Explainability of Self-Supervised Representation Learning for Medical Ultrasound Video

Kangning Zhang **Jianbo Jiao** **J. Alison Noble**
Department of Engineering Science, University of Oxford
kangning.zhang@wolfson.ox.ac.uk

Abstract

This paper concerns how machine learning explainability advances understanding of self-supervised learning for ultrasound video. We define the explainability as capturing anatomy-aware knowledge and propose a new set of quantitative metrics to evaluate explainability. We validate our proposed explainability approach on medical fetal ultrasound video self-supervised learning and demonstrate how it can guide the choice of self-supervised learning method. Our approach is attractive as it reveals biologically meaningful patterns which may instil human (clinician) trust in the trained model.

1 Introduction

Self-supervised (or unsupervised) representation learning models have shown recent success in the field of medical imaging. While some approaches have outperformed human experts [3], lack of algorithm explainability has hindered wide applications [2]. Doctors, medtech regulators, and computer scientists alike need to understand how machine learning algorithms work, and when they may lead to false results [13]. Thus model explanation plays a key role in the evolution of deep learning in medical imaging and in building trust in automatic learning-based decisions [16].

A number of papers concern understanding general deep learning models [16, 1, 17, 15, 11]. Most focus on model visualization, and include permutation-based [15], propagation-based [11, 12] or attention-based methods [18, 10]. The visual-based explanations are generally coarse, especially when the information is diffuse. Quantitative metrics for explainability have been proposed [1, 17], but are still limited. The explainability of self-supervised models, on the other hand, is underexplored.

In this paper, we consider the question: *how can we quantitatively explain the effectiveness of self-supervised representation learning?* We explore the question using medical fetal ultrasound video from mid-pregnancy and define explainability as capturing anatomy-aware knowledge. A set of quantitative explainability metrics are proposed, based on visually salient landmarks [5]. We propose to interpret the quality of representation learning using the quality of landmark CNN feature clustering, with the assumption that landmark CNN features capture anatomy-aware knowledge. Our metrics provide a plausible guide for the choice of appropriate self-supervised learning method, without performing downstream tasks. By using visually salient landmarks, the method also presents a better understanding of the AI-based explanation in a clinically meaningful way.

2 Methodology

The framework of proposed explainability method is summarized in Fig. 1 and consists of four steps:

Firstly, we predict the saliency map and extract local maxima. Sonographers (the clinical operator doing the ultrasound scanning) tend to focus on anatomical informative regions to understand the image [5]. The visual attention is quantified by the saliency map. Here, we use the Saliency-VAM

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

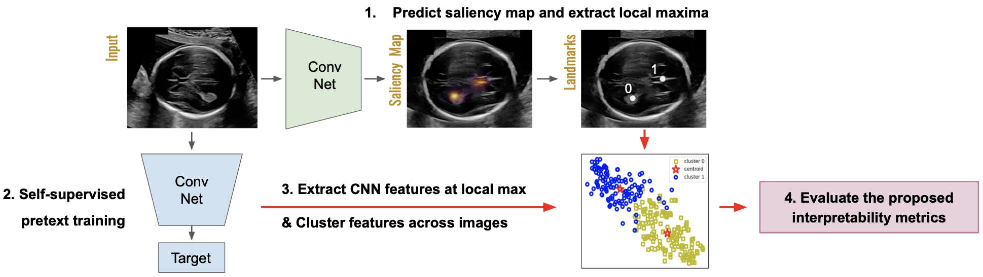
1. Predict saliency map and extract local maxima
Input
Conv Net
Saliency Map
Landmarks
1
0
cluster 0
centroid
cluster 1
2. Self-supervised pretext training
Conv Net
Target
3. Extract CNN features at local max
& Cluster features across images
4. Evaluate the proposed interpretability metrics

Figure 1: Proposed explainability method framework for self-supervised learning on medical ultrasound video, where the input ultrasound is shown in the top left.

model [5] to predict the saliency map and extract the corresponding local maxima, where the highest visual attention is located. Such local maxima can be considered as visually salient landmarks, and are proved to correspond to key anatomical structures in US scans [4].

The second step is self-supervised pre-training according to defined pretext tasks: temporal sequence sorting [8], rotation prediction [7] and pace prediction [14]. These models are trained with a 3D ResNet-18 [7] (with blocks adapted according to the Saliency-VAM model) as backbone network architecture with a SGD optimizer of momentum 0.9 and weight decay 1×10^{-4}.

