Deep Learning Based Disease Detection Using Domain Specific Transfer Learning

Steven A. Hicks[3], Pia H. Smedsrud[1], Pål Halvorsen[1, 2, 3], Michael Riegler[1, 2, 3]
[1] Simula Research Laboratory [2] University of Oslo
[3] Simula Metropolitan Center for Digital Engineering

ABSTRACT

In this paper, we present our approach for the Medico Multimedia Task as part of the MediaEval 2018 Benchmark [13]. Our method is based on convolutional neural networks (CNNs), where we compare how fine-tuning, in the context of transfer learning, from different source domains (general versus medical domain) affect classification performance. The preliminary results show that fine-tuning models trained on large and diverse datasets is favorable, even when the model's source domain has little to no resemblance to the new target.

1 INTRODUCTION

In an effort to explore how medical multimedia can be used to create performant and efficient classification algorithms, in the 2018 Multimedia for Medicine Task the participants explore the challenge of automatically detecting diseases found in the gastrointestinal (GI) tract using as little data as possible [13]. The challenge presents four tasks, of which we decided to focus on the task for *classification of diseases and findings* and the task for *fast and efficient classification.*

2 APPROACH

As the current state-of-the-art method for solving most computer vision tasks involves various implementations of deep neural networks, we decided to base our approach on this class of algorithms, specifically CNNs. However, due to the limited size of the development dataset [11, 12], training a CNN from scratch would most likely yield subpar results. Therefore, to resolve this issue, we fine-tune the weights of networks previously trained on larger datasets using the limited data that we have to fit our specific domain (classification of images taken from the GI tract). This technique is commonly referred to as transfer learning (TL), and has been shown to work well across different domains [5, 6, 16].

For this challenge, we hypothesized that adapting the weights of a model trained on data similar to our own (medical images) would yield better results than that of models trained on data with little resemblance, both in terms of time to convergence and classification performance. To test this hypothesis, we compared models trained for the purpose of gaining high-scores on the ImageNet challenge [4] to models trained for medical image classification.

For the classification task, all models were measured by the requirements given, namely matthews correlation coefficient (MCC) and the number of samples used for training. Runs submitted to the efficiency task were evaluated based on their classification throughput, i.e., the time it takes for the model to classify an image.

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

2.1 Transfer Learning From ImageNet

For the approach of fine-tuning models based on ImageNet [4], we simply used the pre-trained networks available in our deep learning libraries of choice, Keras [3] (with a Tensorflow [1] backend) and Pytorch [10]. Both libraries include several popular CNN architectures trained on 1,000 categories containing objects from every day life. As for our method of fine-tuning, we found that simply replacing the classification block, and tuning across the entire network without freezing any layers gave the best results, both in terms of classification performance and time to convergence.

2.2 Transfer Learning From a Medical Dataset

For the medical domain based fine-tuning approach, we trained two models from scratch on a custom medical dataset, consisting of a combination of two openly available medical datasets, LapGyn4 [8] and Cataract-101 [15]. They contain a total of 57,134 images spread across 31 classes, taken from laparoscopy and eye cataract surgeries, respectively. Between this custom dataset and the supplied Medico development dataset, the only overlapping classes are the classes for detecting instruments. Similar to how we trained the ImageNet models, we fine-tuned across the entire network without freezing any layers.

2.3 Additional Training Techniques

In addition to our main hypothesis, we applied various techniques to offset the common issue of overfitting, which can be especially problematic when training on smaller datasets. Techniques used include weighting the loss function based on class size, various data augmentation techniques [17], regularization of the classification block [7, 9], and resampling the dataset by extending the minority class on some base assumptions [2]. First, as the development dataset is small and highly imbalanced with class sizes ranging from 4 to 613 samples, we weighted the network error based on class size. Second, we applied image pre-processing techniques such as rotation, zooming, flipping, shifting and scaling. Third, we applied some L1 and L2 regularization on the classification block of each trained network. Fourth, we observed that the minority class was the only one which did not contain pictures from within the human body. Based on this, we extended the class *out of patient* by adding 14 additional images depicting a typical office environment, including objects such as windows, computers and people to name a few. Note that these techniques were used for all runs.

2.4 Techniques for Efficient Classification

Our approach for the *fast and efficient classification* task, we simply reused the models trained for the classification task to see which models were most efficient. We quickly observed that models implemented in PyTorch [10] had a much higher frames per second (FPS) than their Tensorflow [1] based counterparts, largely due to the difference of how tensors are laid out between the two frameworks. This led to some models being re-implemented in PyTorch and re-evaluated.

3 RESULTS AND ANALYSIS

The initial evaluation of our internal experiments was done using 3-fold cross-validation, where each run was scored by averaging the macro-average classification scores of each model split. A complete overview of the internal runs for both tasks are shown in Table 1. Based on these initial findings, we selected four runs for the *classification of diseases and findings* task (Table 2) and three runs for the *fast and efficient classification* task (Table 3) as official runs to be submitted to the event organizers.

