a neural network can output nonsensical weights to optimize performance on training data. To ensure the attention mechanism produces attention salient weights we first construct an attention matrix $\alpha \in \mathbb{R}^{196\text{x}T}$ from the probability distributions over the $14\text{x}14$ spatial dimensions for each time step. As described in [10] we calculate the alpha regularizing term, L_{alpha} from α as follows;

$$L_{xu} = \sum_{i}^{N}(1 - \sum_{t}^{C}\alpha_{ti})^2 \tag{1}$$

$$L_{SAL} = \frac{1}{C}\sum_{t=0}^{C}(\frac{max_i(\alpha_{ti}) - mean_i(\alpha_{ti})}{mean_i(\alpha_{ti})}) \tag{2}$$

$$L_{TD} = \frac{1}{N}\sum_{i=0}^{N}(\frac{std_t(\alpha_{ti})}{mean_t(\alpha_{ti})}) \tag{3}$$

Where L_{xu} is the alpha regularising term in [18], t represents the time axis, i represents the probability distribution axis, max_i is the maximum value, $mean_i$ is the mean value along the column axis, std_t is the standard deviation and $mean_t$ is the mean along the row axis of α_{ti}. L_{xu} ensures all image regions receive attention over the course generating each CUI, L_{SAL} ensures attention mechanism produces salient attention maps at each time step and L_{TD} ensures that the attention mechanism is not biased to any particular image region over the course of generation. The final alpha term can thus be written as;

$$L_{alpha} = \lambda_1 C_{xu} + \frac{\lambda_2}{\max(\delta, C_{SAL})} + \frac{\lambda_3}{\max(\delta, C_{TD})} \tag{4}$$

Where λ_1, λ_2 and λ_3 are hyper-parameters to scale the representation of each term. δ is used to avoid zero division and exploding gradients in the initial training steps. This loss term is then added to the total cross entropy loss and we perform standard back-propagation with ADAM optimisation [9]

During training, we implement a teacher forcing training protocol [17] where we feed the ground truth CUI to the LSTM at every time step. During inference, the ”START” token is fed into the LSTM network to get the probability distribution for the first CUI; the index with the highest probability estimate is used to generate the CUI for that time step. This process is repeated until a terminal ”END” sequence token is produced or maximum time steps T have passed.

3.2 Model B

Model B is a standard implementation of our model. The protocol is identical to A except we restrict the dimension of the CUI feature vectors to $d = 300$. This was due to concerns that an embedding size of 512 may over-fit to the training distribution.

3.3 Model C

Model C is a standard implementation of our model. The protocol is identical to A except we restrict the dimension of the concept vectors to $D = 512$. This was due to concerns that an LSTM hidden state size of 1024 may over-fit to the training distribution.

3.4 Model D

Model D seeks to address the problem of cumulative error resulting in a bias towards learning samples with longer CUI sequences. In the standard implementation of our model, the maximum error for each sample is constrained by the number of CUIs in the set. This is because cross entropy error is calculated on a per concept basis (at each time step), not a per sample basis. To ensure each sample has equal weighting in the objective function, we divide the error at every time step by the total number of CUIs for each sample and multiply the result by the maximum number of CUIs (19). This ensures that every sample has the same theoretical maximum error and that the error incurred for each incorrect concept is relative to the total number of concepts in the set. Aside from the new weighted cross entropy loss function, Model D is otherwise identical to Model A.

3.5 Model E

Model E assesses the performance of our standard implementation without any constraints on our attention mechanism. Here, we use the standard implementation described in Model A except only the cross entropy error is used to train the network. This was done to ensure the alpha regularisation strategy is appropriate for this task and not over regulating our network.

3.6 Model F

Model F assesses the performance of our standard implementation without the visual sentinel. Here, we use a similar implementation to that described in Model A; however, we remove the step of estimating the visual gating scalar S_t and allow the LSTM to be conditioned on the unscaled C_t vector. This can be interpreted equally representing features from X_t and C_t at every time step; meaning that the network no longer has the capability of dynamically assessing the importance of visual features for each CUI. This was done to ensure that the gating scalars produced by the network in Model A actually resulted in improved outcomes with regards to performance on the validation dataset.

