Matching the Clinical Reality: Accurate OCT-Based Diagnosis From Few Labels

V. Melnychuk [1, 2] E. Faerman [2] I. Manakov [2] T. Seidl [2]

[1]Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

[2]Ludwig Maximilian University of Munich, Germany

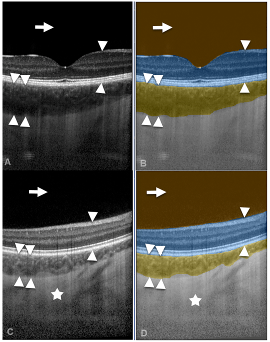
KDAH-CIKM-2020, 19-23 Oct 2020

Table of Contents

Motivation: Few Labels Problem

Supervised learning is difficult to apply in the medical domain:

- high cost of data labelling:
  - requires experts with domain knowledge
  - more fine-grained problem formulations (e.g. volume level vs. slice level) $->$ exponential growth of cost
- epistemic uncertainty: data with high inter-annotator agreement is required [6]



Automated OCT image compartmentalization

Motivation: Usage of Unlabelled Data

Transfer Learning is often used in few labels setting:

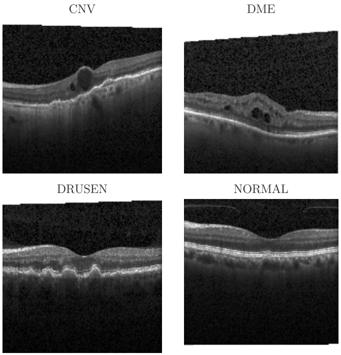
- the possibility of knowledge transfer to medical data is questionable
- ignorance of (abundant) **unlabeled data**

SOTA **Semi-supervised learning** (SSL) algorithms show a promising results on the benchmark datasets $->$ Incentives to employ SSL

Methodology: Dataset

For image classification task we use the **UCSD dataset** published by Kermany et al. [5]:

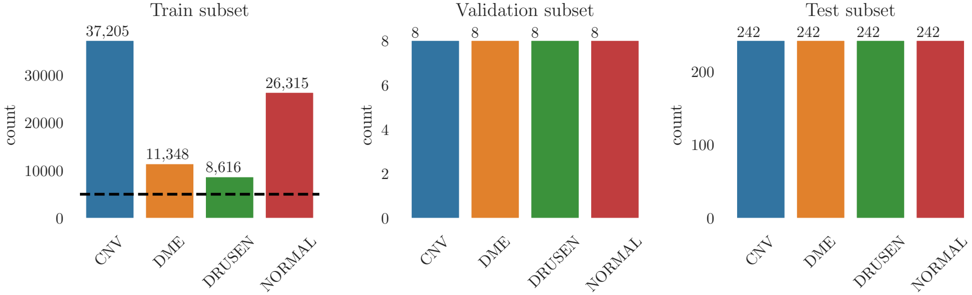
- 84K labelled optical coherence tomography (OCT) b-scans
- 4 classes: "normal", "drusenoid" (DRUSEN), "choroidal neovascularization" (CNV) and "diabetic macular edema" (DME)
- median image size: 496×512 pixels



Sample from UCSD dataset

Methodology: Labelled / Unlabelled Images

Train/validation/test splits are taken from Kaggle. We vary the number of **labelled data**, which we sample randomly and in a balanced way from the training subset



Count plots for dataset split. Dashed line shows labelled-unlabelled data split: upper part = unlabelled subset

Methodology: Realistic Evaluation

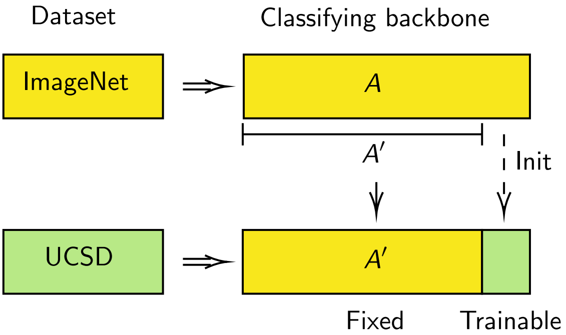
Our work follows the principles of the fair SSL evaluation framework, defined by Oliver et al. [7]:

