Weakly Supervised Captioning of Ultrasound Images

Mohammad Alsharid[1(✉)], Harshita Sharma[1], Lior Drukker[2], Aris T. Papageorgiou[2], and J. Alison Noble[1]

[1] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
mohammad.ali.alsharid@gmail.com

[2] Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK

Abstract. Medical image captioning models generate text to describe the semantic contents of an image, aiding the non-experts in understanding and interpretation. We propose a weakly-supervised approach to improve the performance of image captioning models on small image-text datasets by leveraging a large anatomically-labelled image classification dataset. Our method generates pseudo-captions (weak labels) for caption-less but anatomically-labelled (class-labelled) images using an encoder-decoder sequence-to-sequence model. The augmented dataset is used to train an image-captioning model in a weakly supervised learning manner. For fetal ultrasound, we demonstrate that the proposed augmentation approach outperforms the baseline on semantics and syntax-based metrics, with nearly twice as much improvement in value on *BLEU-1* and *ROUGE-L*. Moreover, we observe that superior models are trained with the proposed data augmentation, when compared with the existing regularization techniques. This work allows seamless automatic annotation of images that lack human-prepared descriptive captions for training image-captioning models. Using pseudo-captions in the training data is particularly useful for medical image captioning when significant time and effort of medical experts is required to obtain real image captions.

Keywords: Image captioning · Fetal ultrasound · Data augmentation

1 Introduction

Image captioning generates a textual description of the spatial information present in an image [9]. There has been growing interest recently in medical image captioning [24], including for ultrasound (US) imaging [8,35,36].

Consider the future where wearable medical technology and seamless sensors provide consumers with basic information about their health [32]. In this future, portable US probes would connect to a user's smart mobile device to capture medical selfies [32] and real-time automated captioning would provide a textual

G. Yang et al. (Eds.): MIUA 2022, LNCS 13413, pp. 187–198, 2022.
https://doi.org/10.1007/978-3-031-12053-4_14

description of the medical selfie content. While the US probes for this realisation exist today, the technical capability of generating such captions needs to be worked on. This paper takes a preliminary step towards actualizing it.

Data: image-caption pairs and caption-less images with anatomical class labels
Result: caption-less images annotated with pseudo-captions
while *caption-less image* i_{cl} *exists* **do**
  calculate similarity between i_{cl} and every image with real caption label;
  retrieve caption of most similar image;
  perform parts-of-speech tagging on caption;
  extract words that have been tagged as nouns;
  input label and nouns to seq2seq model to generate pseudo-caption;
  annotate i_{cl} with pseudo-caption;
end
train captioning model

Algorithm 1: The entire pseudo-caption preparation pipeline.

We are interested in understanding the principles of designing US image and video captioning algorithms. Our work is application-agnostic in that the generated captions could be of use to different users, such as amateur observers of ultrasound images. A clinical motivation for automatic US image and video captioning stems from US images being difficult for non-experts to interpret. However, image-to-text translation may encourage greater US use for simple tasks if users do not have domain-specific knowledge of US, for instance, to communicate simple diagnostic findings in text format rather than to expect users to directly interpret the content of an US image. Until recently, methods introduced for automated image captioning typically relied on purely supervised learning requiring large-scale datasets of image-caption pairs for training [11,29,30,33].

A challenge more specific to medical image captioning is the domain-specific knowledge required to manually prepare image-caption pairs, which makes dataset preparation resource intensive. In contrast, preparing medical imaging datasets for image classification problems is more common and easier to perform. This raises an interesting question: *can we circumvent limitations of the available data for medical image captioning to train captioning models without solely relying on manually prepared image-caption pairs and by leveraging image classification datasets?* In this work, we investigate the potential of using an existing class-labelled classification dataset to augment the small number of available image-caption pairs and train an image captioning model through weakly supervised learning.

