

Omni-Supervised Learning: Scaling Up to Large Unlabelled Medical Datasets

Ruobing Huang, J. Alison Noble, Ana I. L. Namburete

Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, United Kingdom

Abstract. Two major bottlenecks in increasing algorithmic performance in the field of medical imaging analysis are the typically limited size of datasets and the shortage of expert labels for large datasets. This paper investigates approaches to overcome the latter via *omni-supervised learning*: a special case of semi-supervised learning. Our approach seeks to exploit a small annotated dataset and iteratively increase model performance by scaling up to refine the model using a large set of unlabelled data. By fusing predictions of perturbed inputs, the method generates new training annotations without human intervention. We demonstrate the effectiveness of the proposed framework to localize multiple structures in a 3D US dataset of 4044 fetal brain volumes with an initial expert annotation of just 200 volumes (5% in total) in training. Results show that structure localization error was reduced from 2.07 ± 1.65 mm to 1.76 ± 1.35 mm on the hold-out validation set.

1 Introduction

Recent years have witnessed machine learning revolutionizing the field of computer science. This data-driven technology is capable of processing large-scale datasets and in fact, is nourished by increasing availability of data. Many applications, such as natural language processing and image recognition, have since benefited from this feature as a vast amount of data and annotations are obtainable online via crowd-sourcing. This mechanism is not easily reproduced in the medical field, for two principal reasons. Firstly, there are the relatively limited size of medical datasets; collecting medical data is relatively difficult, costly, and may depend on the morbidity. Sharing/transferring medical datasets is usually restricted due to ethical and/or privacy concerns. Secondly, the shortage of expert annotations for large datasets. In most scenarios, human labelling is tedious and time-consuming. Furthermore, the labelling of medical data can require specialized knowledge and skills. As a result, comprehensive labelling for many medical datasets is infeasible owing to the scarcity and costliness of expert resources.

This paper seeks to overcome the latter by proposing a family of algorithms that scale up from a small annotated dataset to potentially infinite unlabelled data, requiring no additional human intervention. We tackle this omni-supervised learning problem by an iterative training/prediction paradigm that distils knowledge from different models and available data. The accuracy of the framework

is lower-bounded by that of a model solely trained on the small annotated set and can be continuously boosted by automatically generated labels.

We demonstrate this general framework for a localization task on a large 3D US dataset of 4044 fetal brains volumes. Starting from a small labelled subset of 200 volumes (5% of the whole dataset), the framework gradually incorporates unlabelled data and generates labels for the **full** dataset.

2 Related work

Semi-supervised learning is a class of machine learning techniques that falls between unsupervised and supervised learning. It attempts to use unlabelled data to improve the performance of the model trained with a smaller annotated data. Self-training is one type of semi-supervised method in which model prediction is used as ground truth to train a new model. However, a naïve implementation of self-training is meaningless as it provides no information gain. A number of methods have been proposed to address this problem. One of the most intuitive approaches is active learning, where predictions are screened and adjusted by human experts before model retraining [1]. This can greatly reduce the amount of data to be annotated but requires expert resource along the process.

Radosavovic et al. proposed *data distillation* to tackle self-training [2]. It generates annotations for unlabelled data by aggregating predictions of perturbed versions of one example using one trained model. These generated labels are then used to train new models and the process is iterated. It is intuitive, as perturbing data (augmentation) can generate useful information that is known to help training (avoid over-fitting). Meanwhile, it is accepted that averaging predictions from different models outperform the results produced by a single model [3]. This idea is extended to transfer knowledge from trained models (teacher models) to a new model (i.e. student model) which is termed as *model distillation* [4]. It is typically accomplished by first combining predictions of an ensemble of teacher models as soft targets to train new models. It can enhance model performance which suggests model distillation can also extract useful signal to help self-training.

This paper, for the first time, combines model distillation and data distillation to build and evaluate an integrated semi-supervised framework. The framework is applied to tackle a real-world medical imaging problem, addressing the challenges of limited annotated data. This work does not emphasize building a specific, sophisticated base model, but rather on exploring a general method that is readily extended to different applications.

