A COURSE-FOCUSED DUAL CURRICULUM FOR IMAGE CAPTIONING

Mohammad Alsharid[1*] *Rasheed El-Bouri*[1*] *Harshita Sharma*[1] *Lior Drukker*[2]
Aris T. Papageorghiou[2] *J. Alison Noble*[1]

[1]Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[2]Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK

ABSTRACT

We propose a curriculum learning captioning method to caption fetal ultrasound images by training a model to dynamically transition between two different modalities (image and text) as training progresses. Specifically, we propose a course-focused dual curriculum method, where a course is training with a curriculum based on only one of the two modalities involved in image captioning. We compare two configurations of the course-focused dual curriculum; an image-first course-focused dual curriculum which prepares the early training batches primarily on the complexity of the image information before slowly introducing an order of batches for training based on the complexity of the text information, and a text-first course-focused dual curriculum which operates in reverse. The evaluation results show that dynamically transitioning between text and images over epochs of training improves results when compared to the scenario where both modalities are considered in equal measure in every epoch.

***Index Terms*—** Image captioning, curriculum learning, fetal ultrasound, image description, meta-learning

1. INTRODUCTION

Fetal ultrasound (US) image captioning is the process of generating text that describes the content of a fetal ultrasound image. It shares some similarities with the more established natural image captioning [1] but is typically characterized by small dataset size and a lack of corresponding textual descriptions describing the content of the images. A limited number of expert annotators available for the formulation of good quality captions makes it hard to undertake large-scale annotation processes as typically used in natural image captioning [2]. This points to the need for solutions for medical imaging tasks that involve natural language processing to efficiently utilize smaller-scale data.

Fetal ultrasound image interpretation is a very challenging task. From a clinical point of view, being able to automatically caption ultrasound videos may benefit trainees learning how to scan. Trainees often watch ultrasound videos offline; however, these are rarely captioned. Addressing this phenomenon is potentially the real clinical use of the techniques and approaches that we propose in this paper.

Previous work [3] addressed this problem using a dual curriculum by 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 this paper, we explore a natural extension of using a linear combination of the complexity metrics of a single multi-modal data sample. 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.

The modalities are clearly different and complementary, and one may be more challenging (e.g. data samples deviate greatly from the mean) to learn by the network than the other. The aim of this work is to quantitatively determine the best weighting combination and explore the advantages, or otherwise, of doing this. It answers the question: Would unequal weighting be beneficial? To the best of our knowledge, this is the first work to investigate training image captioning models with different curricula built during each epoch with a changing weighting for each modality.

1.1. Related Work

Image captioning is made possible through either text retrieval or text generation [1, 4]. Retrieval is when a pre-existing caption (or a combination of captions) is retrieved from the training dataset and associated with a new unseen image [5]. The caption chosen is decided by first determining which of the images in the training set have the highest similarity scores with the unseen image in question.

Generation, as its name suggests, is the process whereby captions are generated. These generated descriptions might not exist in the training set, but the vocabulary used will come from words encountered in the training set [6].

In [7], a teacher network is used to specify to a student network what sub-task to work on first in a curriculum learning framework that consists of teacher-student networks. Curriculum learning has found its way into medical imag-

*Equal contribution.

ing as researchers have adopted it to overcome limitations in class imbalance [8] and issues caused by data that is weakly-labelled [9].

In [6], a fundamental approach to captioning ultrasound images with curriculum learning was considered. In the current work, we take this further, by having unequal influence on the two modality-specific metrics that together constitute the previously introduced dual curriculum. Each metric represents the inherent complexity of one of the two modalities present in image captioning.

2. METHODS

The dual curriculum ranks the data samples in the training set according to the sum of two metrics H_1 and H_2 in an effort to determine the inherent complexity in a multi-modal sample. Each data sample consists of two elements, an image and a corresponding text caption. H_1 is the complexity that the image possesses in relation to the mean using the Wasserstein distance as shown in Eqn. 1:

$$H_1 = \inf_{\pi \in \Gamma(\hat{x}_n, \hat{u})} \int_{R \times R} |m - n| d\pi(m, n) \tag{1}$$

where Γ represents the collection of all measures on an $R \times R$ space over all joint distributions of the random variables m and n with the softmaxed distributions of the image features $\hat{x}_n$ and the mean of the image features $\hat{u}$ serving as marginals.