Related Works. In the literature, a few studies have addressed the problem of limited image captioning datasets. An image captioning approach that does not require readily available image-caption pairs is introduced in [16]. Another

approach identifies concepts in images and tries to find captions that are semantically most similar to those identified concepts [20]. Other studies attempt to first train models on image-caption pairs in one domain, and then transfer the learned knowledge into a second domain that is lacking in paired data [12,37]. Some works identify the concepts exhibited in an image, for example by object identification, and then proceed to build sentence templates around the identified objects in the image [18,19,22,23,31].

However, our work is different from these studies as we do not use a visual concept or object detector. We also do not use the annotated data to train one. A visual concept can refer to an object present in an image and a property pertaining to it (e.g. 'red ball'). In [14], the authors use K-nearest neighbours to identify which image in the dataset is most similar to a target image in a text retrieval-based captioning approach. However, the downside of using text retrieval is that all possible captions are retrieved from the training dataset of image-caption pairs. With text generation, however, novel unseen captions that do not exist in the training dataset can be generated, as proposed in our work. Data augmentation can be considered as a model regularization technique [10]. Some regularization techniques used in NLP include word dropout (WD), where several words in a sequence are randomly chosen to be dropped [17]. Another technique is SwitchOut (SO) which randomly swaps some words in a sequence with other words that exist in the training vocabulary [34]. We compare the proposed augmentation approach with these regularization techniques.

Another different way to make up for the lack of data, as was done in [6], involves using a curriculum learning based approach, where a dual curriculum is used in ranking data points according to their entropy in the image and text modalities and summing their contributions equally to create a new overall ranking. In [7], a natural extension of [6] was investigated where a linear combination of the complexity metrics of a single multi-modal data sample was used. This means that rather than assuming that both metrics contribute equally to the arrangement and ordering of batches in every epoch, one of the complexity metrics is more influential than the other in a given epoch.

Motivation and Contribution. We automatically prepare pseudo-captions for caption-less images by leveraging existing image-caption pairs and anatomically labelled images to use in training the fetal US image captioning model via weakly supervised learning. In the proposed method, nouns are identified in the retrieved caption and then along with the anatomical class label of the target image are fed into a model that generates a sentence given certain keywords. In other work, these keywords would be real concepts that come with the data or could be obtained through object detection. However, in this work, we do not have real concepts associated with our data, and so, we rely on weak labels by generating pseudo-captions from extracted nouns that serve in lieu of concepts acquired through object detection, thereby circumventing two requirements: (1) having to train a visual concept detector, and (2) curating and annotating a dataset to train a visual concept detector on.

Furthermore, the approach makes it possible to introduce an aspect of potential novel caption generation that would be missing if we solely relied on text retrieval, allowing for greater diversity in obtained captions for fetal ultrasound images. Our approach consists of four steps: (1) text retrieval, (2) noun extraction, (3) pseudo-caption creation from anatomical labels and extracted nouns, and (4) caption generation through an image captioning framework.

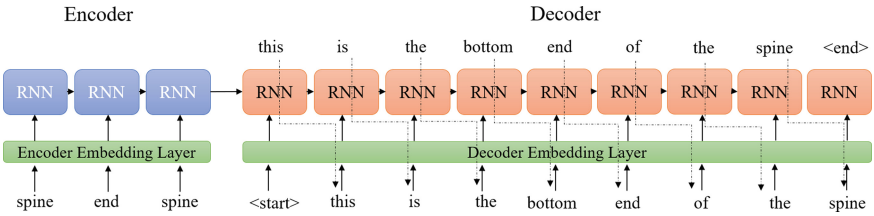

Fig. 1. The sequence-to-sequence model architecture translates the sequence of the anatomical label and the nouns into a pseudo-caption. In the input sequence, the first 'spine' is the label, and 'end' and the second 'spine' are the extracted nouns. Both the encoder and decoder consist of 100 LSTM units. The word embedding size is 300.