3 Methods

The proposed framework can be summarized as follows: 1) Building a group of teacher/base models and training them with the manually labelled dataset respectively; 2) Perturbing unlabelled data (i.e. via geometric transformations) to generate multiple copies of each image; 3) Applying the trained models on these transformed images; 4) Ensembling predictions from different models and

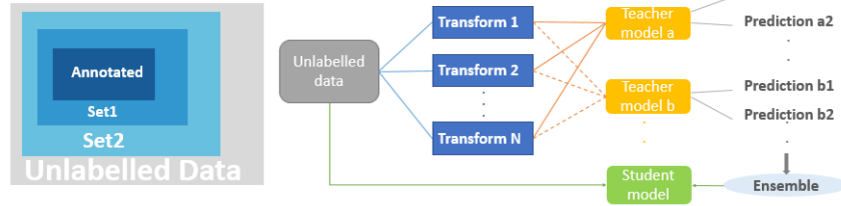


Fig. 1: Schematic of the proposed framework. The figure in the left shows the framework starts from a small annotated subset, to train the base models, and gradually expand to the full unlabelled set. The flow-chart shows unlabelled data is transformed to generate multiple of copies, and are sent to different base models for evaluation. The predictions are aggregated to generate new labels to train student models.

different transformations via weighting or averaging; 5) Training a student model using the mixture of generated and manual labels (see Fig. 1).

One of the key parts of our framework is model distillation. The general principle is that ‘soft targets’ (predicted class probability: $0 \leq p \leq 1$) provide richer information than the ‘hard targets’ (e.g. 0/1 binary score). The predicted probability includes additional information of similarity between inputs and targets thus it is more informative in training new models. Moreover, an ensemble of models is usually superior in classification accuracy than any single component. It suggests different models might be complementary to each other and combining their predictions can be advantageous in self-training. Note that the base models (also referred to as teacher models) can be more complicated, while the student models usually have a compact size to ensure fast inference.

Another contribution of our framework is the incorporation of data distillation. Perturbations of inputs can produce a useful signal for self-training. It does not modify network structures and is simple to implement. Here we also highlight the importance of selecting suitable types of transformations, especially for medical datasets. Later experiments show that certain transformations can be more informative than others in a specific application.

After automatic annotations are generated for the unlabelled data, the results are merged with the initial annotated dataset as the new training set. This new training set can be used to fine-tune the base models or train new student models from scratch. In practice, fine-tuning the model usually encourages faster convergence, but it might be limited when the base model converges to a poor local extrema. We investigate this point further in Sect.4.

4 Experiments on Structure Localization

Clinical task definition We evaluate the proposed method on structure detection in 3D fetal brain neurosonography: a complex task in a challenging imaging modality. A standard fetal 3D neurosonography examination requires identification and evaluation of several key brain anatomies; namely, the lateral ventricles (LV), cavum septi pellucidi (CSP), thalami (Tha), cerebellum (CE), brain stem (BS) and eye (Eye) (Fig.2). Identifying these structures in ultrasound (US) is

non-trivial as: 1) image quality is greatly affected by speckle, skull calcification and the position of the US probe with respect to the brain; 2) developing brain structures change continuously over gestation, both in size and appearance; 3) the position and orientation of the fetal head are highly variable, and commonly observed in reverse positions (see Fig 3). As a result, interpreting a 3D fetal US volume is time-consuming and requires a high-level of expertise. An automatic method to localize brain structures across a large gestational age (GA) range is desirable to lessen the clinical burden of interpreting 3D scans and assist routine evaluation.

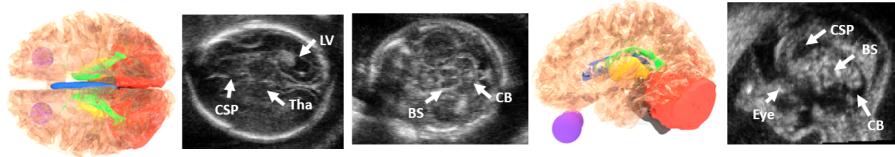


Fig. 2: Key brain structures. Schematics of CSP (blue), LV (green), TH (yellow), CB (red) and BS (grey), Eye (purple) are shown in axial and sagittal views. Examples of the structures shown in an US volume are displayed accordingly.

Datasets 4044 3D US fetal brain volumes were acquired to a standard clinical 3D acquisition protocol and gestational age ranged between 20 and 30 weeks. As the fan-shaped US beam is bordered by a large black region, each volume was cropped at the centre of size $160 \times 160 \times 160$ (with isotropic voxels of $0.6 \times 0.6 \times 0.6 \text{ mm}^3$) and to homogenize the data dimensions. 388 volumes were annotated and separated into a set of 200 to train the initial teacher model, and 188 were held out for validation. The remaining unlabelled 3656 volumes were divided into sets of 403, 811, 2242 respectively for self-training. In total, annotating the 388 volumes took approximately **120** hours of expert time; Manual annotation of the full dataset would require **1251** hours (over 30 weeks of work, given a 40 hrs/weeks work schedule) which is not feasible.

