Captioning Ultrasound Images Automatically

Mohammad Alsharid[1], Harshita Sharma[1], Lior Drukker[2], Pierre Chatelain[1], Aris T. Papageorghiou[2], and J. Alison Noble[1]

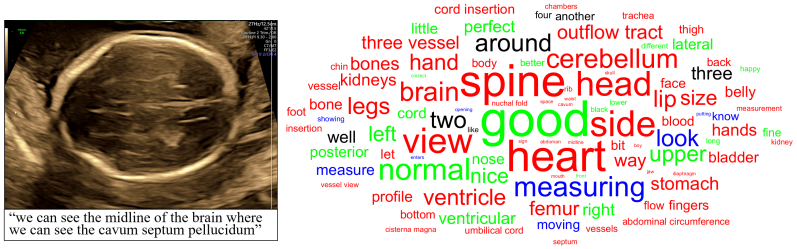
[1] Institute of Biomedical Engineering, University of Oxford, UK
[2] Nuffield Dept. of Women's & Reproductive Health, University of Oxford, UK
mohammad.alsharid@eng.ox.ac.uk

Abstract. We describe an automatic natural language processing (NLP)-based image captioning method to describe fetal ultrasound video content by modelling the vocabulary commonly used by sonographers and sonologists. The generated captions are similar to the words spoken by a sonographer when describing the scan experience in terms of visual content and performed scanning actions. Using full-length second-trimester fetal ultrasound videos and text derived from accompanying expert voice-over audio recordings, we train deep learning models consisting of convolutional neural networks and recurrent neural networks in merged configurations to generate captions for ultrasound video frames. We evaluate different model architectures using established general metrics (*BLEU*, *ROUGE-L*) and application-specific metrics. Results show that the proposed models can learn joint representations of image and text to generate relevant and descriptive captions for anatomies, such as the spine, the abdomen, the heart, and the head, in clinical fetal ultrasound scans.

Keywords: Image Description, Image Captioning, Deep Learning, Natural Language Processing, Recurrent Neural Networks, Fetal Ultrasound

1 Introduction

Automatic image captioning combines computer vision with natural language processing to generate a textual statement, called a caption, to represent image content. Image captioning has been widely explored for natural images with benchmark datasets [1], however, most established image-captioning datasets do not include medical images. Preparing medical image captioning benchmarks is challenging for two reasons: (a) describing medical images with specific terminology requires expert knowledge of medical professionals; and (b) the sensitive nature of medical images prevents wide-scale annotation, for instance, using crowd-sourcing services. Therefore, automatic image captioning has not been widely studied on ultrasound images before, the challenge being enhanced by the lack of readily available large datasets of ultrasound images with captions. To the best of our knowledge, this is the first attempt to perform automatic image captioning on fetal ultrasound video frames, using sonographer spoken words to describe their scanning experience.



(a) (b)

Fig. 1: (a) Example of a fetal ultrasound image with sonographer description. (b) Word cloud of most frequently occurring words in sonographer vocabulary. Red, green, and blue represent nouns, adjectives, and verbs, respectively. The size of a word in the word cloud is proportional to its frequency of use.

As part of routine care, pregnant women are offered a detailed fetal anomaly ultrasound scan at approximately 20 weeks of gestation to identify any fetal malformations. While developing an ultrasound image-captioning model, we can analyse the vocabulary used by sonographers to describe the scans, reflecting their experience during the scan process in terms of visual content and performed scanning actions. The aim of our work is to learn joint image-text representations to describe ultrasound images with rich vocabulary consisting of nouns, verbs, and adjectives. The current work is application agnostic, but a potential application of the work is its use as an educational tool that communicates descriptions of anatomical views of interest to the subjects and sonography trainees. An example of an image and its caption is shown in Fig. 1a. The word cloud in Fig. 1b shows the most common spoken sonographer words used to describe fetal ultrasound scans in our work.

Related Work. There are currently two established ways to perform image captioning [1,20]: (a) text retrieval where descriptions are stored beforehand and retrieved using scores between stored and queried images [15]; and (b) text generation where novel text descriptions are generated. The latter is achieved using top-down or bottom-up approaches [23]. In the top-down approach, an image is described by translating visual representations to text, and in the bottom-up approach, constituent objects and concepts in an image are described with words that are then combined into sentences using language models [4]. In both cases, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are built from the images and text, respectively [20,22]. We are aware of only two previous ultrasound image captioning works [12,24]. In [12], captions are generated for the ImageCLEF dataset including other radiology images, which also uses a top down deep-learning based text generation approach. In our work, a reduced complexity is achieved using a merged configuration in which image feature vectors are not included as part of the input sequence to the recurrent network. In [24], the image captioning task is performed on adult abdominal

ultrasound with a focus on diseases of the kidney and the gallbladder, where a structure and an associated disease are classified before generating a description with an RNN trained specifically on words of that structure. In contrast, we propose models where representations are jointly learned in a single step. Both [12] and [24] use text reports as a raw source of data. We use sonographer voice-over recordings to describe the videos in real-time, thereby providing a richer description of the spatio-temporal video content.