4 Results

This section provides the results for our own internal evaluations on the validation dataset supplied for this task; these are tabulated in Table. 4. We submitted Model A for evaluation on test data as it achieved the highest F1 score, as shown in Table. 4. We decided to submit Model D as well as it achieved a competitive result with a surprisingly small average concepts per sample; we were curious to see the performance of a more conservative model on the test distribution. Model A and Model D achieved F1 scores of 0.1749349 (rank 22) and 0.1640647 (rank 27) on the test dataset.

Table 4. This table compares the quantitative performance of each of our models on the validation dataset. F1 refers to the average F1 score on the validation dataset. MIN, MAX, and MEAN respectively refer to the minimum, maximum and mean number of concepts generated for each example in the validation set. The highest F1 score is highlighted in bold font.

Model	F1	MIN	MAX	MEAN
A	**0.16**	1	16	4.3
B	0.15	1	11	4.2
C	0.13	0	14	3.7
D	0.15	0	14	0.4
E	0.12	1	9	2.5
F	0.12	1	9	2.9

5 Conclusion and Future Works

The performance of the proposed methods placed the CRADLE group 6^{th} out of 12 participating teams in the ImageCLEF 2019 medical concept detection task. The baseline architecture “Model A” achieved the highest performance. “Model B” and “Model C” did not improve the F1 score which suggests that the proposed dimensionality of hidden state and word embedding vectors in “Model A” is not resulting in over-fitting to the training distribution.

As evident by the reduced performance of “Model D”, resolving disparities in concept distributions by normalising per-sample error has an adverse effect on training. This contrary to what was hypothesized in Section 3.4. In retrospect, normalizing per-sample error in fact forms a bias towards samples with fewer concepts. This is because the “disproportionate” increase in per-sample error for longer concept sequences would occur at time steps where operations are exclusive to those longer sequences. Once the “END” token is generated for a sample, the error at latter time steps for this sample should in fact be zero. The normalization methods described in Section 3.4 would unfairly disadvantage longer sequences by reducing the relative error at each time step. Subduing error in operations common to all samples to resolve disparities in total error due to exclusive operations in longer samples is counter-productive and is likely to explain the reduced performance of “Model D”.

The reduced performance of ‘Model E” confirms that unregulated attention mechanisms result in reduced performance and that the general constraints described in Section 3.1 are capable of improving attention and overall performance. “Model F” achieved one of the lowest F1 scores, highlighting the importance of regulating the weight of visual features depending on the visual connotations of each CUI. Future work will attempt to address the challenges described in Section 2 by studying the association of CUI semantic type to visual connotation. This will be achieved by retrieving CUI meta-data from the UMLS metathesaurus [2].

6 Acknowledgements

This project has been funded by Sullivan Nicolaides Pathology and the Australian Research Council (ARC) Linkage Project [Grant number LP160101797].

References

1. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Proceedings of the AMIA Symposium. p. 17. American Medical Informatics Association (2001)
2. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research 32(suppl_1), D267–D270 (2004)
3. Browne, A.C., McCray, A.T., Srinivasan, S.: The specialist lexicon. National Library of Medicine Technical Reports pp. 18–21 (2000)
4. Chollet, F., et al.: Keras. `https://keras.io` (2015)
5. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association 23(2), 304–310 (2015)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
7. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks 18(5-6), 602–610 (2005)
8. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia

retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)

9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Maksoud, S., Wiliem, A., Zhao, K., Zhang, T., Wu, L., Lovell, B.C.: Coral8: Concurrent object regression for area localization in medical image panels. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2019)
11. Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., Khudanpur, S.: Recurrent neural network based language model. In: Eleventh annual conference of the international speech communication association (2010)
12. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1717–1724 (2014)
13. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2019 concept prediction task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland (September 09-12 2019)
14. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology objects in context (roco): A multimodal image dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, pp. 180–189. Springer (2018)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Soldaini, L., Goharian, N.: Quickumls: a fast, unsupervised approach for medical concept extraction. In: MedIR workshop, sigir (2016)
17. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural computation 1(2), 270–280 (1989)
18. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. pp. 2048–2057 (2015)