Base model design One of the most popular current methods in object detection is the R-CNN [5] and its variants [6], which consists of a region proposal part (RPP) and a region classification part. The state-of-the-art RPP uses a sliding-window scheme [6] which is not well-suited for fetal neurosonography as scans are taken from different angles. Our task seeks orientation-sensitive predictions instead of axis-aligned boxes. Moreover, as we work with a 3D modality, an exhaustive search for RPP in full 3D space is excessively computationally expensive.

Alternatively, we found the task could be framed as a segmentation problem which can be solved using the well-known 3D U-Net [7], which provides a more straight-forward and unified approach than a two-step framework [5]. As a 3D U-Net supports volumetric input, it is able to incorporate global information that might be hard to obtain in an individual region. Moreover, it provides a denser supervision that enables model distillation. One major obstacle in transfer learning knowledge for localization is that it is usually defined as a regression

problem (to the desired coordinates). A single value of the predicted coordinate does not carry inheritable probability information, which is indispensable in model distillation. Our approach transforms the task into a voxel-wise classification problem that naturally produces probability heat-maps (output voxel values $\in [0, 1]$) for each class. These maps are the soft targets that carry rich information that can be passed on to new models. For simplicity, we consider two base models, namely \mathcal{M}_{CE} and \mathcal{M}_{Dice} , that have identical network architecture but were trained using different loss functions: binary cross-entropy (CE) and dice similarity loss (Dice), respectively.

Multi-Transform Inference Many types of geometric transformation can be used in data distillation, such as cropping, flipping, and rotation. Radosavovic et al. used scaling and horizontal flipping [2] to improve the model. Here we investigate the influence of geometric transformation type in more detail. Specifically, we retrain the model on labels generated using two different groups of transformation: \mathcal{T}_t : flipping and translation. Input volumes were flipped horizontally and vertically. To generate realistic input, each raw volume was translated by $t = -10, 0, 10$ in each orthogonal direction. In total, this resulted in 7 perturbed versions of an input. \mathcal{T}_r : flipping and rotation. Here, input volumes were rotated in the axial and the coronal views by $\pm 10^\circ$. This group also had 7 perturbed versions of each input. For simplicity, predictions were aggregated via averaging across different perturbations and different models for all the experiments.

Implementation details Each 3D U-Net model contained four convolutional (CONV) and down-sampling layers and four CONV and up-sampling layers. The kernel numbers for the first two CONV layers are 16, 32, and 64 for all remaining CONV layers. Kernel size is $3 \times 3 \times 3$ voxels. The feature maps were fed into six sigmoid layers to yield the bounding box masks for each target. Model training was done end-to-end simultaneously via the Adam optimizer with an initial learning rate of 10^{-3} (decayed by a factor of 0.1 every 15 epochs). Batch-normalization, ReLU, and max-pooling were used after each linear CONV layer. On average, a 3D volume was processed in 1.3 secs on an 11GB RAM workstation with one NVIDIA GTX 1080 TI.

5 Results and conclusion

Base models The two base models: \mathcal{M}_{CE} , \mathcal{M}_{Dice} were trained from scratch using 200 manually annotated volumes. The first two rows of Tab. 1 report their accuracy on the held-out validation set. The two models performed similarly in finding the centre of targets. \mathcal{M}_{Dice} outperformed \mathcal{M}_{CE} in IoU metric but scored inferiorly in estimating size. This can be explained as the CE loss identifies the class of each voxel (local evaluation): more voxels are correctly labelled, which leads to more accurate volume estimation. While the Dice loss evaluates the gross overlap (global evaluation) and the IoU metric quantifies this property. Given the complexity of fetal brain anatomy, an ideal model should utilize information on both local appearance and global information. Thus the combination of their predictions should assist training new models.

Transformation type To evaluate the influence of the geometric transformation in data distillation, we compared prediction accuracy of the two student models, \mathcal{T}_t , \mathcal{T}_r , which were learned from labels generated using transform groups \mathcal{T}_t and \mathcal{T}_r , respectively (as defined in Sect. 4). In each case, the training set consisted of 200 manually labelled volumes and 403 automatically annotated volumes. Comparing the performance of \mathcal{T}_t , \mathcal{T}_r with \mathcal{M}_{Dice} , Tab.1 shows localization accuracy was enhanced in both cases for all evaluated metrics. Moreover, \mathcal{T}_r outperformed \mathcal{T}_t . This is as expected as the targets in fetal neurosonography have large orientation variations. Perturbing data rotationally may produce informative signals for the data distillation. This highlights the importance of selecting a data-specific transformation to distil knowledge from unlabelled medical data. We opt for transformation \mathcal{T}_r for all following experiments.