H_2 is the complexity that the caption possesses represented through the tf-idf (term-frequency inverse-document frequency) scores of its constituent words [10] in Eqn. 2:

$$H_2 = \sum_{w} \frac{\#(w \in c)}{\Sigma_{w_i}^{W} \#(w_i \in c)} \ln \frac{\#(c \in C)}{\#(c \; if \; w \in c)} \tag{2}$$

where $\#()$ is a count function that counts the number of times a word w is repeated in a caption c which is made up of all its constituent words W. c is only one caption of all captions C that exist in the training dataset. The tf-idf score provides a measure of the meaningfulness that a word possesses. For each caption, the tf-idf scores of its constituent words are summed, allowing us to incorporate the length of a caption into its complexity. After summation, the data samples are ordered for training such that the data samples with the lowest summed scores start the curriculum and constitute the bulk of the earlier batches. In this work, curriculum learning, as in [3], attempts to alleviate the dependence of the results on the specifics and the hyperparameters of the model being trained and depend instead on the intrinsic nature of the data and its complexity as explained earlier.

The vanilla dual curriculum sums the two complexity metrics after min-max normalisation as shown in Eqn. 3 where s is one data sample:

$$d_{dc} = \frac{H_{1s} - min(H_1)}{max(H_1) - min(H_1)} + \frac{H_{2s} - min(H_2)}{max(H_2) - min(H_2)} \tag{3}$$

Each data sample has its own d_{dc} score, and the curriculum is ordered such that the earlier batches consist of data samples with the lowest d_{dc} scores.

This paper investigates unequal weighting of the two 'courses' (image and text) in the curriculum. Specifically, we explore two variants of the course-focused dual curriculum (CF-DC); an image-first course-focused dual curriculum (I1-CF-DC) and a text-first course-focused dual curriculum (T1-CF-DC). The equation for the CF-DC is shown in Eqn. 4:

$$d_{cf} = w_1 \frac{H_{1s} - min(H_1)}{max(H_1) - min(H_1)} + w_2 \frac{H_{2s} - min(H_2)}{max(H_2) - min(H_2)} \tag{4}$$

In I1-CF-DC, w_1 starts at 1.0 and proceeds to decrease linearly per epoch to a value of 0.0, while w_2 starts at 0.0 and proceeds to increase at the same rate. Effectively, $w_1(Epoch = 0) = 1.0$ and $w_1(Epoch = N) = 0.0$ where N is the final epoch with $w_2 = |1 - w_1|$ throughout. With T1-CF-DC, it is the other way around. In the T1-CF-DC approach, we start with placing a higher emphasis on the complexity metric of the text information rather than that of the image information. So, w_1 starts at 0.0, and w_2 starts at 1.0, and then, the weights linearly increase and decrease respectively by a factor of 0.1 after a run through the curriculum. This process continues until the final run of the curriculum where w_1 would be 1.0 and w_2 would be 0.0.

Unlike the vanilla dual curriculum, a course-focused dual curriculum is effectively a meta-curriculum, a hierarchical curriculum that prepares and orders different curricula for model training first based on the weighting associated with a modality (w_1 and w_2) and then based on the complexity of the data samples according to H_1 and H_2.

Another way to update w_1 and w_2 is to evaluate the model with the validation set after every curriculum iteration and to observe whether the model has scored better with syntax-focused metrics (B1 [11], RL [12], GB[13]) or semantics-focused metrics (ARS [6], F1). w_1 is decreased and w_2 is increased if semantics-focused metrics are higher than syntax-focused metrics or vice-versa. This is similar to the approach followed in [14] where performance on the validation set is used to guide the automatic creation of a curriculum.

2.1. Image Captioning Model Architecture

Fig. 1 shows the image captioning model that is used in this work which is the late merge captioning model used in [3, 6]. It consists of two branches. The right branch handles the image information and consists of a fine-tuned VGG-16 and fully-connected layers. The left branch embeds each tokenised word with a Word2vec embedding vector before passing it through a recurrent neural network. A flat feature vector from each branch is concatenated together and a prediction for the next word in the sequence is made based on the concatenated output.

