Using Temporal Similarity in Contrastive Learning for Multi-class Kidney Ultrasound Segmentation

Rohit Singla
School of Biomedical Engineering
University of British Columbia
Vancouver, BC
rsingla@ece.ubc.ca

Cailin Ringstrom
School of Biomedical Engineering
University of British Columbia
Vancouver, BC
ceringstrom@ece.ubc.ca

Victoria Lessoway
School of Biomedical Engineering
University of British Columbia
Vancouver, BC

Janice Reid
School of Biomedical Engineering
University of British Columbia
Vancouver, BC

Christopher Nguan
Urologic Sciences
University of British Columbia
Vancouver, BC

Robert Rohling
Electrical and Computer Engineering
University of British Columbia
Vancouver, BC

Abstract

Creating ground truth segmentations for medical imaging is labour and time intensive. While promising, contemporary contrastive learning techniques commonly overlook the ultrasound domain. We investigate the potential benefits of using ultrasound's real-time trait through different contrastive learning sampling strategies in multi-class semantic segmentation. First, we perform a head-to-head label efficiency comparison between two state of the art algorithms, one for contrastive learning and the other fully supervised, to demonstrate the efficiency gains from contrastive learning. Next, we leverage the notion of temporal coherency which is the notion that frames within an ultrasound cine that are close together share structural similarities. Using data from over 500 patients, our preliminary results indicate that temporal partitioning has potential improvements to the learned embeddings. Future work is needed to investigate the changes to intra-class compactness and inter-class separability for these embeddings, as well as identifying downstream tasks which may benefit the most from temporal coherency.

1 Introduction

Semantic segmentation is a fundamental task in the field of medical image analysis.[1] It involves densely assigning a category to each pixel within the image, providing structured spatial information. Segmentation plays a significant role in visualization, image-guidance for interventions and surgery, clinical decision-making and diagnosis, and quantification of spatial relationships. With modern machine learning approaches, particularly the supervised algorithms, this task requires an abundance of fine-grained annotations to train and evaluate models so that they may learn representations directly from the data itself.[2] In the medical imaging field, these labels are difficult to acquire. They require significant effort to manually annotate, are time-intensive to produce, and require clinical expertise from clinicians or allied health professionals to produce. Even then, inter-rater variability on interpretation is highly varied.[3-6] While efforts to explore alternatives in generating the fine-grained

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

annotations exist, including using crowd-sourcing from novices or active learning techniques, these are pre-dominantly areas of research.[7-9] Alternatives to avoiding the effort of large data sets have focused on novel learning algorithms, such as self-supervised learning or generative adversarial networks, as well as data augmentations.[7] Given this, there remains significant interest in achieving high label efficiency - maintaining high accuracy with few labels.

Contrastive learning, a variant of self-supervised learning, is a promising avenue for this. Generally, a contrastive learning network aims to learn a lower-embedding space wherein the representations of similar samples are close in distance to one another, whereas the representations of dissimilar samples are farther, capturing invariance and covariance. Commonly, the first component frequently involves training a large task-agnostic feature extractor or encoder (with a projection head) using unlabelled images and augmentations of those images as similar pairs. The next component stage involves fine-tuning based on the available labelled data, and yet another component may involve knowledge distillation. Augmentations may include colour augmentation, image rotation and translation, and other geometric transformations. The development of modality-specific augmentations is an active area of interest within medical imaging such as in ultrasound, computed tomography, and magnetic resonance imaging.[10-12]

The application of contrastive learning to semantic segmentation has garnered significant interest in the last few years. Zhao *et al.* used a pixel-wise, label-based contrastive loss for pre-training, Alonso *et al.* utilized a memory bank and contrastive learning module, and Xie *et al.* also use a pixel-wise contrastive loss with the addition of a propagation consistency scheme.[13–15] In the medical imaging sub-field, Pandey *et al.* incorporated a consistency regularization scheme to aid in their contrastive learning segmentation, You *et al.* incorporated a voxel-wise representation, and Chaitanya *et al.* demonstrated the use of contrastive learning at both the global and local scale for volumetric medical images.[16-18] While these techniques all have achieved remarkable results, they commonly overlook one imaging modality: ultrasound.

Ultrasound is real-time, non-ionizing and non-invasive in nature. Despite being the first line imaging modality for numerous organs, including the kidney, the availability of ultrasound data compared to other modalities is scarce. Optimizing our ability to learn from limited samples is even more vital with ultrasound than with other modalities. In particular, unlike computed tomography and magnetic resonance imaging, ultrasound is a video-like modality which has a temporal nature to it. Leveraging this temporal characteristic may enhance contrastive learning-based semantic segmentation.

In this work, we aim to demonstrate the benefit of ultrasound-specific adaptations in the use of contrastive learning for multi-class semantic segmentation. We first compare the label efficiency of contrastive learning compared to a fully supervised network to justify the need to utilize the temporal property. We then systematically incorporate a binary and linear temporal coherency sampling strategy for generating a) positive pairs only and b) both positive and negative pairs. We incorporate this sampling strategy at different stages of the contrastive learning framework to comprehensively explore benefits in learning global and local features.

