Published in final edited form as:

Med Ultrasound Preterm Perinat Paediatr Image Anal (2020). 2020 October ; 12437: 75–84.

doi:10.1007/978-3-030-60334-2_8.

A Curriculum Learning Based Approach to Captioning Ultrasound Images

Mohammad Alsharid^{#1}, Rasheed El-Bouri^{#1}, Harshita Sharma¹, Lior Drukker², Aris T. Papageorgiou², J. Alison Noble¹

¹Institute of Biomedical Engineering, University of Oxford, UK

²Nuffield Dept. of Women's & Reproductive Health, University of Oxford, UK

[#] These authors contributed equally to this work.

Abstract

We present a novel curriculum learning approach to train a natural language processing (NLP) based fetal ultrasound image captioning model. Datasets containing medical images and corresponding textual descriptions are relatively rare and hence, smaller-sized when compared to the datasets of natural images and their captions. This fact inspired us to develop an approach to train a captioning model suitable for small-sized medical data. Our datasets are prepared using real-world ultrasound video along with synchronised and transcribed sonographer speech recordings. We propose a “dual-curriculum” method for the ultrasound image captioning problem. The method relies on building and learning from curricula of image and text information for the ultrasound image captioning problem. We compare several distance measures for creating the dual curriculum and observe the best performance using the Wasserstein distance for image information and tf-idf metric for text information. The evaluation results show an improvement in all performance metrics when using curriculum learning over stochastic mini-batch training for the individual task of image classification as well as using a dual curriculum for image captioning.

Keywords

Image captioning; Curriculum learning; Fetal ultrasound

1 Introduction

Image captioning involves the generation of words constituting a descriptive caption given an image of interest. Image captioning has proven to be popular in computer vision, and this fact is illustrated by the number of datasets consisting of natural images and their corresponding captions made available as large-scale datasets [4]. Medical image captioning is an area of growing interest because of the potential to automate reporting and assist non-experts in interpreting images. Generating clinically meaningful captions requires medical

Correspondence to: Mohammad Alsharid.

mohammad.alsharid@eng.ox.ac.uk.

experts for the annotation process; however, such experts are an expensive and often scarce resource. Using crowd-sourcing services, as used in some natural image captioning tasks, is not an option since laypeople do not possess the professional knowledge to describe clinical images. In addition, patient privacy and confidentiality prevent the widespread sharing of sensitive medical data. Unfortunately, these facts leave researchers with relatively small-sized datasets of medical images and their corresponding captions. Therefore, in this paper, we consider a training regimen more suited and tailored for small-sized multi-modal medical data of images and text. The contribution of this paper is the development of algorithms to prepare suitable curricula using curriculum learning techniques from a relatively small-sized training dataset and the demonstration of how this leads to improved deep learning-based image classification and captioning models for fetal ultrasound images. To the best of our knowledge, this is the first work to propose a curriculum learning technique to prepare curricula based on both medical image and text data concurrently for the task of medical image captioning.

Randomly shuffling the training dataset throughout the epochs of training is standard practice in deep learning. This randomness is meant to help regularise model training. The goal of regularisation is to prevent the values of model parameters from getting stuck in local minima of the optimisation function.

In curriculum learning, the defining concept is the order in which data samples are introduced. Thus, curriculum learning also attempts to perform regularisation but in a more organised, less random way. Curriculum learning was introduced in [3] to mimic the way that learning occurs in humans. Humans are trained with a curriculum that begins with easy concepts and progresses to harder, more ‘complex’ topics [9]. In the literature, a number of ways to approach and define curriculum learning have been proposed [5]. In this paper, curriculum learning is independent of any specifics of the model architecture or its hyperparameters that determine how it is trained. The only aspect under consideration is how the training batches are prepared and presented to the network. Curriculum learning techniques address the fact that optimising the parameters of a neural network is a non-convex problem. Therefore, a function’s global minimum is not necessarily reachable, and the lowest possible error cannot be obtained during normal stochastic mini-batch gradient descent training [3, 5].