- the same classifying backbone across all experiments: **Wide ResNet-50-2** [11]
- SSL methods are compared with well-fine-tuned transfer learning / fully-supervised models
- unlabelled data from the same distribution
- realistically small validation subset (32 images)

Experiments: Transfer Learning

We use an ImageNet pre-trained network with 2 settings:

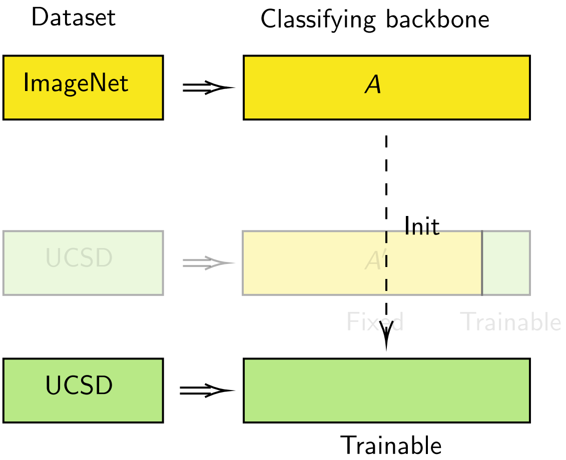
- **Feature extraction**. Freezing all parameters except the last FC layer



Experiments: Transfer Learning

We use an ImageNet pre-trained network with 2 settings:

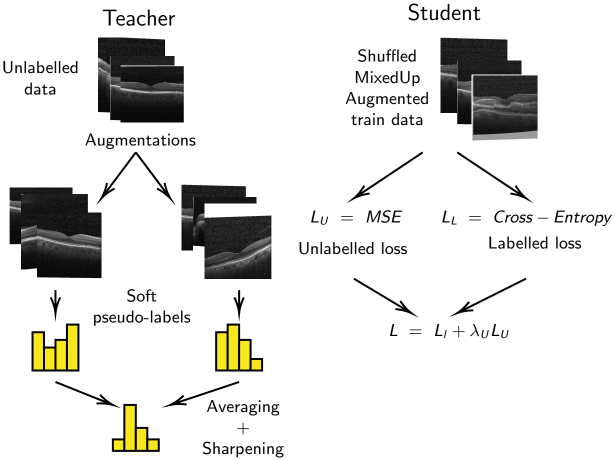
- **Feature extraction**. Freezing all parameters except the last FC layer
- **Fine-tuning**. Using the pre-trained network as the initialization, all parameters are trainable



Experiments: Semi-supervised Learning

MixMatch [2] (2019) – teacher-student architecture:

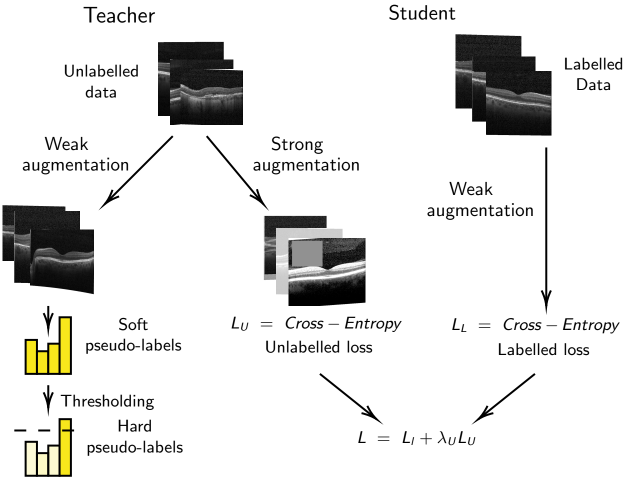
- weak augmentations (flip-and-shift) $->$ consistency regularization
- soft pseudo-labeling of unlabelled augmented data with sharpening
- images and targets Mix-Ups $->$ linear behavior between training samples
- optional improvements: parameters EMA, linear rump-up for λ_U