2 Methods

Data Acquisition and Processing. Full-length routine fetal anomaly ultrasound scan videos acquired by an expert sonographer were available for the research from the PULSE study [3]. We had the sonographer retrospectively record voice-overs in English for five anonymised videos with a mean duration of 37 minutes (range: 20-56 minutes). A total of 160 minutes of audiovisual content was recorded. From the full-length videos, freeze frames were automatically detected. The sonographers freeze a frame when they find a suitable view of interest for diagnostic examination, which are the anatomical standard planes. The display frame was automatically cropped to include only the anatomical view. The speech recordings were pre-processed for anonymisation and then transcribed using Google Cloud Speech (GCS) API [6]. GCS is designed for natural language, but the recordings contain additional medical vocabulary. The transcriptions contained a few errors which were corrected by manual post-processing. ELAN, a multimedia annotator for audiovisual content, was used to synchronize video contents with generated transcriptions and to correct erroneous text [19]. After the transcribed words were manually checked, grouped, and synced, a file containing the captions with start and end times was produced to automatically align video frames with captions. Fig. 2 shows the process of creating image-caption pairs. The raw text was cleaned by removing punctuation, replacing numeric characters with their word equivalents, and removing stall words (e.g. 'so yeah', 'well'). Special tokens denoted a caption's start and end. The resulting caption length varied between 1–22 words, with a vocabulary of 158 unique words and distribution of adjectives, determiners, nouns, and verbs is 12.7%, 22.2%, 28.0%, and 16.0%, respectively. The remaining 21.1% are prepositions, pronouns, adverbs, and other parts-of-speech. Hence, the combined dataset was composed of real-world fetal anomaly ultrasound video freeze frames and their associated captions.

Model Architecture. Image captioning often involves a CNN to encode image information followed by an RNN as a decoder to generate text [22]. However, to reduce computational complexity, an RNN was used solely as the textual feature extractor and the encoded image information from a CNN was combined with the textual features in merged configurations [20,21]. The model diagram is shown in Fig. 3. One branch of the model is a CNN based on the VGGNet16 [18] architecture, pre-trained on the ImageNet dataset and fine-tuned on fetal ultrasound standard planes of the abdomen, face, brain, femur, heart,

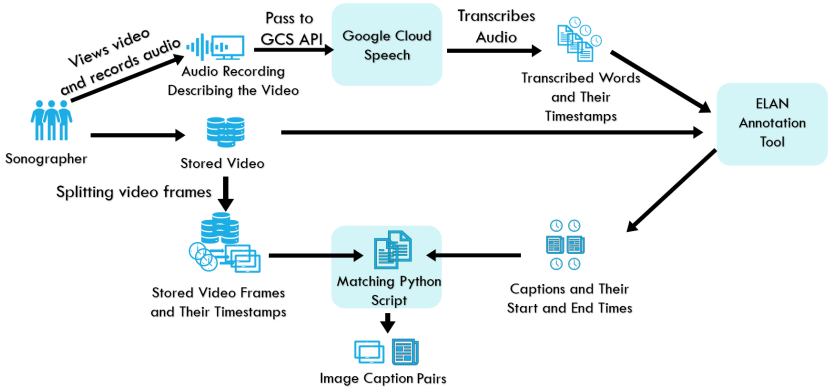


Fig. 2: Data acquisition and processing pipeline

spine, and placenta. The other branch represents the text encoding part of the model, including an embedding layer, which embeds the words in the sequence into a vector, followed by an RNN.