Prioritizing runs for submissions was done by looking at which experiments achieved the highest metric relative to the task at hand (MCC or FPS). Additionally, we wanted to submit a variety of different models, e.g., even though our fine-tuned medical based models did not perform as well as their ImageNet based counterparts, we still wanted to submit a run for official evaluation. For this same reason, we also submitted a model which was trained on a significantly limited development dataset, i.e., a model trained on only 657 samples.

3.1 Classification Subtask Results

Looking at the results for the classification task (Table 2), we see that the best performing run is the *3-Averaged DenseNet169*. This was expected as it constitutes the averaged output of the best performing model from our internal experiments. Furthermore, as shown in our internal runs, the ImageNet based model beats the medical based on by approximately 10% when comparing MCC scores. We believe these results may be due to the difference in variety and size between the two datasets used to train the base models. Due to limited time and resources, we were only able to train a small variety of networks on the medical dataset, and we believe there is more work to be done in this aspect.

Somewhat surprisingly, the submitted model which was trained on a severely limited training set (657 samples), *(Tiny) DenseNet2010*, was still able to retain a relatively high MCC score. We believe this is due to the similarity between images within the same class, and how each class is quite visually distinct (with a few exceptions). This is supported by the confusion matrix shown in Table 4, where we see the model fails on just a few categories.

3.2 Efficiency Subtask Results

Looking at Table 3, we see the official results for the efficiency subtask. Note that all models submitted to this task were implemented in PyTorch. Of the three models, AlexNet was the most performant by quite a large margin. We believe this is due to the networks depth and complexity, i.e., the number of layers and parameters. Additionally, the model's MCC score is relatively high, considering that AlexNet is rather simple compared to models we used for the classification

Internal Classification Evaluation Results

Method	MCC	F1	REC	PREC	SPEC	ACC	FPS
ImageNet Based Transfer Learning							
InceptionResnetV2	0.857	0.858	0.866	0.991	0.851	0.983	31
ResNet50	0.866	0.869	0.874	0.995	0.864	0.991	100
ResNet18	0.866	0.880	0.994	0.995	0.882	0.989	323
AlexNet	0.878	0.885	0.901	0.993	0.880	0.986	**1015**
DenseNet169	**0.915**	**0.922**	**0.931**	**0.995**	**0.918**	**0.991**	**45**
VGG11	0.901	0.908	0.923	0.995	0.905	0.990	624
(Tiny) DenseNet201	0.864	0.876	0.906	0.993	0.873	0.987	58
Medical Based Transfer Learning							
DenseNet169	0.792	0.798	0.830	0.991	0.795	0.983	52
InceptionResnetV2	0.802	0.814	0.843	0.989	0.807	0.979	30

Table 1: The classification performance results of our internal experiments. Note that the displayed metrics are averages across K-splits generated through cross-validation.

Official Classification Evaluation Results

Method	MCC	F1	REC	PREC	SPEC	ACC
ImageNet TL DenseNet169	0.927	0.931	0.931	0.931	0.995	0.991
Medical TL InceptionResNetV2	0.830	0.841	0.841	0.841	0.989	0.980
3-Averaged DenseNet169	**0.935**	**0.939**	**0.939**	**0.939**	**0.996**	**0.992**
Tiny Dataset DenseNet201	0.890	0.897	0.897	0.897	0.993	0.987

Table 2: The official classification performance results as provided by the Medico task organizers.

Official Efficiency Evaluation Results

Method	MCC	F1	REC	PREC	SPEC	ACC	FPS
ResNet18	0.892	0.899	0.899	0.899	0.993	0.987	323
VGG11	0.907	0.913	0.913	0.913	0.994	0.989	624
AlexNet	**0.882**	**0.890**	**0.890**	**0.890**	**0.993**	**0.986**	**1015**

Table 3: The official efficiency results as provided by the Medico task organizers.

Confusion Matrix for 3-Averaged DenseNet169

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	512	0	0	0	1	0	0	0	0	0	4	7	0	0	0	9
B	1	452	84	0	0	0	0	0	0	0	0	0	0	0	0	0
C	1	103	477	0	0	0	0	0	0	0	0	0	0	0	0	0
D	1	0	0	499	41	0	0	0	0	0	2	0	0	0	0	41
E	0	0	0	51	522	0	0	0	0	0	1	0	0	0	0	15
F	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0
G	1	1	2	0	0	0	555	0	0	0	1	0	0	0	0	0
H	0	0	0	0	0	0	0	490	4	0	0	0	0	0	0	0
I	1	0	0	0	0	0	0	0	1961	0	0	0	0	0	0	1
J	1	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0
K	17	0	0	5	0	0	6	0	0	0	357	13	0	6	0	68
L	5	0	0	1	0	0	0	0	0	0	9	564	0	0	0	3
M	1	0	0	0	0	0	0	16	0	0	0	0	1065	0	0	0
N	1	0	0	0	0	0	0	0	0	0	0	0	0	183	1	1
O	0	0	0	0	0	0	0	0	0	0	0	0	0	3	396	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	135