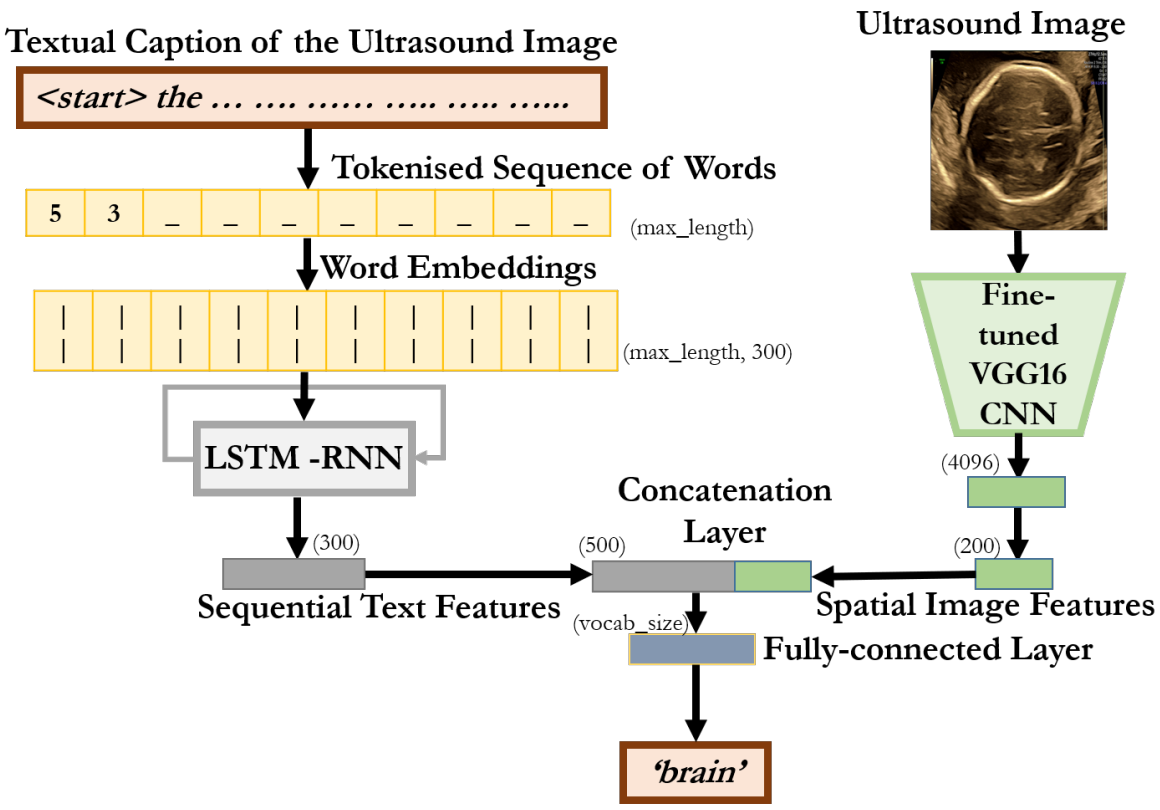


Fig. 1. The image captioning model. 'max_length' represents the maximum number of words in a caption.

3. EXPERIMENTS

3.1. Dataset and Data Preparation

As part of the PULSE study [15], information from full-length fetal ultrasound scans are acquired including ultrasound video and audio data. The audio data had been transcribed into text [6] from ten audio-video sequences of which image-caption pairs are prepared. Those ten audio-videos have a mean duration lasting 37 minutes.

Data from seven of the ten videos were used for training. Two of the ten served as a validation set, and the tenth video was the test set. This arrangement gave us 12,808 image-caption pairs to train and validate models on and 9,979 image-caption pairs for model testing. This tenth video that we have reserved as the test set is almost twice as long as most of the videos that went into the training set. Its large number of available image-caption pairs made it suitable to be the one video set aside for testing.

Pre-processing of the data samples included cropping the images to remove the US system's GUI, resizing the US images to 224×224 pixels as VGG16 expects, removing punctuation from the captions, and converting uppercase letters to lowercase.