2 Methods

Before discussing each step of the proposed method, we briefly describe the available images, captions, and annotations. The images are video frames extracted from second-trimester fetal US scan videos and each image has an image-level anatomical class label associated with the main visible anatomical structure in the image. The associated label is 'abdomen', 'head', 'heart', or 'spine'. These labels are made available through the work of [26]. The images are split into two categories.

The first category consists of images from scan videos with textual descriptions from the transcribed speech of sonographers. These form the training dataset of real image-caption pairs (caption-labelled). The process of annotating and pre-processing this caption-labelled dataset is a cumbersome process. The second category consists of images from different scan videos that form the dataset of the caption-less images (caption-free) or the anatomically labelled image classification dataset. A high-level description of the method is given in Algorithm 1. The subsections below describe each part of the pseudo-caption creation pipeline.

Text Retrieval. We calculate the cosine similarity between caption-less images and every image in the training dataset on the feature domain. Features were extracted from a VGG-16 CNN [27]. The cosine similarity is defined as:

$$similarity_{cos} = \boldsymbol{x_n} \cdot \boldsymbol{t} / |\boldsymbol{x_n}||\boldsymbol{t}| \tag{1}$$

where $\boldsymbol{x_n}$ is the image feature vector of a data sample from the real image-caption pair dataset and $\boldsymbol{t}$ is the image feature vector of the caption-less image. We retrieve the caption of the image with the highest cosine similarity.

Noun Extraction. After retrieving the caption of the image most similar to the caption-less image, we extract its nouns through the TextBlob Python library [4]. When provided with a string, a list of tuples is returned. Each tuple consists of a word in the original string and its parts-of-speech tag. From the returned list, we create a sublist consisting solely of the nouns that exist in the original string, and we use the parts-of-speech tags to identify those nouns. We hypothesise that these nouns adequately represent the inherent concepts associated with the image feature vector in question and, therefore, are important to use when creating pseudo-captions. Our assumption is based on previous work such as [28] which considers centering the sentence syntax around nouns and [5] which considers nouns to be important in determining the context. This phenomenon coincides with what we have observed in the sonographer recordings, in which most of the verbs (such as 'looking', 'seeing', and 'measuring') are often less correlated to the clinical interpretation of the images. The process by which the input to the pseudo-caption is prepared is demonstrated by Eqs. 2 and 3.

$$C_w = \begin{cases} \{TB(w)\} & if w \in N \\ \{\} & otherwise \end{cases} \tag{2}$$

$$X = \bigcup_{w}^{W} (C_w) \cup \{\alpha\} \tag{3}$$

where C represents a concept associated with a word w, $TB(w)$ represents the parts-of-speech tag of w, N represents the set of tags associated with nouns (including 'NN', 'NNS', 'NNP', 'NNPS'), X represents the extracted nouns and anatomical label α of a data sample, W are all the words in the caption of a single data sample.

Pseudo-Caption Creation. We train an encoder-decoder sequence-to-sequence model to transform a sequence of extracted nouns and the corresponding anatomical label of the caption-less image to a pseudo-caption. This is similar to the con2sen model [16] which generates a pseudo-caption from the objects detected in an image of interest. To train the model (shown in Fig. 1), we perform noun extraction on the real captions. The extracted nouns and the anatomical labels serve as the input while the real captions are the target outputs. Training and deploying the model to create pseudo-captions for caption-less images takes 12 min on a machine with an NVIDIA GeForce GTX 1080.