Model	Loss	Training size	Cen Err(mm)	Vol Err(%)	3D IoU (%)
\mathcal{M}_{CE}	CE	200	2.07 ± 1.65	22.8 ± 19.6	57.9 ± 15.2
\mathcal{M}_{Dice}	Dice	200	2.00 ± 1.57	24.3 ± 17.4	59.5 ± 13.6
\mathcal{T}_t	Dice	603	1.96 ± 1.60	17.6 ± 14.4	60.6 ± 15.1
\mathcal{T}_r	Dice	603	1.94 ± 1.56	17.1 ± 13.7	60.9 ± 14.9
Fine tuned	Dice	603	1.96 ± 1.74	17.3 ± 14.1	60.9 ± 15.2
Our method	Dice	603	1.90 ± 1.54	16.2 ± 14.5	61.8 ± 15.1

Table 1: Model performance. The predictions and annotations were compared by evaluating the distance between their centre (Cen Err), the average volume difference (Vol Err) and the 3D Intersection over Union (IoU).

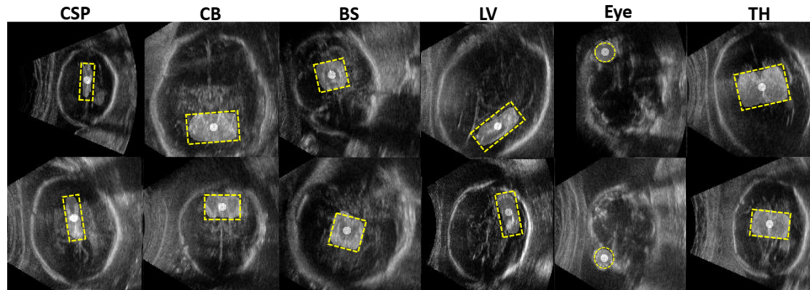


Fig. 3: Target structures viewed on US slices from random subjects (ground truth - yellow box, prediction - transparent overlay). The centre of each structure is plotted as a white dot. The image contrast, fetal head size, and orientation, vary dramatically. Speckle and acoustic shadows also influence structure visibility.

Model re-training To compare fine-tuning with training from scratch, we fine-tuned a model based on \mathcal{M}_{Dice} . The weights of the first two layers were fixed and the initial learning rate was set to be 0.5×10^{-4} . To compare with a model trained from scratch (\mathcal{T}_r), the **Fine tuned** model was trained using the same dataset as \mathcal{T}_r for consistency. Tab. 1 (row 4,5) suggests that retraining from scratch resulted in slightly better performance. This shows local optimum trapping might have

larger effects on model learning, which agrees with the findings reported in [2]. All other experiments are conducted by fully retraining.

Full framework Next, we report the result of the full framework (last row in Tab.1), that combines model distillation (using \mathcal{M}_{CE} and \mathcal{M}_{Dice}), and data distillation (using \mathcal{T}_r). The final model is the best at localizing the six targeted structures. Furthermore, the full framework outperformed the model only using data distillation (row 4, \mathcal{T}_r). This shows that information learned by different models can be effectively combined for self-training.

Model	Training size	Cen Err(mm)	Vol Err(%)	3D IoU (%)
$\mathcal{M}_{Dice}1$	200	2.07 ± 1.65	22.8 ± 19.6	57.9 ± 15.2
$\mathcal{M}_{Dice}2$	603	1.90 ± 1.54	16.2 ± 14.5	61.8 ± 15.1
$\mathcal{M}_{Dice}3$	1414	1.78 ± 1.48	15.8 ± 15.0	62.8 ± 13.6
$\mathcal{M}_{Dice}4$	3856	1.76 ± 1.35	15.6 ± 14.0	62.8 ± 13.0

Table 2: Model performance with increasing training set size. The predictions and annotations were evaluated in the same manner as Tab.1.