Related Works

There have only been a few studies investigating the captioning of ultrasound images. A convolutional neural network (CNN) based captioning model for second trimester fetal ultrasound images that fuses text and image information for the next word generation is described in [2]. In that method, random shuffling of training data was used without any optimization based on sample complexity. We improve on the findings of [2] using a curriculum learning approach for small-sized datasets; however, a direct comparison with that paper is not possible due to the use of different datasets (details in Section 3). An image captioning framework to caption ultrasound images of the abdomen is built in [23]. Their framework begins with a classifier to identify the structure of interest in the given abdominal ultrasound image, before passing the ultrasound image to the encoding convolutional neural

network (CNN) of the captioning model of that structure. In contrast, in the current paper, the classifier is effectively also responsible for encoding the image information for the captioning model. Examples of using curriculum learning exist in computer vision [18] and natural language processing [6, 22, 10] as well as a few in biomedical imaging applications, such as classification of lesions in chest X-rays [17] to address weakly labeled data and detection of cardiac MR motion artefacts [14] to deal with class imbalance. In this paper, we consider techniques to build curricula to aid in the captioning of second trimester fetal ultrasound images. To the best of our knowledge, this paper is the first work towards optimizing results in medical image captioning, a multi-modal task, for small-sized real-world datasets, by proposing the training of the deep neural network in an organised manner through a dual curriculum.

2 Methods

Curriculum Learning and Distance Measures

To understand curriculum learning, we first define a metric, H , that encodes some notion of complexity (examples of H are defined in the paragraphs below). We then split the training dataset into batches according to this complexity such that $H[B_i] < H[B_j]$, where B is a batch of data from our training set, i is an index of an earlier batch, and j is an index of a later batch. These batches are ordered for presentation to the network and training according to the choice of curriculum. There is no current consensus on what a general curriculum looks like as a curriculum is usually tailored to the problem at hand. For example, for some problems, models achieve greater performance when trained with ‘anti-curricula’ (presentation of data from high ‘complexity’ or entropy to low) [3], [13]. In this paper, we experiment with various configurations and curriculum strategies.

It has been hypothesized that curriculum learning performs a similar function to numerical continuation methods [3], whereby a complex surface is approximated by a smooth version before increasing the complexity of that surface to become more similar to the original [1]. Entropy is a measure of randomness and uncertainty. Data samples with low entropy with respect to the mean of the samples are ‘easier’ to learn from. In terms of the curriculum, by optimising with a low entropy, ‘easy’ to learn batch, we provide a smoother version of the prediction error surface. By training sequentially with data in higher ‘complexity’ batches, we progressively increase the complexity of the surface while already having some parameters within the domain of a minimum which takes the solution closer to the global minimum, thus giving us model weights that can provide better results. In this work, when building an image-based curriculum, we experiment with the Mahalanobis distance [11], the cosine similarity and the Wasserstein distance as H metrics. We perform experiments using a forward curriculum (low H to high H). The data samples for image focused tasks are image feature vectors extracted through a finetuned VGG16 [21].

The Mahalanobis distance d_{ml} measures how far, or how different, an image feature vector x_n is from the mean of the samples μ . It is the multi-dimensional generalisation of standard deviation from the mean and given in Eqn. 1,

$$d_{ml} = ((\mathbf{x}_n - \mathbf{u})^T S^{-1} (\mathbf{x}_n - \mathbf{u}))^{(1/2)} \quad (1)$$

where S is the covariance matrix.

The cosine similarity d_{cos} is defined in Eqn. 2. We use a fixed reference (mean) vector to compare the dot products of model inputs with.

$$d_{cos} = \mathbf{x}_n \mathbf{u} / |\mathbf{x}_n| |\mathbf{u}| \quad (2)$$

The Wasserstein distance d_{ws} is defined as the distance between two distributions and is mathematically defined in Eqn. 3,

$$d_{ws} = \inf_{\pi \in \Gamma(\hat{x}_n, \hat{u})} \int_{R \times R} |m - n| d\pi(m, n) \quad (3)$$

where Γ denotes the collection of all measures on a space, $R \times R$, over all joint distributions of the random variables m and n with marginals $\hat{x}_n$ and $\hat{u}$ respectively. In this case, $\hat{x}_n$ and $\hat{u}$ are the softmaxed distributions of our features and the mean of the features respectively. The use of the Wasserstein distance arose to determine the optimal transport of piles of sand, represented by the distributions $\hat{x}_n$ and $\hat{u}$ [15]. To use a Wasserstein distance-based curriculum, we first apply a softmax function on the image feature vectors, representing them as distributions. Then, the distance between the distribution of every image vector and the distribution of the mean feature vector is computed.

For preparing text-focused curricula, we used term frequency-inverse document frequency (tf-idf) score [19], as calculated by Eqn. 4:

$$d_{ti} = \frac{\#(w \in c)}{\sum_{w_i} \#(w_i \in c)} \ln \frac{\#(c \in C)}{\#(c \text{ if } w \in c)} \quad (4)$$

where d_{ti} is a word's tf-idf score, $\#()$ is a count function, w is a word, c is a caption, W are all the words in caption c , and C are all the captions in the dataset. It 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.