Table 1. Quantitative Results

	Syntax-Focused			Semantics-Focused	
Metric	*B1*	*RL*	*GB↓*	*ARS*	*F1*
NC	0.18	0.37	—	0.23	0.77
DC [1]	0.12	0.42	1.60	**0.43**	**0.97**
I1-CF-DC	**0.34**	**0.49**	**0.89**	0.39	0.96
T1-CF-DC	0.30	0.46	1.01	**0.43**	**0.97**

4. RESULTS AND DISCUSSION

Table 1 compares the results of CF-DC trained models introduced in this paper with the results of a DC trained model and a traditionally trained model (NC for no curriculum) [3]. Table 1 shows that one approach (I1-CF-DC) performs better with respect to metrics focused on evaluating sentence quality or the syntax (B1, RL, GB), and one (DC) is better with respect to metrics that deal with semantics and anatomic relevance (ARS, F1). However, T1-CF-DC has similar performance metric scores to DC on the semantics-focused metrics while improving scores on the syntax-focused metrics.

The question that arises is why T1-CF-DC is the best performing method. In the vanilla dual curriculum (DC) where w_1 and w_2 both have fixed values of 1.0 throughout training,

we are effectively trying to merge two different curricula together. There is nothing to guarantee that one curriculum is not interfering with another's progress. There might be instances of forgetting as described in [7]. In I1-CF-DC, the earlier focus on the image-associated complexity metric when assigning a dual curriculum score to a data sample loses its effect. The more complex and widely varied information in the text makes it more likely to forget what the model has learned from first primarily learning from the image-focused early part of the curriculum.

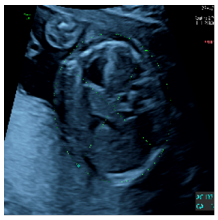
However, for T1-CF-DC, the data samples that are first encountered are primarily determined based on how 'easy' or 'hard' they are on the text-associated complexity metric. There was less forgetting that had to happen when later dealing with the image-focused part of the curriculum. The reasons for this are twofold. There is a greater variety and therefore complexity in the text information. We can imagine that a captioning model is effectively an image classifier with four possible classes connected with a language model that has hundreds of possible 'classes' that it needs to learn, each associated with a single word in the training vocabulary. The second reason is the fact that even when dealing with easy examples according to the text complexity metric, the model is still learning the image-based semantics through the anatomical vocabulary. The words of the vocabulary carry semantic meaning relevant to the image information while the opposite (learning the image information doesn't lead to learning the words) does not necessarily hold. These two reasons explain why the T1-CF-DC trained model had the most balanced scores across the different metrics.

B1 (or BLEU-1) [11] and RL (ROUGE-L) [12] are metrics commonly used when evaluating image captioning models. They look at the overlap in words between generated captions and the ground truth. GB (GrammarBot) [13] is an API that we have used to determine the number of grammatical mistakes that may occur on average in a generated caption.

Table 2 compares the performance of course-focused dual curricula on complex data. A complex data sample in an image captioning problem is an image-caption pair with an image feature vector that is at a high Wasserstein distance from the mean of the image feature vectors and a caption that is long (in terms of number of words) and consists of more infrequent words as determined by the tf-idf scores. To prepare this table, we took the test set, ranked each data sample by its complexity, and took the more complex half of the test set to evaluate on. Effectively, the same phenomenon observed in the previous experiment holds true here as well with I1-CF-DC performing better with syntax-focused metrics while T1-CF-DC performing better with semantics-focused metrics.

5. CONCLUSION AND FUTURE WORK

In this work, we have explored variants of the dual curriculum for fetal image captioning. A text-first course-focused dual curriculum was found to perform best. In the future, we will investigate whether this conclusion holds true for the temporal equivalent of video clip captioning.



GT: this is where we measure the abdominal circumference
DC: us measure around the baby's belly
I1-CF-DC: let us measure around the baby's belly
T1-CF-DC: measuring the abdomen is like putting measuring tape around the baby's waist

Fig. 2. Qualitative results for an abdomen image. GT is for 'Ground Truth'. This is the caption that was provided by the sonographer directly. DC is for 'Dual Curriculum'. This is the caption that was generated by a model that was trained with a dual curriculum. I1-CF-DC is for 'Image-First Course-Focused Dual Curriculum'. This is the caption that was generated by a model that was trained with an I1-CF-DC. T1-CF-DC is for 'Text-First Course-Focused Dual Curriculum'. This is the caption that was generated by a model that was trained with a T1-CF-DC.

