Multi-task Knowledge Distillation for Eye Disease Prediction

Our implementation of Knowledge Distillation on Multi-task Learning

Mateusz Sperkowski, Szymon Rećko, Malwina Wojewoda

May 29, 2022

Faculty of Mathematics and Information Science
WARSAW UNIVERSITY OF TECHNOLOGY

Paper Reproduction

“Non-reproducible single occurrences are of no significance to science.”

Karl Popper (1959), “The logic of scientific discovery”, p. 66

Chelaramani, S., Gupta, M., Agarwal, V., Gupta, P., and Habash, R. (2021). Multi-task knowledge distillation for eye disease prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3983–3993

Table of Contents

MTKD

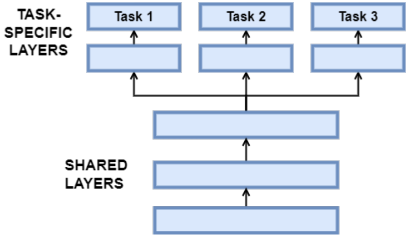

Figure: Multi-task Learning

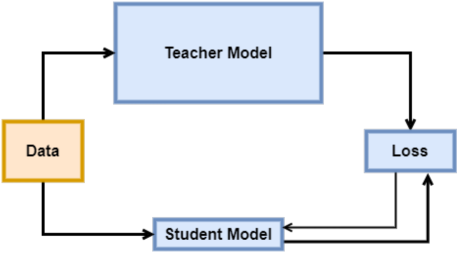

Figure: Knowledge distilation

Pipeline Architecture

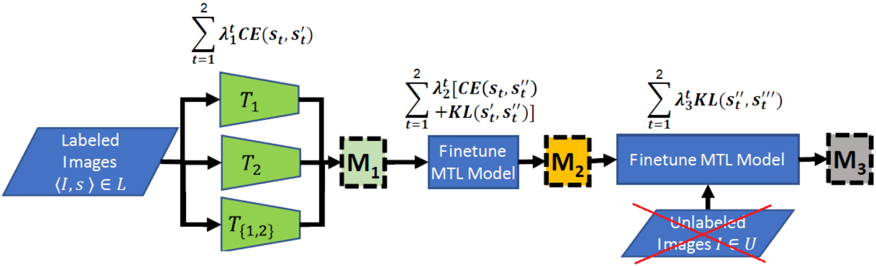

Figure: Training phases of the MTKD Pipeline

What worked and what didn't

- We managed to implement the Pipeline from the paper with reasonable changes.
- We found labeled retinal fundus images similar to those used by the authors, with 2 similar tasks.
- We didn't manage to implement two other tasks.
- We didn't manage to thoroughly experiment on many hyperparameters and pretrained models.
- We had a problem with getting good results on model trained using KD.
- Crucial elements that were not described in the article:
  - The shared layers from pretrained ResNet-50.
  - Weights in the sum of losses for Model 2.
  - Precise method of ensembling the teachers.

Results

	Task 1				Task 2			
	Balanced Accuracy		F1		Balanced Accuracy		F1	
Model	Train	Test	Train	Test	Train	Test	Train	Test
M1[1]	0.911	0.822	0.864	0.715	-	-	-	-
M1[2]	-	-	-	-	0.889	0.811	0.888	0.789
M1[1,2]	0.843	0.758	0.799	0.755	0.874	0.854	0.832	0.772
Ensemble	0.946	0.861	0.920	0.791	0.870	0.837	0.828	0.743
M2	0.857	0.2	0.803	0.291	0.883	0.456	0.821	0.322
M3	0.2	0.197	0.042	0.116	0.385	0.327	0.324	0.291

Table: Results of the trained models

MTKD

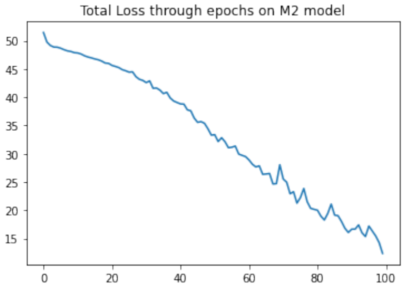

Figure: Loss on M2

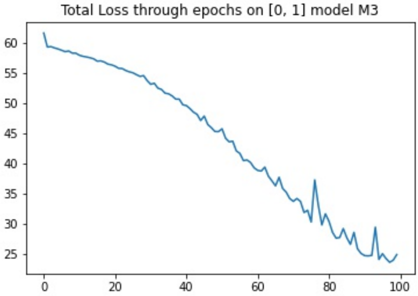

Figure: Loss on M3

Faculty of Mathematics and Information Science
WARSAW UNIVERSITY OF TECHNOLOGY

Pros & Cons of the Architecture

Pros :

- Possibility of using a small labelled dataset together with a larger unlabeled one.
- Potentially less costs connected to labeling a dataset.
- Teacher ensemble method is more effective than just a single teacher for KD.

Cons :

- High computing requirements for all task combinations.
- Large number of hyperparameter choices.
- High requirements for in memory storage of the weights.

Possible future expansions of project

Possible paths for improvements of the pipeline model.

- Expand model to the unused tasks.
- Generalize code for any dataset/model architectures.
- Test the architecture on other hyperparameters.

What we've learned

- Structure & implementation of MTL and KD architectures
- Joining multiple loss functions
- Code and data sharing are crucial for reproduction
- Hyperparameters tuning

Summary

- Reproduction is a crucial part of the scientific method and every scientist should strive to ensure that their paper is reproducible. In case of Computer Science, sharing the code is of upmost importance.
- We successfully analized and implemented a Multi-task Knowledge distillation pipeline.
- It's worth to consider joining multiple transfer learning methods for a single application.

Feel free to send us questions (eg. MS Teams).
Thanks for listening!

References I

Chelaramani, S., Gupta, M., Agarwal, V., Gupta, P., and Habash, R. (2021).
Multi-task knowledge distillation for eye disease prediction.
In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3983–3993.