The remaining two steps are 3) clustering self-supervised model features at visually salient landmarks across images using K-means, and 4) measuring the clustering quality with the proposed metrics (below). We assume that in order to learn explainable representations, the model should focus on anatomically informative regions. Therefore, the corresponding features should well-cluster to align the visually salient landmarks. Hence, we use the quality of clustering to express the explainability of the learned representations.

Here, we introduce three quantitative metrics to measure clustering quality: the Silhouette Coefficient [9], cluster Compactness and Uniqueness. We use the Euclidean norm as a distance metric, partition extracted CNN features $\{x_i\}_{i=1}^{n}$ into K clusters.

Silhouette coefficient (S) aims to explain the consistency of clustering and quantifies how similar a data point is to its cluster compared with the other clusters [9]. S is defined as

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}}, \tag{1}$$

where $a(x_i)$ is the mean distance between x_i and other $x_j (j \neq i)$ in the same cluster C_l and $b(x_i) = \min_{k \neq l} \frac{1}{|C_k|} \sum_{x_o \in C_k} ||x_i, x_o||_2^2$, i.e. the minimum mean distance to other clusters.

Compactness (D_C) quantifies the inter-class deviation, which evaluates how similar an object is to its own cluster. Here, τ is a scaling factor, $\hat{\mu_k}$ is the centroid of k-th cluster C_k. D_C is defined as

$$D_C = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} \frac{||\mathbf{x}_i - \hat{\mu_k}||_2^2}{\tau}. \tag{2}$$

Uniqueness (D_U) quantifies the intra-class deviation. It is defined to consider both distance and angle separation. We use the cosine similarity to emphasize the direction (of cluster centroid to data mean $\bar{\mathbf{x}}$) and $\hat{\mu}_\mathbf{1} - \bar{\mathbf{x}}$ as a base reference vector. D_U is defined as

$$D_U = \frac{1}{n} \sum_{k=1}^{K} |C_k| \frac{\exp\{||\hat{\mu}_\mathbf{k} - \bar{\mathbf{x}}||_2^2 - \frac{(\hat{\mu}_\mathbf{k} - \bar{\mathbf{x}}) \cdot (\hat{\mu}_\mathbf{1} - \bar{\mathbf{x}})}{||\hat{\mu}_\mathbf{k} - \bar{\mathbf{x}}|| || \hat{\mu}_\mathbf{1} - \bar{\mathbf{x}}||}\}}{\tau}. \tag{3}$$

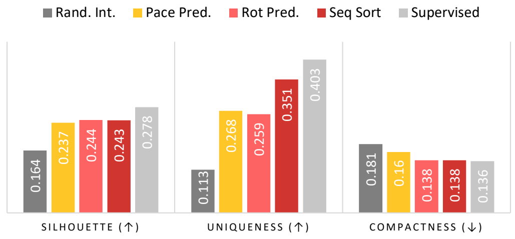

Figure 2: Performance of the proposed metrics

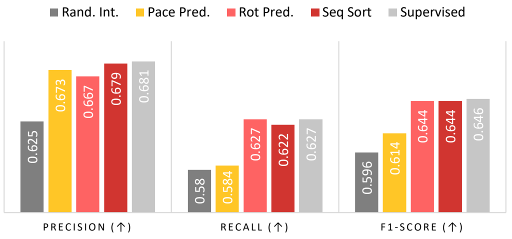

Figure 3: Evaluation results on downstream task

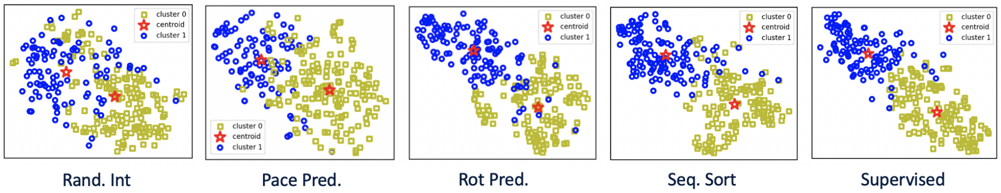

Figure 4: The clustering results illustrated by t-SNE.

3 Experiments and results

3.1 Data

We use a large-scale routine clinical fetal ultrasound dataset (Research Ethics Committee Reference 18/WS/0051) in our experiments. 342 scans (715,968 frames) are used for training self-supervised learning models. These models are trained on a single NVIDIA GTX 1080 Ti GPU. 135 scans are used to train and test the downstream task with three-fold cross-validation. A set of 214,970 clinician-labelled video frames are used to train a supervised anatomy classification model as an upper bound reference model. An additional 122 ultrasound scans are used for independent testing of the proposed metrics.