Caption Generation. Captions are generated for fetal US frames through an image captioning model architecture, as shown in Fig. 2. The image captioning

model can be described as a late merge captioning model [6,8,29,30], where image and textual information are merged towards the end of the model to generate the next word in the sequence. The text-focused branch in Fig. 2 is in the left half of the figure. It shows that each word in the input sequence is given a token before being passed through an embedding layer. The embedding layers uses weights from a Word2vec model that has been pretrained on the GoogleNews corpus [1]. The sequence is then passed through a recurrent neural network. The right branch of the captioning model consists of a VGG16 convolutional neural network (CNN) that has been fine-tuned on fetal US images of the same gestational age. The CNN is followed by two fully connected layers. From both branches, a flat feature vector is obtained. The vectors are concatenated together to predict the next word. The process is repeated until the maximum possible length is reached or a special end token ('<end>') is generated. The framework also includes an image feature vector classifier, identical to the right branch of the captioning model, that classifies an image to one of four possible classes, 'abdomen', 'head', 'heart', and 'spine'. In the framework, four variants of the captioning model are trained, one for each of the four structures. By having a separate captioning model for each anatomical structure, generated captions are more likely to be relevant to the ultrasound image. During inference, once

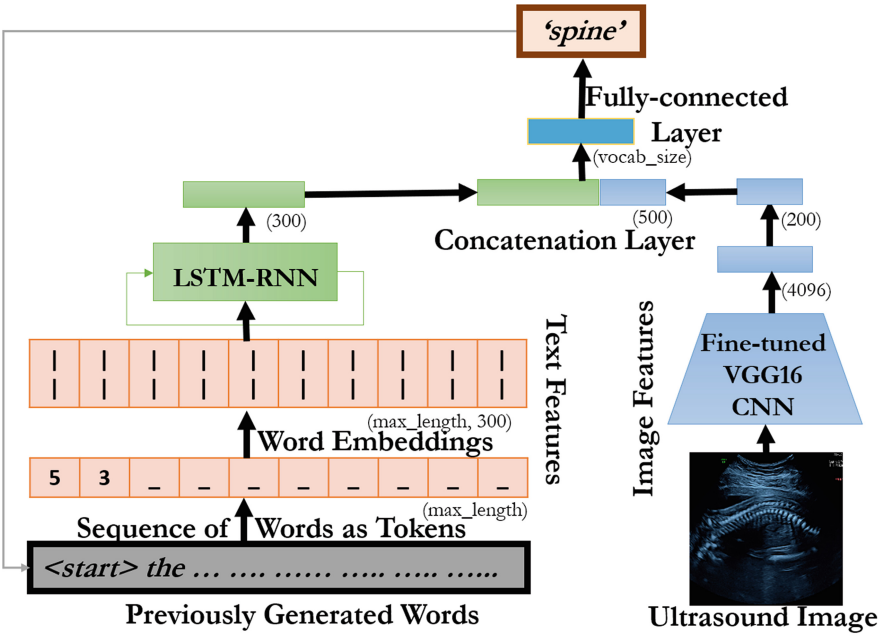

Fig. 2. The image captioning model. Max length represents the maximum number of words a caption of this anatomical structure could consist of. The LSTM-RNN consists of 300 units.

the image feature vector classifier classifies an image, the appropriate captioning model is initiated. Also, at inference, previously generated words are used as input to generate the next word in the sequence. The models are trained with a categorical cross entropy loss. The Adam optimization algorithm [21] is used with a learning rate of 0.001. For the model trained with our approach, we used the real image-caption pairs augmented by the image-pseudo-caption pairs.

3 Experiments

Datasets and Data Preparation. The used data came from the PULSE study [15]. As part of the study, videos of full-length fetal US scans were acquired along with their accompanying audio recordings. For our paper, only part of the data was available. Breakdown of the data used is shown in Table 1. In total, 18 videos are used in this paper. The mean video length is 32 min with standard deviation of 14 min. The shortest video was 8 min long, while the longest was 56 min long. Manually annotating these videos with captions is a time-consuming process. The real image-caption pairs were prepared by transcribing the audio recordings of sonographers from 10 full-length scan videos following the approach of [8]. In this way, we have obtained thousands of image-caption pairs that are rich and representative of the second-trimester fetal ultrasound scan to train the image captioning models (Table 1). As all the scans follow a countrywide scanning protocol, the semantic contents (anatomical categories) are consistent across the different scans, hence, we were able to extract multiple diverse examples of each category from the full-length scan videos (number of samples of each category shown in Table 1).