Features are extracted from the ultrasound video frames using the fine-tuned CNN. A textual caption is encoded by an embedding layer followed by a recurrent layer. The branches are merged, followed by a fully connected and decision-making layer. The model configurations generate the next word in a caption at every step as the probability distribution over the words in the vocabulary. We comparatively evaluated different embeddings, namely, word2vec embedding trained on the Google News corpus [7], GloVe embedding trained on Wikipedia-2014 corpus [17], and a plain random initialization. Word2vec is a shallow neural network trained to predict the context around a given word in a skip-gram model [14]. GloVe incorporates word co-occurrence probabilities with the idea that words occurring together often enough are likely to hold underlying semantic meaning. The embedded word vectors are learnt by an RNN consisting of a Long Short-Term Memory (LSTM) unit [9] or a Gated Recurrent Unit (GRU) [2]. GRUs have less trainable parameters than LSTMs and require fewer operations, which makes GRUs more efficient to train, scaling down well to smaller datasets. The two branches produce tensors of different lengths (200 and 300, respectively) that are joined together by merging. We compare two merge methods, namely, concatenation and addition. In concatenation configurations, text and image features vectors of unequal length are combined to deliberately force the model towards relying more on the text branch when generating the next word in a sequence to have a textually well-structured generated caption. However, output vectors of an equal length of 300 are used in addition configurations.

Training Process. Sixty-five percent of the total data was used for training and thirty-five percent for testing. For training the deep learning models, we excluded captions that do not describe one of the four anatomical structures of highest interest, *i.e.*, head, heart, spine, and abdomen. These anatomical

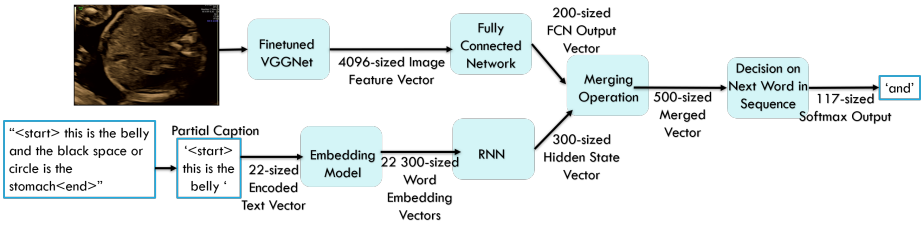


Fig. 3: Image captioning model (concatenation configurations)

classes were selected due to having the most representation in the collected dataset compared to other anatomical classes, and they form 40%, 22%, 20%, and 18% of the data, respectively. From caption pre-processing, vocabularies of each of the four anatomical classes of interest were obtained. Lexical diversity scores, specifically MTLD [13], to measure word variety in each vocabulary were obtained as 21.2, 20.3, 17.9, and 21.8, respectively. To address class imbalance, an equal number of unique captions were included for each anatomical class, namely, abdomen, head, heart, and spine. In addition to class imbalance, caption imbalance is observed as some captions correspond to more than one video frame because sonographers spend a different amount of time looking at different fetal structures.

The training set consisted of 2,240 image-caption pairs and the validation set consisted of 560 image-caption pairs. The images were resized to 224×224 pixels. Each image in the dataset was augmented twice; first by rotating by an angle between -30°and 30°, and second by horizontally flipping the image. Pre-trained VGGNet16 was first fine-tuned on ultrasound images. During training of image captioning models, 'teacher forcing' was applied where the ground truth sequences of increasing length were used at every step rather than the sequence of the words generated by the model at previous steps [5]. The model was called in a recursive fashion with the sequence of generated words so far being iteratively fed into the model at every time-step, along with the corresponding image. This process continued until the model generated a special end token or the maximum caption length was reached. Adam optimization [10] and categorical cross-entropy loss were applied during training. Early stopping was used to stop training when validation loss did not improve for more than five epochs. Dropout (rate between 0.4 and 0.5) was used to reduce overfitting. During inference, the model relied on its previously generated words to generate the next word.

Evaluation Metrics. Different model configurations were compared using the established general metrics *BLEU* (Bilingual Evaluation Understudy) and *ROUGE-L* (Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence), a grammar score *GB* (GrammarBot) [8], a classification metric *Class. F1* (F1 score), and an anatomical description metric *ARS* (Anatomical Relevance Score). Objective metrics *BLEU* [16] and *ROUGE-L* [11] are calculated between the ground truth captions and the generated captions. These two metrics are commonly used when evaluating image captioning mod-

els but may lead to lower values when the pair of captions do not show exact matches. Hence, for our caption generation task, grammar-based, classification-based, description-based, and subjective metrics were additionally evaluated. To evaluate captions grammatically, the average number of grammatical mistakes in a generated caption was calculated. Classification F1 scores were calculated by determining the caption class as the class of the vocabulary that has the highest overlap with the predicted caption. We devised an anatomical relevance score (ARS) by matching words in a generated caption with the terminology of the anatomical class of interest. For example, an image of an abdomen may have a ground truth caption about the ribs, but if the generated caption describes the stomach, it is not an erroneous caption. ARS is calculated using Equations 1, 2 and 3