3.2 Quantitative evaluation

Quantitative evaluation results of the proposed metrics are presented in Fig. 2. A baseline model of random initialization is included as a lower bound, together with a supervised model as an upper bound. Results show the upper and lower bound models have the best (e.g. 0.403 for Uniqueness) and worst (0.113) clustering quality. Among the tested self-supervised learning models, the sequence sorting (Seq. Sort) model has the highest clustering quality (0.351). Therefore, the Seq. Sort model is selected as the recommended approach for representation learning accordingly. Fig. 4 shows the t-SNE plots of clustering results, which align well with the findings above.

3.3 Effectiveness validation

To further validate the effectiveness of the proposed metrics, we adopt the self-supervised learning evaluation practice, by fine-tuning the pre-trained self-supervised models on a downstream task: standard plane detection, a 14-class classification task [6]. Fig. 3 shows that Seq.Sort performs the best (e.g. 0.679 for Precision), while Rand.Int. and Supervised serve the lower (0.625) and upper (0.681) bounds. The results are in agreement with the performance reported in Fig. 2.

4 Conclusion

In this paper, we address the problem of self-supervised representation learning explainability for medical ultrasound video, by proposing explainable quantitative metrics. Experimental results demonstrate that our proposed quantitative metrics work well in explaining the anatomy-aware knowledge captured during representation learning. Although showcased with ultrasound video, the only assumption is that the landmarks are available. Hence this approach is well-suited in other medical imaging applications of self-supervised learning.

Potential social negative impacts

The paper describes a technical solution aimed at advancing explainability of self-supervised learning for medical fetal ultrasound video. The proposed method reveals biologically meaningful patterns and the ability of a CNN to capture anatomy-aware knowledge via representation learning. Although the proposed metrics output statistics that reveal explainability, this is unlikely to be sufficient to communicate with patients alone without any formal clinical explanations. The purpose of the research is to advance knowledge of design of methodology that may aid ML developers, and possibly clinicians and medical device regulators to build trust in a trained model. The utility of the method has yet to be formally evaluated for clinical use.

Acknowledgment The authors would like to thank the support from the ERC Project PULSE ERC-ADG-2015 694581, and the EPSRC Programme Grant Visual AI EP/T028572/1 and the EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1).

References

[1] David Bau et al. “Network dissection: Quantifying interpretability of deep visual representations”. In: *CVPR*. 2017.

[2] Joy Buolamwini and Timnit Gebru. “Gender shades: Intersectional accuracy disparities in commercial gender classification”. In: *Conference on fairness, accountability and transparency*. 2018, pp. 77–91.

[3] Liang Chen et al. “Self-supervised learning for medical image analysis using image context restoration”. In: *Medical image analysis* 58 (2019), p. 101539.

[4] Richard Droste et al. “Discovering Salient Anatomical Landmarks by Predicting Human Gaze”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1711–1714.

[5] Richard Droste et al. “Ultrasound image representation learning by modeling sonographer visual attention”. In: *IPMI*. 2019.

[6] Jianbo Jiao et al. “Self-supervised Contrastive Video-Speech Representation Learning for Ultrasound”. In: *MICCAI*. 2020.

[7] Longlong Jing et al. “Self-supervised spatiotemporal feature learning via video rotation prediction”. In: *arXiv preprint arXiv:1811.11387* (2018).

[8] Hsin-Ying Lee et al. “Unsupervised representation learning by sorting sequences”. In: *ICCV*. 2017.

[9] Peter J Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[10] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *ICCV*. 2017.

[11] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (2013).

[12] Jost Tobias Springenberg et al. “Striving for simplicity: The all convolutional net”. In: *arXiv preprint arXiv:1412.6806* (2014).

[13] Erico Tjoa and Cuntai Guan. “A survey on explainable artificial intelligence (XAI): towards medical XAI”. In: *arXiv preprint arXiv:1907.07374* (2019).

[14] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. “Self-supervised Video Representation Learning by Pace Prediction”. In: *ECCV*. 2020.

[15] Matthew D Zeiler and Rob Fergus. “Visualizing and understanding convolutional networks”. In: *ECCV*. 2014.

[16] Quan-shi Zhang and Song-Chun Zhu. “Visual interpretability for deep learning: a survey”. In: *Frontiers of Information Technology & Electronic Engineering* 19.1 (2018), pp. 27–39.

[17] Quanshi Zhang et al. “Interpreting cnn knowledge via an explanatory graph”. In: *arXiv preprint arXiv:1708.01785* (2017).

[18] Bolei Zhou et al. “Learning deep features for discriminative localization”. In: *CVPR*. 2016.