The dual curriculum, as used in the captioning task, uses two metrics of complexity, one for each data modality, namely, image H_1 and text H_2 . Each metric ranks the data samples s in the dataset S differently. With dual curriculum, we combine these rankings. The scales used by the different metrics are not the same, so it was necessary to first normalise each metric before obtaining a combined measure of 'difficulty'. This allows us to take into account the difficulty associated with both images and text as shown in Eqn. 5:

$$d_{dc} = \frac{H_{1s} - \min(H_1)}{\max(H_1) - \min(H_1)} + \frac{H_{2s} - \min(H_2)}{\max(H_2) - \min(H_2)} \quad (5)$$

Image Classification Model

Before considering the image captioning task, we first consider which distance metric works best for the simpler task of fetal ultrasound image classification. This image classification model is later used as the image information branch of the image captioning model shown in Fig. 1. In other words, the upper branch of the image captioning model serves as the base of the image classification model. Specifically, the problem is to classify an ultrasound image given four classes; the abdomen, the head, the heart, and the spine. We consider a pre-trained VGG16 [21] that has been fine-tuned on fetal ultrasound images. The images and labels used in fine-tuning the VGG16 are obtained from video clips in the US scan videos that were manually labelled based on the viewed anatomies [20]. The VGG16 architecture provides image feature vectors that can serve as input for a shallow 3-layer image classifier. We compare four variants of the model for evaluation of distance measures trained using: 1) mini-batches with stochastic gradient descent and random sampling, 2) a Mahalanobis distance based image curriculum (MD-IC), 3) a cosine similarity based image curriculum (CD-IC), 4) a Wasserstein distance based image curriculum (WD-IC).

Image Captioning Model

Fig. 1 shows the image captioning model. The image captioning model consists of two separate branches for image and text modalities, combined in a late fusion configuration using a concatenation layer followed by a fully-connected classification layer [2]. The upper branch followed by a fully-connected layer with softmax activation represents the image feature classifier model. “?” represents a varying number of words depending on the vocabulary of the anatomical structure that the specific captioning model would correspond to. For the image captioning task, we explored three different dual-curriculum options (MD-DC, CD-DC, WD-DC) that combine metrics discussed in the previous section with the tf-idf score. Four captioning models were trained, one for each anatomical structure. During training the partial caption comes from the ground truth caption, whereas at inference, the model would rely on its previously generated words. At inference, the US image was first classified using the trained US image classifier. Once the associated anatomical structure of the image was determined, the corresponding captioning model was deployed to caption the image. Posing the captioning problem this way places a high importance on first correctly identifying the structure.

Training Process

In all experiments, the loss function was categorical cross entropy, and Adam optimisation was used. When training, the number of batches was 32 for the classification task and eight for the captioning task. In the captioning tasks, the words were embedded with fine-tunable Word2vec embedding vectors that were pre-trained on the Google News corpus [12]. Models were trained for 150 epochs with early stopping. During training, teacher forcing

was used where partial captions consisted of the ground truth words rather than the words the model generated at previous steps, but at inference, the captioning model had to rely on its previously generated words to generate the next word in the caption. Dropout was applied throughout with a rate between 0.4 and 0.5.

3 Experiments

Dataset and Data Preparation

The image-caption pairs were obtained from ten full-length fetal ultrasound scans as part of the PULSE study [8]. The mean length of the videos is 37 minutes. The UK Research Ethics Committee (Reference 18/WS/0051) approved this study. Written informed consent was given by all participating pregnant women and sonographers. The images were obtained from ultrasound video frames [8] with audio additionally recorded. The audio recordings were transcribed to provide the captions for the images in the dataset [2]. In this work, we use a total of ten full-length transcribed videos, as opposed to five videos in [2]. Seven video scans were used for training, two were used for validation, and one was used for testing. The training and validation data were subsampled to ensure an equal distribution of the four different anatomical structures in the dataset. A total of 12808 image-caption pairs were used for training and validation. The test set consisted of 9979 image-caption pairs. All images were cropped to remove the user interface, keeping only the ultrasound image, and then resized to 224×224 pixels. All punctuation was dropped from the text, and all letters made lowercase.