2 Methods

We adopt the network from Chaitanya *et al.* for volumetric segmentation as our baseline contrastive learning model.[18] This network involves three stages. The first is an pre-training stage where unlabelled images are used to train an encoder connected to a projection head. Augmented versions of a reference image are considered similar positive pairs, while all other images are considered negative pairs with said reference. The Information Noise Contrastive Estimation (InfoNCE) loss is used. The second stage is the pre-training of the decoder which replaced the projection head. The InfoNCE loss is modified to work on regions within the image such that local features are learned. In the third stage, supervised learning using a small amount of labels is used to fine-tune the network.

For binary temporal coherency, we expand the definition of positive pairs to include other images and their augmentations within a set margin from the reference. From within a set range of the reference image, we randomly select images to treat as positive pairs. All other images are considered dissimilar. For linear temporal coherency, rather than a set margin from the reference, images from the reference are weighted on a scale of 0 to 1 of similarity. The closer an image is to the reference, the higher its weighted similarity. Here, we use the traditional contrastive learning loss.

Table 1: Mean DSC across all four classes between nnU-net and baseline contrastive learning network across a range of percentages. At 30%, the contrastive learning network reaches the same mean DSC as the nnU-net.

	Percentage of Labels Used											
	1%	**5%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
nnU-net	0.333	0.375	0.439	0.516	0.521	0.549	0.544	0.565	0.554	0.569	0.557	0.570
Chaitanya	0.276	0.443	0.504	0.563	0.572	0.561	0.582	0.575	0.582	0.565	0.558	0.571

3 Results

We report on two sets of experiments for multi-class segmentation in kidney ultrasound. Kidney ultrasounds were acquired and anonymized from our local institution over five years. Data collection was approved by our institution's Research Ethics Board. We use 514 fine-grained polygon annotations of the kidney, its cortex, medulla, and central echogenic complex. Annotations were generated by two expert sonographers with over 20 years of experience. We additionally use 7000 unlabelled cines with an average of 200 frames each for pre-training. The conventional 80/20 training/testing split was used in fine-tuning of the contrastive learning network, and in the nnU-net. Computation was performed on a single GPU (NVIDIA Tesla V100 32GB) using an internal institutional cluster.

In experiment 1, we trained a nnU-net model [19] using 514 labelled images for segmentation. This is a data-adaptive network that represents the state of the art in segmentation. We compared it to the original Chaitanya network, and report the average Dice Sorenson Coefficient (DSC) across all classes. Both networks were trained with 1%, 10%, 25%, 50%, 75% and 100% of labels. Table 1 reports these results.

In experiment 2, we evaluate binary coherency for both InfoNCE and contrastive loss. We evaluate linear temporal coherency only for contrastive loss. In all cases, we applied partitioning to the encoder only, decoder only, and to both encoder and decoder. All models were evaluated using 10%, 50%, and 100% of labels.

4 Discussion and Conclusion

Our preliminary results indicate that i) contrastive learning may be used in place of fully supervised networks to reduce laborious annotation burden using only 30% of labels to achieve similar results, ii) the use of temporal coherency has mixed gains in improving accuracy at lower percentage of labels, iii) the temporal sampling using hard negatives demonstrates more consistent performance increases. Further work is needed to investigate how best to utilize ultrasound's temporal nature in learning robust embeddings for classes, as well as which downstream tasks may benefit from such techniques.

5 Potential Negative Societal Impact

While this approach may beneficial, there are several potential negative societal effects. The first is that contrastive learning may displace workers who's job function is to generate such detailed annotations. In the medical imaging subfield, there are already a limited number of these workers who have suitable expertise and training. Another potential negative is the use of 'under-trained' algorithms, wherein algorithms may be trained without utilizing all 100% of the labelled data available; doing so may increase the uncertainty of such algorithms. Finally, the sampling strategies proposed remain ignorant to the patient demographics used, and may perpetuate underlying biases hidden in the data; the use of metadata for sampling may be a worthwhile investigation.

References

[1] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Laak, J. A. W. M. van der, Ginneken, B. van and Sánchez, C. I., "A survey on deep learning in medical image analysis," Med Image Anal 42, 60–88 (2017).

[2] Ronneberger, O., Fischer, P. and Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation," MICCAI, 234–241 (2015).

Table 2: Evaluation of temporal coherency sampling strategies using positive pairs only on mean Dice Sorenson Coefficients. The variation that is closest to or higher than the baseline is bolded.