Table 4: (A) Ulcerative colitis; (B) esophagitis; (C) normal z-line; (D) dyed-lifted polyps; (E) dyed resection margins; (F) out of patient; (G) normal pylorus; (H) stool inclusions; (I) stool plenty; (J) blurry nothing; (K) polyps; (L) normal cecum; (M) colon clear; (N) retroflex rectum; (O) retroflex stomach; (P) instruments.

4 CONCLUSION

In this paper, we presented the work done as part of the Medico Multimedia Task where we participated in two of the four available subtasks. Our main hypothesis for this challenge was that fine-tuned models with a medical source domain would perform better than fine-tuned ImageNet models, when used for medical disease detection. Furthermore, with a goal of submitting to the efficiency task, we measured the FPS of the models. Based on our internal experiments and the official evaluation metrics received from the event organizers, we conclude that a large and varied dataset takes precedence over how similar the source domain is to the target. Additionally, we found that networks of lesser depth and complexity were generally more efficient. We admit that these results may be anecdotal, but we believe this requires more research to fully explore the potential of our approach.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015). https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2017. A systematic study of the class imbalance problem in convolutional neural networks. *Computing Research Repository* abs/1710.05381 (2017). arXiv:1710.05381 http://arxiv.org/abs/1710.05381

[3] François Chollet and others. 2015. Keras. https://keras.io. (2015).

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

[5] H. G. Kim, Y. Choi, and Y. M. Ro. 2017. Modality-bridge Transfer Learning for Medical Image Classification. *ArXiv e-prints* (Aug. 2017). arXiv:cs.CV/1708.03111

[6] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. 2018. Do Better ImageNet Models Transfer Better? *Computing Research Repository* abs/1805.08974 (2018). arXiv:1805.08974 http://arxiv.org/abs/1805.08974

[7] Anders Krogh and John A. Hertz. 1992. A Simple Weight Decay Can Improve Generalization. In *Advances in Neural Information Processing Systems 4*, J. E. Moody, S. J. Hanson, and R. P. Lippmann (Eds.). Morgan-Kaufmann, 950–957. http://papers.nips.cc/paper/563-a-simple-weight-decay-can-improve-generalization.pdf

[8] Andreas Leibetseder, Stefan Petscharnig, Manfred Jürgen Primus, Sabrina Kletz, Bernd Münzer, Klaus Schoeffmann, and Jörg Keckstein. 2018. Lapgyn4: A Dataset for 4 Automatic Content Analysis Problems in the Domain of Laparoscopic Gynecology. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys '18)*. ACM, New York, NY, USA, 357–362. https://doi.org/10.1145/3204949.3208127

[9] Pushparaja Murugan and Shanmugasundaram Durairaj. 2017. Regularization and Optimization strategies in Deep Convolutional Neural Network. *Computing Research Repository* abs/1712.04711 (2017). arXiv:1712.04711 http://arxiv.org/abs/1712.04711

[10] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[11] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. Nerthus: A Bowel Preparation Quality Video Dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference.* ACM, 170–174.

[12] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference.* ACM, 164–169.

[13] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Thomas de Lange, Kristin Ranheim Randel, Duc-Tien Dang-Nguyen, Mathias Lux, and Olga Ostroukhova. 2018. Medico Multimedia Task at MediaEval 2018.

[14] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Carsten Griwodz, Thomas de Lange, Kristin Ranheim Randel, Sigrun Losada Eskeland, Duc-Tien Dang-Nguyen, Olga Ostroukhova, Mathias Lux, and Concetto Spampinato. 2017. A Comparison of Deep Learning with Global Features for Gastrointestinal Disease Detection. In *MediaEval*.

[15] Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. 2018. Cataract-101: Video Dataset of 101 Cataract Surgeries. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys '18)*. ACM, New York, NY, USA, 421–425. https://doi.org/10.1145/3204949.3208137

[16] Chuen-Kai Shie, Chung-Hisang Chuang, Chun-Nan Chou, Meng-Hsi Wu, and Edward Y. Chang. 2015. Transfer representation learning for medical image analysis. 2015 (08 2015), 711–714.

[17] Luke Taylor and Geoff Nitschke. 2017. Improving Deep Learning using Generic Data Augmentation. *Computing Research Repository* abs/1708.06020 (2017). arXiv:1708.06020 http://arxiv.org/abs/1708.06020