Performance Evaluation

To evaluate the model performance, we use the *F1* score, *BLEU-1* (*B1*) [16], *ROUGE-L* (*RL*) [7], and *ARS* (the Anatomical Relevance Score) as introduced in [2]. *BLEU-1* and *ROUGE-L* have been slightly modified for the task at hand as follows. Traditionally, both these metrics check that two captions are similar based on their constituent words with no regard to the actual semantic relevance of the generated words to those in the ground truth caption. In order to prevent relatively high scores being obtained when no relevant anatomical terminology is generated, we calculate *B1* and *RL* by modifying them according to the anatomical classification performance. Each image has a ground truth label (one of the four anatomical structures), hence, when evaluating a captioning model, the *B1* and *RL* scores for the corpus (i.e the image-captioning test set of that structure) are multiplied by the percentage of images that are correctly classified by the image classification model of the same curriculum. If the percentage of correctly classified images is high (e.g. 90%) for that structure, the *B1* and *RL* scores for that structure will only be slightly reduced. If the percentage of images that are misclassified is high, the *B1* and *RL* scores will drop. This will penalise models for generating good captions but for the wrong structure. We then take the mean of the scores of each structure.

4 Results and Discussion

The results for the image classification task are shown in Table 1 (P : Precision, R : Recall). We observe that the WD-IC trained model outperformed the traditionally trained (no curriculum) model and the CD-IC trained model. The latter two relatively struggled. The MD-IC trained model performed well and was only slightly inferior in terms of accuracy to the WD-IC model. For the classification problem, the WD-IC and MD-IC models both performed comparably well, and we see this behavior extended into the captioning problem.

Fig. 2 shows the distribution of structures in ten batches prepared by an MD-IC, a CD-IC, and a WD-IC. For visualisation purposes, we assume that the data is split into ten batches. A higher batch number would mean 'harder' data samples in the batch. From this visualisation, we can observe class imbalance in the batches. Preliminary experiments in image classification showed that the curricula struggled when batches had a disproportionate number of samples of the different structures as in Fig. 2. Future work will investigate this observation. To alleviate this problem, for the rest of the experiments reported in this paper, batch preparation was done through the curricula as follows. Rather than having a single mean of the whole dataset from which a distance measure is calculated for each data sample, each structure had its own mean. In each of these balanced batches, for each structure, we had data samples with the lowest entropy relative to that structure's mean instead of the data samples with the lowest entropy relative to the mean of the whole dataset.

Qualitative results of the image captioning task are shown in Fig. 3. Quantitative results are summarized in Table 2. We notice an improvement when compared to the traditionally trained model. WD-DC is superior to other curricula for three out of four metrics. The reason RL is lower for WD-DC than for MD-DC is discussed as follows. RL looks for the longest common subsequence (series of words) between the ground truth and generated captions, whereas $B1$ looks for single words in common. There might be a certain multiple-word-long way to phrase captions that exists in the training set that the model trained with MD-DC seems to be better at emulating; however, not phrasing the captions in this way is not indicative of loss of relevant information because otherwise we would have noticed ARS and $F1$ to be lower for WD-DC than for MD-DC.

5 Conclusion and Future Work

Curriculum learning helps to train more accurate models than using a traditional stochastic mini-batch preparation process does for image captioning. We built curricula for image classification and image captioning tasks suitable for small-sized medical datasets. In particular, we found the Wasserstein-distance-based curriculum and the Mahalanobis-distance-based curriculum to be good options for automatic fetal ultrasound image captioning. In the future, we will investigate if curriculum learning could help obtain comparable results while using a simpler model with less trainable parameters for real time video captioning applications.

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. ReB is supported by an EPSRC Industrial Strategy Challenge Fund PhD studentship.