Table 1. Breakdown of the data used in this work

Dataset	Scans	Pairs	Abdomen	Head	Heart	Spine
Train (real)	7	23,558	3,085	11,145	6,573	2,755
Val (real)	2	17,471	3,003	8,479	4,482	507
Test (real)	1	14,601	2,032	3,710	7,078	1,543
Train (pseudo)	8	2,721	184	614	1,297	553
Entire dataset	18	58,351	8,304	23,948	19,430	5,358

The images from the image-pseudo-caption pairs cover a higher variance having included a step size of 16 between each sampled frame from the scan video. As part of standard pre-processing practice, the images were cropped to remove the user interface and then resized to 224×224 pixels. The text had its punctuation removed, and all letters were made lowercase. The longest video contained the greatest number of samples and, so, was held out as the test set.

Performance Evaluation. The image-captioning evaluation metrics can be divided into two categories: syntax-focused metrics and semantics-focused metrics. The syntax-focused metrics include *BLEU-1 (B1)* [25], *ROUGE-L (RL)* [13] and GrammarBot (*GB*) [3]. $B1$ and RL look at the degree of overlap between the words of the ground truth and the words of the generated caption. However, they do not take into consideration whether the overlapping words hold any semantic value with respect to the image.

The semantics-focused metrics include the Anatomical Relevance Score (*ARS*) [8] and $F1$ score. *ARS* rewards the model for generating words that are not necessarily synonymous to the words in the ground truth caption, but are anatomically relevant by looking at the degree of word overlap between the words in the generated caption and the vocabulary associated with the anatomical structure depicted in the image of interest. *GB* simply measures the quality of the grammar in the generated captions.

4 Results and Discussion

We trained multiple models to compare our approach (including image-pseudo-caption pairs in the training dataset) with a baseline (no text augmentation or regularization) and other regularization techniques (WD [17] and SO [34]). The quantitative results are shown in Table 2. When comparing between the performance of a model trained on the original dataset and a model trained on the augmented dataset, we can immediately see an increase in *B1*, *RL*, and *ARS* metrics when incorporating the image-pseudo-caption pairs with the exception of $F1$ score where models trained both with and without pseudo-captions obtained the high score 0.97. However, from the *ARS* score, we notice that the model that was trained with pseudo-captions score higher. A higher *ARS* score translates to the model producing higher softmax probabilities for terms that are relevant to an anatomical structure (for e.g. 'brain' and 'nuchal' for the head), so with a high ARS, we can say more confidently that the model generates words that are relevant. With regards to the *B1* score, the models trained with pseudo-captions obtained a score 0.30 which falls within the range that [2] would describe as being 'understandable and good'. On the other hand, the set of models trained without pseudo-captions obtained a *B1* score of 0.13. A score of 0.13 falls within the range that [2] would describe as 'hard to get the gist (of)'. So, our method generates captions with a better sentence structure.

We can see this behavior exemplified in the generated captions in Fig. 3 where qualitative results for four examples are shown. With more data to learn from, a model is more likely to learn an understandable sentence structure, and it is pleasing to see that the pseudo-captions although not provided directly by a sonographer can still help a model to score higher on the metrics, justifying the inclusion of pesudo-captions in the training of a captioning model. All the syntax-focused metrics are better with our proposed approach except *GB*; however, that can be explained by the fact that WD drops words from the sequences

Table 2. Quantitative results comparing our proposed augmentation approach with other regularization techniques

Methods	Syntax-focused			Semantics-focused	
	B1	*RL*	*GB*↓	*ARS*	*F1*
No pseudo-captions (baseline)	0.17	0.24	1.26	0.39	**0.97**
Word dropout token level	0.01	0.03	**0.52**	0.05	0.34
Word dropout vector level [17]	0.07	0.17	0.84	0.03	0.03
SwitchOut [34]	0.11	0.20	2.55	0.08	0.85
Synonym swapping	0.11	0.20	2.62	0.09	0.73
With pseudo-captions (ours)	**0.30**	**0.42**	1.14	**0.77**	**0.97**