Scaling up to the full dataset In a *supervised* setting, deep learning surpasses other machine learning techniques as it can be continuously improved given more training data. Here we show similar results using our *semi-supervised* framework. Visual results refer to Fig. 3. Table 2 shows that model performance scaled with training set size. In Tab. 2, $\mathcal{M}_{Dice}4$ used all the available unlabelled data, and achieved the best accuracy. Compare to $\mathcal{M}_{Dice}1$, it successfully boosted the model performance by nearly 7% in predicting volume size, 5% in 3D IoU, and is 0.7 mm more accurate in centre-point localization on average. It shows that the proposed framework can exploit unlabelled data to benefit subsequent retraining. $\mathcal{M}_{Dice}3$ and $\mathcal{M}_{Dice}4$ had similar mean accuracy while the latter had smaller variance. This suggests the framework performance might saturate given a limited number of base models and data transformation types. While a direct comparison is not possible (dataset is not publicly available), our results are comparable with fully-supervised approaches [8]. Furthermore, our best model $\mathcal{M}_{Dice}4$ outperformed the results reported in the recently published [9], thus adding credibility to our baseline model.

To visually evaluate model performance on the 3856 unlabelled data, we used Procrustes analysis to align the centre of the detected structures thereby registering their US volumes accordingly (see Fig. 4). After alignment, the mean volume corresponds better with the anatomical diagram (all shown in coronal view). It suggests the final model can be extended to build a rigid registration tool and create a brain atlas for further analysis and processing.

To conclude, this paper has presented an original semi-supervised framework that can efficiently scale up to a large medical dataset given a small annotated subset. Validation experiments were carried out on a large 3D US dataset, containing 4044 fetal brain volumes, for a multi-stream localization task. The method has potential to be applied to other tasks to greatly reduce the expert resource required for labelling large-scale medical datasets.

Acknowledgement

We acknowledge the *Intergrowth-21st* study [10] for the image datasets. This work was supported by the National Institutes of Health (NIH) through National Institute on Alcohol Abuse and Alcoholism (NIAAA) (2 U01 AA014809-14), the Royal Academy of Engineering Research Fellowship, and the EPSRC Programme Grant Seebibyte (EP/M013774/1).

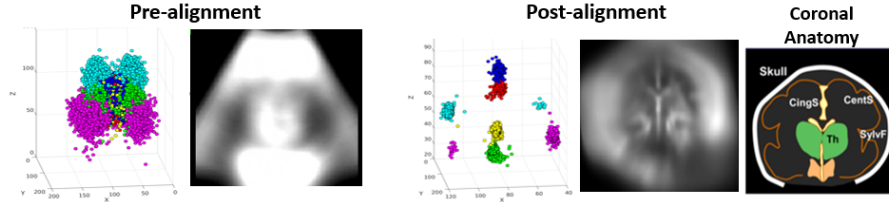


Fig. 4: Rigid registration of all unlabelled volumes using predicted structure locations. Each color in the point clouds represents a structure: blue-CSP, red-TH, yellow-BS, green-CB, cyan-LV, purple-Eye. The mean volumes of the unlabelled volumes before and after alignment are shown accordingly. The figure on the right is a schematic of the fetal brain in the coronal view.

References

1. Y. Gur, M. Moradi, H. Bulu, Y. Guo, C. Compas, and T. Syeda-Mahmood, “Towards an efficient way of building annotated medical image collections for big data studies,” in *MICCAI Workshop*, pp. 87–95, Springer, 2017.
2. I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, “Data distillation: Towards omni-supervised learning,” *arXiv preprint arXiv:1712.04440*, 2017.
3. G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
4. C. Bucila, R. Caruana, and A. Niculescu-Mizil, “Model compression: Making big, slow models practical,” in *Proc. of the 12th International Conf. on Knowledge Discovery and Data Mining (KDD06)*, 2006.
5. R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *CVPR*, 2014.
6. K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 2980–2988, IEEE, 2017.
7. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3d u-net: learning dense volumetric segmentation from sparse annotation,” in *MICCAI*, pp. 424–432, Springer, 2016.
8. M. Sofka, J. Zhang, S. Good, S. K. Zhou, and D. Comaniciu, “Automatic detection and measurement of structures in fetal head ultrasound volumes using sequential estimation and integrated detection network (idn),” *IEEE TMI*, vol. 33, no. 5, pp. 1054–1070, 2014.
9. R. Huang, W. Xie, and J. A. Noble, “Vp-nets: Efficient automatic localization of key brain structures in 3d fetal neurosonography,” *Medical image analysis*, vol. 47, pp. 127–139, 2018.

10. A. T. Papageorgiou *et al.*, “International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the intergrowth-21 st project,” *The Lancet*, vol. 384, no. 9946, pp. 869–879, 2014.