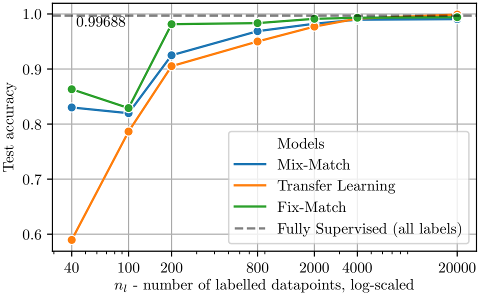
Experiments: Semi-supervised Learning

FixMatch [8] (2020) – teacher-student architecture:

- weak augmentations (flip-and-shift) and strong augmentations (e.g. affine trasformations, color-jittering)
- hard pseudo-labeling of unlabelled weakly-augmented data
- threshold considers only confident pseudo-labels
- parameters EMA



Experiments: Comparison



Best models, test performance after two-fold hyperparameter search

Method	n_l	Accuracy	Notes
Kermany et al. [5]	All	96.6%	Original paper
Alqudah [1]	All	97.1%	Extended UCSD with 5 classes
Wu et al. [10]	All	97.5%	
Chetoui et al. [3]	All	98.46%	
Tsuji et al.[9]	All	99.6%	
WideResNet-50-2 (with EMA)	All	**99.69%**	With EMA decay ($\beta_{\text{EMA}} = 0.999$)
He et al. [4]	835	87.25% *	*Average precision

Reported test accuracies for UCSD dataset

Experiments: Additional Findings

Parameters Exponential Moving Average (EMA) is an inherent part of Fix-Match and an optional for MixMatch:

- we observe learning curves to be more stable for both train and validation subsets
- validation subset is well-chosen $->$ variability could be advantageous
- on UCSD – no obvious advantage of its usage

Transfer learning approaches:

- fine-tuning outperforms feature extraction approach in all label settings
- original models are trained on the dataset with RGB channels $->$ better adaptability to monochrome images in full network fine-tuning

Conclusion

- we demonstrate the efficacy of **MixMatch** and **FixMatch**, when applied to an ophthalmological diagnostic problem on OCT data
- achieving over 80% on as little as 40 labelled samples
- both algorithms outperform transfer learning in the few labelled data settings

References I

[1] A. M. Alqudah.
Aoct-net: a convolutional network automated classification of multiclass retinal diseases using spectral-domain optical coherence tomography images.
Medical & biological engineering & computing, 58(1):41–53, 2020.

[2] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel.
Mixmatch: A holistic approach to semi-supervised learning.
In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019.

[3] M. Chetoui and M. A. Akhloufi.
Deep retinal diseases detection and explainability using oct images.
In *International Conference on Image Analysis and Recognition*, pages 358–366. Springer, 2020.

[4] X. He, L. Fang, H. Rabbani, X. Chen, and Z. Liu.
Retinal optical coherence tomography image classification with label smoothing generative adversarial network.
Neurocomputing, 2020.

[5] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al.
Identifying medical diagnoses and treatable diseases by image-based deep learning.
Cell, 172(5):1122–1131, 2018.

References II

[6] T. A. Lampert, A. Stumpf, and P. Gançarski.
An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation.
IEEE Transactions on Image Processing, 25(6):2557–2572, 2016.

[7] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow.
Realistic evaluation of deep semi-supervised learning algorithms.
In *Advances in neural information processing systems*, pages 3235–3246, 2018.

[8] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel.
Fixmatch: Simplifying semi-supervised learning with consistency and confidence.
ArXiv, abs/2001.07685, 2020.

[9] T. Tsuji, Y. Hirose, K. Fujimori, T. Hirose, A. Oyama, Y. Saikawa, T. Mimura, K. Shiraishi, T. Kobayashi, A. Mizota, et al.
Classification of optical coherence tomography images using a capsule network.
BMC ophthalmology, 20(1):1–9, 2020.

[10] J. Wu, Y. Zhang, J. Wang, J. Zhao, D. Ding, N. Chen, L. Wang, X. Chen, C. Jiang, X. Zou, et al.
Attennet: Deep attention based retinal disease classification in oct images.
In *International Conference on Multimedia Modeling*, pages 565–576. Springer, 2020.

[11] S. Zagoruyko and N. Komodakis.
Wide residual networks.
CoRR, abs/1605.07146, 2016.

References III