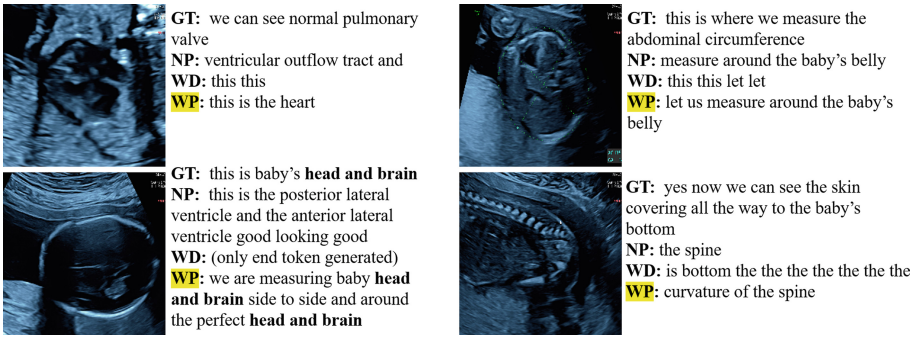

Fig. 3. Qualitative results for different images. GT stands for 'Ground Truth' as spoken by a sonographer. 'NP' stands for model trained with 'No Pseudo-captions'. 'WD' stands for model regularized with 'Word Dropout'. 'WP' stands for model trained 'With Pseudo-captions' (our proposed method).

that make up the training samples, and so, the model ends up learning to generate shorter sentences. With shorter sentences, the number of grammatical mistakes decreases. WD and SO have notably lower scores on the semantics-focused metrics. This phenomenon can be explained by the fact that, with WD and SO, anatomically relevant words could be dropped or switched out during training. Even with a WD rate of only 0.2, a caption of, say, five words will lose one word, and that could be the word that refers to the anatomical content. There is a high chance of losing semantically meaningful words in shorter captions when using WD and similar techniques.

5 Conclusion and Future Work

In this paper, we describe a novel approach to create pseudo-captions for fetal US images that lack any captions by leveraging an existing smaller image captioning dataset and an image classification dataset. The practical usefulness of

models trained with the extra pseudo-captions allows for better interpretation and relaying of information to laypersons who may be observing an ultrasound scan. We show that these image-pseudo-caption pairs can improve the performance of weakly learnt image captioning models for the fetal US image captioning task. In the future, we will investigate the applicability of the approach for other medical imaging modalities.

5.1 Future Work

One of the future tasks is to use pseudo-captions with other tasks, such as cross-modal retrieval. We intend to work on a transformer-based retrieval of ultrasound images. The architecture would consist of a vision transformer and a BERT encoder to extract an image feature representation and text feature representation. The pooled output from the BERT encoder captures the overall information of the input sequence. To enable retrieval, the cross-modal model would be trained with a ranking loss with the aim of minimizing the distance between the two feature representations of a positive pair and maximizing the distance between the two feature representations of a negative pair in the feature space.

References

1. Google code archive (2018). https://code.google.com/archive/p/word2vec/
2. Evaluating models — automl translation documentation (2020). https://cloud.google.com/translate/automl/docs/evaluate
3. Grammarbot (2020). https://www.grammarbot.io/
4. Textblob (2020). https://textblob.readthedocs.io/en/dev/
5. Context analysis in NLP: why it's valuable and how it's done (2021). https://www.lexalytics.com/lexablog/context-analysis-nlp
6. Alsharid, M., El-Bouri, R., Sharma, H., Drukker, L., Papageorghiou, A.T., Noble, J.A.: A curriculum learning based approach to captioning ultrasound images. In: Hu, Y., et al. (eds.) ASMUS/PIPPI -2020. LNCS, vol. 12437, pp. 75–84. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60334-2_8
7. Alsharid, M., El-Bouri, R., Sharma, H., Drukker, L., Papageorghiou, A.T., Noble, J.A.: A course-focused dual curriculum for image captioning. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 716–720. IEEE (2021)
8. Alsharid, M., Sharma, H., Drukker, L., Chatelain, P., Papageorghiou, A.T., Noble, J.A.: Captioning ultrasound images automatically. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 338–346. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_37
9. Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., et al.: Automatic description generation from images: a survey of models, datasets, and evaluation measures. J. Artif. Intell. Res. **55**, 409–442 (2016)
10. Burkov, A.: The Hundred-Page Machine Learning Book, pp. 100–101. Andriy Burkov (2019)
11. Chen, L., et al.: SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5659–5667 (2017)