$$CS_k = \left(\sum_{i=1}^{L(W^c)} \mathbf{1}_{V_k}(w_i^c) \right)^{-1} \sum_{i=1}^{L(W^c)} \mathbf{1}_{V_k}(w_i^c) p_i \tag{1}$$

$$SS_c = \begin{cases} \max\limits_{k \in K} CS_k & \text{if} \arg\max\limits_{k \in K}(CS_k) = GT_c \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$ARS = \frac{1}{C} \sum_{i=1}^{C} SS_c \tag{3}$$

where CS_k is a score that a caption has in relation to the anatomical class k, K is the set of four anatomical classes, V_k is the vocabulary set of class k, L is the length of a caption W^c consisting of words w_i^c with softmax probabilities p_i, $\mathbf{1}_V(\cdot)$ is an indicator function which returns 1 if w_i is in V and 0 otherwise, SS_c is a score that only considers CS_k if it has the ground truth anatomical class GT_c, and C is the total number of captions in the set.

3 Results

Quantitative Evaluation. Table 1 presents quantitative evaluation results for different model configurations. An overall score is obtained by calculating the mean of the scores (GB was normalised and inverted). The overall best performing model was the Fine-tunable-Word2vec-LSTM-Concatenation configuration, which is used to demonstrate anatomical evaluation in Table 2 and Fig. 4. For a subjective measure, Likert Scale based evaluations are performed where a medical professional was asked to give a score of 0 ('No'), 1 ('Neutral'), or 2 ('Yes') in response to the following statements about a generated caption, namely it: (1) accurately describes the image; (2) has no incorrect information; (3) is grammatically correct; (4) is relevant for this image. For each caption, the responses were averaged. These scores are reported in Table 2 as LSS (Likert Scale Scores). Knowing the original image class and resulting caption classes, we plot the confusion matrix for the best performing configuration in Fig. 4.

Table 1: Evaluation Results of Model Configurations

Word Embedding	RNN	Merge Mode	*BLEU-4*	*ROUGE-L*	*GB*↓	*Class. F1*	*ARS*	Overall
Fine-Tunable GloVe	LSTM	Concatenation	0.066	0.536	1.091	0.809	0.680	0.385
		Addition	0.081	0.580	0.9	0.948	0.686	0.397
	GRU	Concatenation	0.081	0.585	**0.889**	0.502	0.455	0.261
		Addition	0.094	0.561	0.923	0.529	0.449	0.268
Fine-Tunable Word2vec	LSTM	Concatenation	**0.105**	**0.594**	1.214	0.970	0.536	**0.427**
		Addition	0.045	0.546	0.929	0.679	0.506	0.297
	GRU	Concatenation	0.080	0.523	1.200	0.764	0.594	0.376
		Addition	0.086	0.539	1.077	0.609	0.476	0.307
Pretrained Word2vec	LSTM	Concatenation	0.085	0.574	1.200	0.921	0.567	0.413
		Addition	0.063	0.529	1.267	0.641	0.537	0.348
	GRU	Concatenation	0.066	0.530	1.100	0.768	**0.718**	0.385
		Addition	0.062	0.545	0.917	0.714	0.648	0.334
Random Initialisation	LSTM	Concatenation	0.075	0.560	1.222	**0.975**	0.564	0.422
		Addition	0.091	0.536	1.188	0.805	0.539	0.362
	GRU	Concatenation	0.067	0.507	1.308	0.763	0.632	0.394
		Addition	0.084	0.525	0.857	0.625	0.547	0.287