Loss	Variation	Stage	Percentage of Labels		
			10%	**50%**	**100%**
InfoNCE	–	–	0.442	0.547	0.544
InfoNCE	Binary with Positive Pairs	Encoder	0.453	0.512	0.530
InfoNCE	Binary with Positive Pairs	Decoder	**0.467**	0.525	0.541
InfoNCE	Binary with Positive Pairs	Both	0.450	0.503	0.534
Contrastive	Binary with Positive Pairs	Encoder	0.426	0.532	0.553
Contrastive	Binary with Positive Pairs	Decoder	0.435	0.526	**0.542**
Contrastive	Binary with Positive Pairs	Both	0.431	**0.535**	0.534
Contrastive	Linear with Positive Pairs	Encoder	0.413	0.515	0.538
Contrastive	Linear with Positive Pairs	Decoder	0.437	0.510	0.541
Contrastive	Linear with Positive Pairs	Both	0.399	0.505	**0.542**
Contrastive	Binary with Negative Pairs	Encoder	0.445	0.515	0.535
Contrastive	Binary with Negative Pairs	Decoder	0.433	0.522	0.541
Contrastive	Binary with Negative Pairs	Both	**0.460**	**0.530**	0.536
Contrastive	Linear with Negative Pairs	Encoder	0.424	0.512	**0.555**
Contrastive	Linear with Negative Pairs	Decoder	0.420	0.522	0.547
Contrastive	Linear with Negative Pairs	Both	0.428	0.524	0.527

[3] Nir, G., Hor, S., Karimi, D., Fazli, L., Skinnider, B. F., Tavassoli, P., Turbin, D., Villamil, C. F., Wang, G., Wilson, R. S., Iczkowski, K. A., Lucia, M. S., Black, P. C., Abolmaesumi, P., Goldenberg, S. L. and Salcudean, S. E., "Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts," Med Image Anal 50, 167–180 (2018).

[4] Ridge, C. A., Yildirim, A., Boiselle, P. M., Franquet, T., Schaefer-Prokop, C. M., Tack, D., Gevenois, P. A. and Bankier, A. A., "Differentiating between Subsolid and Solid Pulmonary Nodules at CT: Inter- and Intraobserver Agreement between Experienced Thoracic Radiologists," Radiology 278(3), 888–896 (2015).

[5] Sahli, Z. T., Sharma, A. K., Canner, J. K., Karipineni, F., Ali, O., Kawamoto, S., Hang, J., Mathur, A., Ali, S. Z., Zeiger, M. A. and Sheth, S., "TIRADS Interobserver Variability Among Indeterminate Thyroid Nodules: A Single-Institution Study," J Ultras Med 38(7), 1807–1813 (2019).

[6] Dong, Y., Zhou, C., Zhou, J., Yang, Z., Zhang, J. and Zhan, W., "Breast strain elastography: Observer variability in data acquisition and interpretation," Eur J Radiol 101, 157–161 (2018).

[7] Pesteie, M., Abolmaesumi, P. and Rohling, R. N., "Adaptive Augmentation of Medical Data Using Independently Conditional Variational Auto-Encoders," Ieee T Med Imaging 38(12), 2807–2820 (2019).

[8] Heim, E., Roß, T., Seitel, A., März, K., Stieltjes, B., Eisenmann, M., Lebert, J., Metzger, J., Sommer, G., Sauter, A. W., Schwartz, F. R., Termer, A., Wagner, F., Kenngott, H. G. and Maier-Hein, L., "Large-scale medical image annotation with crowd-powered algorithms," J Medical Imaging 5(3), 034002–034002 (2018).

[9] Yang, L., Zhang, Y., Chen, J., Zhang, S. and Chen, D. Z., "Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation," Lect Notes Comput Sc 10435, 399–407 (2017).

[10] Lee, L. H., Gao, Y. and Noble, J. A., "Principled Ultrasound Data Augmentation for Classification of Standard Planes," Arxiv (2021).

[11] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J. and Greenspan, H., "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," Neurocomputing 321, 321–331 (2018).

[12] Hao, R., Namdar, K., Liu, L., Haider, M. A. and Khalvati, F., "A Comprehensive Study of Data Augmentation Strategies for Prostate Cancer Detection in Diffusion-Weighted MRI Using Convolutional Neural Networks," J Digit Imaging, 1–15 (2021).

[13] Zhao, X., Vemulapalli, R., Mansfield, P., Gong, B., Green, B., Shapira, L. and Wu, Y., "Contrastive Learning for Label-Efficient Semantic Segmentation," Arxiv (2020).

[14] Alonso, I., Sabater, A., Ferstl, D., Montesano, L. and Murillo, A. C., "Semi-Supervised Semantic Segmentation with Pixel-Level Contrastive Learning from a Class-wise Memory Bank," Arxiv (2021).

[15] Xie, Z., Lin, Y., Zhang, Z., Cao, Y., Lin, S. and Hu, H., "Propagate Yourself: Exploring Pixel-Level Consistency for Unsupervised Visual Representation Learning," Arxiv (2020).

[16] Pandey, P., Pai, A., Bhatt, N., Das, P., Makharia, G., AP, P. and Mausam., "Contrastive Semi-Supervised Learning for 2D Medical Image Segmentation," Arxiv (2021).

[17] You, C., Zhao, R., Staib, L. and Duncan, J. S., "Momentum Contrastive Voxel-wise Representation Learning for Semi-supervised Volumetric Medical Image Segmentation," Arxiv (2021).

[18] Chaitanya, K., Erdil, E., Karani, N. and Konukoglu, E., "Contrastive learning of global and local features for medical image segmentation with limited annotations," Arxiv (2020).

[19] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]