12. Chen, T.H., Liao, Y.H., Chuang, C.Y., Hsu, W.T., Fu, J., Sun, M.: Show, adapt and tell: adversarial training of cross-domain image captioner. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 521–530 (2017)
13. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain, pp. 56–60 (2004)
14. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., et al.: Language models for image captioning: the quirks and what works. arXiv preprint arXiv:1505.01809 (2015)
15. Drukker, L., et al.: Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. Sci. Rep. **11**(1), 1–12 (2021)
16. Feng, Y., Ma, L., Liu, W., Luo, J.: Unsupervised image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4125–4134 (2019)
17. Gal, Y., Ghahramani, Z.: A theoretically grounded application of dropout in recurrent neural networks. arXiv preprint arXiv:1512.05287 (2015)
18. Guadarrama, S., et al.: YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2712–2719 (2013)
19. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2012–2019. IEEE (2009)
20. Hendricks, L.A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., Darrell, T.: Deep compositional captioning: describing novel object categories without paired training data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–10 (2016)
21. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Krishnamoorthy, N., Malkarnenkar, G., Mooney, R., Saenko, K., Guadarrama, S.: Generating natural-language video descriptions using text-mined knowledge. In: Proceedings of the Workshop on Vision and Natural Language Processing, pp. 10–19 (2013)
23. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., et al.: Babytalk: understanding and generating simple image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2891–2903 (2013)
24. Lyndon, D., Kumar, A., Kim, J.: Neural captioning for the ImageCLEF 2017 medical image challenges. In: CLEF (Working Notes) (2017)
25. 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)
26. Sharma, H., Drukker, L., Chatelain, P., Droste, R., Papageorghiou, A.T., Noble, J.A.: Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos. Med. Image Anal. **69**, 101973 (2021)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Stuart, L.M., Taylor, J.M., Raskin, V.: The importance of nouns in text processing. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 35 (2013)

29. Tanti, M., Gatt, A., Camilleri, K.: What is the role of recurrent neural networks (RNNs) in an image caption generator? arXiv preprint arXiv:1708.02043 (2017)
30. Tanti, M., Gatt, A., Camilleri, K.P.: Where to put the image in an image caption generator. Nat. Lang. Eng. **24**(3), 467–489 (2018)
31. Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., Mooney, R.: Integrating language and vision to generate natural language descriptions of videos in the wild. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1218–1227 (2014)
32. Topol, E.J.: A decade of digital medicine innovation. Sci. Transl. Med. **11**(498), eaaw7610 (2019)
33. 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)
34. Wang, X., Pham, H., Dai, Z., Neubig, G.: SwitchOut: an efficient data augmentation algorithm for neural machine translation. arXiv preprint arXiv:1808.07512 (2018)
35. Zeng, X.H., Liu, B.G., Zhou, M.: Understanding and generating ultrasound image description. J. Comput. Sci. Technol. **33**(5), 1086–1100 (2018)
36. Zeng, X., Wen, L., Liu, B., Qi, X.: Deep learning for ultrasound image caption generation based on object detection. Neurocomputing **392**, 132–141 (2019)
37. Zhao, W., et al.: Dual learning for cross-domain image captioning. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 29–38 (2017)