Discussion. Table 1 shows that there is no clear superior configuration across the different metrics, but overall, the Fine-tunable-Word2vec-LSTM-Concatenation configuration performs the best across the different metrics. Its generated captions are shown in the supplementary material. It is marginally outperformed in anatomical classification scores by the Random-Initialisation-LSTM-Concatenation configuration but scores higher in *BLEU-4* and *ROUGE-L*, implying the usefulness of pre-trained embeddings to ensure superior caption structuring compared to randomly initialised vectors. Word2vec embeddings were found to be more effective than GloVe embeddings for the fetal ultrasound datasets. It is interesting to note that, in most cases, concatenation performed better than addition, and LSTM units outperformed GRUs, even for our limited datasets. Among the anatomical classes, from Table 2, abdomen and head show low scores in *BLEU-4* and *ROUGE-L* due to having the highest lexical diversity. Spine does well in *ROUGE-L* and *GB* because of its lower lexical diversity, however, *BLEU-4* is zero due to the absence of 4-gram overlaps but *BLEU-3*=0.319 is achieved. From *LSS*, we can see that the heart class is more challenging. In clinical practice, a fetal heart is typically identified by its beating motion (a video clip) rather than a still image. Further, the current captioning system is not trained to distinguish between the different heart views, but the textual description can be heart view specific. Adding more image-caption pairs of distinct heart views may solve this problem. In Fig. 4, it can be seen that all classes are accurately identified; however, the model struggles with 11% of abdomen images, misclassifying them as hearts. On investigation, it was found that for these specific images the stomach bubble has an elongated appearance, which has some resemblance to a heart view or heart chamber.

Table 2: Evaluation results for the different anatomical structures

Structure	*BLEU-3*	*BLEU-4*	*ROUGE-L*	*GB*↓	*Class. F1*	*ARS*	*LSS*
Abdomen	0.000	0.000	0.533	0.667	0.886	0.316	0.625
Head	0.122	0.058	0.479	2.000	**1.000**	0.213	0.625
Heart	0.252	**0.140**	0.581	0.857	0.993	**0.843**	0.500
Spine	**0.319**	0.000	**0.789**	**0.000**	**1.000**	0.771	**1.000**

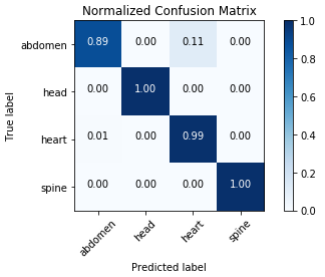


Fig. 4: Confusion Matrix

4 Conclusions

We proposed an automatic image captioning method to describe fetal ultrasound video content from four types of anatomical structures using real-world sonographer vocabularies. The Fine-tunable-Word2vec-LSTM-Concatenation performed best among the different evaluated model configurations. Richer vocabularies and extensions to spatio-temporal models will be considered in future work.

Acknowledgement

We acknowledge the ERC (ERC-ADG-2015 694 project PULSE), the EPSRC (EP/MO13774/1), the Rhodes Trust, and the NIHR BRC funding scheme.

References

1. Bernardi, R., et al.: Automatic description generation from images: A survey of models, datasets, and evaluation measures. IJCAI pp. 4970–4 (2017)
2. Cho, K., et al.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: EMNLP. pp. 1724–34. ACL (2014)
3. Department of Engineering Science, University of Oxford: PULSE, `https://www.eng.ox.ac.uk/pulse/`
4. Elliott, D., Keller, F.: Image description using visual dependency representations. In: EMNLP. pp. 1292–1302 (2013)
5. Goodfellow, I., et al.: Deep learning (2016)
6. Google Cloud: Cloud Speech-to-Text, `cloud.google.com/speech-to-text/`
7. Google Code Archive: Word2Vec (2013), `code.google.com/archive/p/word2vec/`
8. GrammarBot: Grammar Check API, `https://www.grammarbot.io/`
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. NC **9**(8), 1735–80 (1997)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2015)
11. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
12. Lyndon, D., et al.: Neural captioning for the ImageCLEF 2017 medical image challenges. CEUR Workshop Proceedings **1866** (2017)

13. McCarthy, P.M., Jarvis, S.: MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior research methods **42**(2), 381–92 (2010)
14. Mikolov, T., et al.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems (2013)
15. Ordonez, V., et al.: Im2text: Describing images using 1 million captioned photographs. In: Advances in NIPS. pp. 1143–51 (2011)
16. Papineni, K., et al.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on ACL. pp. 311–8. ACL (2002)
17. Pennington, et al.: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–43 (2014)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
19. Sloetjes, H., Wittenburg, P.: Annotation by category-ELAN and ISO DCR. In: LREC (2008)
20. Tanti, M., et al.: What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator? pp. 51–60. ACL (2017)
21. Tanti, M., et al.: Where to put the image in an image caption generator. Natural Language Engineering **24**(3), 467–89 (2018)
22. Vinyals, O., et al.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on CVPR. pp. 3156–3164 (2015)
23. You, Q., et al.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on CVPR. pp. 4651–9 (2016)
24. Zeng, X.H., et al.: Understanding and generating ultrasound image description. Journal of Computer Science and Technology **33**(5), 1086–100 (2018)