References

1. Allgower EL, Georg K. Numerical continuation methods: an introduction. Springer Science & Business Media. 2012; 13
2. Alsharid, M; Sharma, H; Drukker, L; Chatelain, P; Papageorgiou, AT; Noble, JA. Captioning ultrasound images automatically. Springer; International Conference on Medical Image Computing and Computer-Assisted Intervention; 2019. 338–346.
3. Bengio, Y; Louradour, J; Collobert, R; Weston, J. Curriculum learning. Proceedings of the 26th annual international conference on machine learning; 2009. 41–48.
4. Bernardi R, Cakici R, Elliott D, Erdem A, Erdem E, Ikizler-Cinbis N, Keller F, Muscat A, Plank B. Automatic description generation from images: A survey of models, datasets, and evaluation measures. Journal of Artificial Intelligence Research. 2016; 55:409–442.
5. el Bouri R, Eyre D, Watkinson P, Zhu T, Clifton D. Student-teacher curriculum learning via reinforcement learning: Predicting hospital inpatient admission location. arXiv preprint. 2020
6. Cirik V, Hovy E, Morency LP. Visualizing and understanding curriculum learning for long short-term memory networks. arXiv preprint. 2016
7. CY, L. Rouge: a package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches Out; Barcelona, Spain. 2004. 56–60.
8. Department of Engineering Science: Pulse. 2019. <https://www.eng.ox.ac.uk/pulse/>
9. El-Bouri R, Eyre D, Watkinson PJ, Zhu T, Clifton D. Hospital admission location prediction via deep interpretable networks for the year-round improvement of emergency patient care. IEEE Journal of Biomedical and Health Informatics. 2020
10. Liu, C; He, S; Liu, K; Zhao, J. Curriculum learning for natural answer generation. Proceedings of the 27th International Joint Conference on Artificial Intelligence; 2018. 4223–4229.
11. Mahalanobis PC. On the generalized distance in statistics. National Institute of Science of India. 1936
12. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint. 2013
13. Morerio, P; Cavazza, J; Volpi, R; Vidal, R; Murino, V. Curriculum dropout. Proceedings of the IEEE International Conference on Computer Vision; 2017. 3544–3552.
14. Oksuz I, Ruijsink B, Puyol-Antón E, Clough JR, Cruz G, Bustin A, Prieto C, Botnar R, Rueckert D, Schnabel JA, et al. Automatic cnn-based detection of cardiac mr motion artefacts using k-space data augmentation and curriculum learning. Medical image analysis. 2019; 55:136–147. [PubMed: 31055126]
15. Panaretos VM, Zemel Y. Statistical aspects of wasserstein distances. Annual Review Of Statistics And Its Application. 2019; 6(6(ARTICLE)):405–431.
16. Papineni, K; Roukos, S; Ward, T; Zhu, WJ. Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics; 2002. 311–318. Association for Computational Linguistics
17. Park B, Cho Y, Lee G, Lee SM, Cho YH, Lee ES, Lee KH, Seo JB, Kim N. A curriculum learning strategy to enhance the accuracy of classification of various lesions in chest-pa x-ray screening for pulmonary abnormalities. Scientific reports. 2019; 9(1):1–9. [PubMed: 30626917]
18. Ren, Z; Wang, X; Zhang, N; Lv, X; Li, LJ. Deep reinforcement learning-based image captioning with embedding reward. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. 290–298.
19. Sammut, C, Webb, GI. TF-IDF. Springer US; Boston MA: 2010. 986–987.

20. Sharma, H; Droste, R; Chatelain, P; Drukker, L; Papageorghiou, A; Noble, J. Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019); IEEE; 2019. 987–990.
21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint. 2014
22. Tsvetkov, Y; Faruqui, M; Ling, W; MacWhinney, B; Dyer, C. Learning the curriculum with bayesian optimization for task-specific word representation learning. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016. 130–139.
23. Zhu X, Li L, Liu J, Peng H, Niu X. Captioning transformer with stacked attention modules. Applied Sciences. 2018; 8(5):739.

Fig. 1.
The image captioning model. 'n' represents the maximum sequence length.

Fig. 2. Distribution of anatomical structures in an MD-IC (left), a CD-IC (center), and a WD-IC (right) with non-balanced batches (with respect to anatomy class).

GT: 'belly'
NC: 'good heart'
WD-DC: 'this is baby's belly
the black space the black circle
is the baby's stomach'

GT: 'the three vessel view'
NC: 'is the the ventricular
outflow tract'
WD-DC: 'three vessel view'

Fig. 3.

Qualitative results for an abdomen image and a heart image. GT is for 'Ground Truth'. 'NC' is for 'No Curriculum'.

Image Classification**Table 1**

Metric	<i>P</i>	<i>R</i>	<i>F1</i>
No Curriculum	0.83	0.76	0.73
CD-IC	0.63	0.73	0.67
MD-IC	0.94	0.95	0.95
WD-IC	0.96	0.96	0.96

Image Captioning**Table 2**

Metric	<i>BI</i>	<i>RL</i>	<i>ARS</i>	<i>F1</i>
No Curriculum	0.18	0.37	0.23	0.77
CD-DC	0.22	0.34	0.22	0.73
MD-DC	0.25	0.47	0.40	0.96
WD-DC	0.27	0.42	0.43	0.97