GT: this is the left ventricular outflow tract
DC: ventricular outflow tract and right ventricular outflow tract are looking very very good
I1-CF-DC: ventricular outflow tract
T1-CF-DC: this is the left ventricular outflow tract

Fig. 3. Qualitative results for a heart image. GT is for 'Ground Truth'. This is the caption that was provided by the sonographer directly. The full forms of 'DC','I1-CF-DC', 'T1-CF-DC' can be found in the caption of Fig. 2.

Table 2. Comparing CF-DCs on Complex Data

	Syntax-Focused			Semantics-Focused	
Metric	*B1*	*RL*	*GB*↓	*ARS*	*F1*
I1-CF-DC	**0.38**	**0.47**	**0.56**	0.40	0.96
T1-CF-DC	0.33	0.43	0.88	**0.42**	**0.97**

6. COMPLIANCE WITH ETHICAL STANDARDS

This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051) and the ERC ethics committee.

7. ACKNOWLEDGMENTS

We acknowledge the ERC (ERC-ADG-2015 694581 project PULSE), the EPSRC (EP/MO13774/1), the Rhodes Trust, and the NIHR Oxford Biomedical Research Centre (BRC)

funding scheme. ReB is supported by an EPSRC Industrial Strategy Challenge Fund PhD studentship. The authors have no financial conflicts of interest to disclose related to this work.

8. REFERENCES

[1] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, 2016.

[2] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[3] M. Alsharid, R. El-Bouri, H. Sharma, L. Drukker, A.T. Papageorghiou, and J.A. Noble, "A curriculum learning based approach to captioning ultrasound images," in *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pp. 75–84. Springer, 2020.

[4] M. Tanti, A. Gatt, and K.P. Camilleri, "What is the role of recurrent neural networks (rnns) in an image caption generator?," *arXiv preprint arXiv:1708.02043*, 2017.

[5] V. Ordonez, G. Kulkarni, and T.L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Advances in neural information processing systems*, 2011, pp. 1143–1151.

[6] M. Alsharid, , H. Sharma, L. Drukker, P. Chatelain, A.T. Papageorghiou, and J.A. Noble, "Captioning ultrasound images automatically," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 338–346.

[7] T. Matiisen, A. Oliver, T. Cohen, and J. Schulman, "Teacher-student curriculum learning," *IEEE transactions on neural networks and learning systems*, 2019.

[8] I. Oksuz, B. Ruijsink, E. Puyol-Antón, J.R. Clough, G. Cruz, A. Bustin, C. Prieto, R. Botnar, D. Rueckert, J.A. Schnabel, et al., "Automatic cnn-based detection of cardiac mr motion artefacts using k-space data augmentation and curriculum learning," *Medical image analysis*, vol. 55, pp. 136–147, 2019.

[9] B. Park, Y. Cho, G. Lee, S.M. Lee, Y.H. Cho, E.S. Lee, K.H. Lee, J.B. Seo, and N. Kim, "A curriculum learning strategy to enhance the accuracy of classification of various lesions in chest-pa x-ray screening for pulmonary abnormalities," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[10] C. Sammut and G.I. Webb, Eds., *TF–IDF*, pp. 986–987, Springer US, Boston, MA, 2010.

[11] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[12] C.Y. Lin, "Rouge: a package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out. Barcelona, Spain*, 2004, pp. 56–60.

[13] S. Loria, "Grammarbot," 2020, Available at https://www.grammarbot.io/.

[14] R. El-Bouri, D. Eyre, P. Watkinson, T. Zhu, and D. Clifton, "Student-teacher curriculum learning via reinforcement learning: predicting hospital inpatient admission location," *arXiv preprint arXiv:2007.01135*, 2020.

[15] Department of Engineering Science at the University of Oxford, "Pulse," 2019, Available at https://www.eng.ox.ac.uk/pulse/.