SCL-UMD at the Medico Task-MediaEval 2017: Transfer learning based Classification of Medical Images

Taruna Agrawal*, Rahul Gupta*, Saurabh Sahu+, Carol Espy Wilson+
+Speech and Communication Lab, University of Maryland, College Park
taruna3@gmail.com,rahul.1987iit@gmail.com,ssahu89@umd.edu,espy@isr.umd.edu

ABSTRACT

Detecting landmarks in medical images can aid medical diagnosis and is a widely researched problem. The Medico task at MediaEval 2017 addresses the problem of detecting gastrointestinal landmarks, keeping into consideration the amount of training data as well as the speed of the detection system. Since medical data is obtained from real-world patients, access to large amounts of data for training the models can be restricted. We therefore focus on a transfer learning approach, where we can borrow image representations yielded by other image classification/detection systems and then train a supervised learning schemes on the available annotated medical data. We borrow the state of the art deep learning classification schemes (VGGNet and Inception-V3 networks) to obtain representations for the medical images and use them in addition to the provided set of features. A joint model trained on all these features yields a Matthew's Correlation Coefficient (MCC) of 0.826 with an accuracy and F1-score values of 0.961 and 0.847, respectively.

KEYWORDS

Image classification, transfer learning, Convolutional Neural Networks.

1 INTRODUCTION

The Medico task addresses the problem of detecting diseases based on image signals from the gastrointestinal (GI) tract. The goal of the task is to advance the application of machine learning tools within the medical domain, while specifically focusing on the detection of GI landmarks from images. Our approach in this task involves leveraging the established frameworks for the detection of real-world objects from images. Specifically, we borrow the state of the art deep learning models in object classification to aid the classification of medical images. Models such as VGGNet [16] and Inception-V3 [19] contain several convolutional, pooling and fully connected layers and are typically trained on large amounts of datasets. Training on these datasets yield models that can capture various geometrical patterns in the input images and translate them into features vectors that are then consumed by the final soft-max layer for class prediction. We aim to harness the capability of such deep networks by retaining the initial convolution filters and pooling layers in these networks. We then obtain the representations yielded by these networks towards the final layers of these networks. This approach can be particularly useful in the cases with limited amount of training data. Since medical domain data is often obtained from real-world patients, training models on limited resources is a requirement. We motivate our approach by discussions of some related work in the next section, followed by the description of the database and the methodology.

2 RELATED WORK

Transfer learning involves borrowing knowledge from related domains to aid classification in a domain of interest [9]. Transfer learning has been successfully applied in tasks such as human behavioral understanding [1, 8], developing deep architectures [2] and autonomous shaping [4]. Several recent works have leveraged advances in image recognition and detection techniques to improve different but related tasks. Shin et al. [15] provide an overview of CNN architectures, data characteristics and transfer learning. Li et al. [7] perform domain adaptation for object localization, using VGGNet. Zheng et al. [21] provide good practices for CNN feature transfer, as we have used in our work. Other applications that have used VGGNet and Inception-V3 based architectures include Alzheimer's disease classification [14], disambiguation for large scene classification [20] and plant classification [6]. The success of these CNN based transfer learning inspire our experiments.

3 DATABASE

We use the dataset provided as part of the *The 2017 Multimedia for Medicine Task (Medico)* task during MediaEval benchmarking initiative 2017 [10, 13]. The dataset consists of 8000 images of the GI tract which are annotated and verified by experienced medical doctors into eight different anatomical landmarks. We use the suggested split of 4000 images as training set and the remaining images as the testing test for the purpose of our experiments. The training dataset contains a balanced number of instances per class. More details regarding the dataset can be found in [13].

4 METHODOLOGY

Deep learning models have achieved state of the art performance in several image classification related tasks. In particular, Convolutional Neural Networks (CNN) have provided the best performances on tasks such as object classification [17], detection [12] and tracking [5]. Inspired from these developments, we obtain a set of features from popular CNN designs, in addition to the provided set of features. First, we describe the set of features used in our experiments, followed by the classification setup.

4.1 Features

We use an assembly of features provided as part of the challenge as well as a few CNN based features. We discuss these features below.

MediaEval'17, 13-15 September 2017, Dublin, Ireland

*Independent authors.

Table 1: Results obtained using various feature sets. Description of the metrics can be found in [13].

Features used	R_k	Accuracy	F1-score
Baseline + Inception-V3	0.816	0.959	0.838
Baseline + VGGNet	0.785	0.953	0.812
Baseline + Inception-V3 + VGGNet	0.826	0.961	0.847

Baseline features The task provides a set of features extracted on the images such as Tamura, ColorLayout, EdgeHistogram and, AutoColorCorrelogram. Each of these features is a global descriptor of the image. Note that these features quantify a specific property of each image, which may or may not be associated with the final classification task. On the other hand, CNN architectures learn to extract features relevant to the task at hand, although it may be hard to interpret those features. More details regarding these features can be found in the task paper [13].

VGGNet based features We use the VGGNet [16] pre-trained on ImageNet dataset [3] as a feature extractor. Since the ImageNet dataset contains a large number of training samples, we expect the VGGNet dataset to be able to model a large variety of shape patterns in the images. We hypothesize that this characteristic of the VGGNet can be useful in the Medico task. We use the 16 layer configuration of VGGNet to predict the outcomes on the ImageNet dataset (configuration D in Table 1, [16]). After this pre-training, we provide the Medico task images as input to the trained VGGNet. Note that each image is of a different size and is rescaled to 244×244 in order to be fed to the VGGNet network. We use the outputs from the first fully connected layer as features for the classification task at hand. The dimensionality of the outputs from the first fully connected layer is 4096.

Inception-V3 features Similar to the VGGNet based features, we extract features from the Inception-V3 network [19]. Inception-V3 consists of a stack of convolutional layers and pooling stacked together and the features we obtain are obtained from the penultimate layer. We again resize the Medico task images to 139×139 pixels. The dimensionality of the features obtained from the penultimate layer is 2048. We next describe the classification setup to predict the anatomical landmarks.

4.2 Classification setup

We evaluate three different classification setups, with different combinations of features. We train a multi-class Support Vector Machine (SVM) classifier on the following combinations:

- Baseline + Inception-V3 features
- Baseline + VGGNet features and
- Baseline + Inception-V3 + VGGNet features

The hyper-parameters for the SVM classifier was tuned using five fold cross-validation framework on the training dataset. We tuned the SVM box-constraint parameter as well as the kernel. The linear kernel performs the best, suggesting that further non-linear transformation of the features is not required. In the next section, we present the obtained results.

Table 2: Confusion matrix for the best performing system using all the features.

Predictions → True class ↓	1	2	3	4	5	6	7	8
1: Polyps	448	13	1	0	0	0	23	33
2: Normal-cecum	22	478	0	0	0	0	0	27
3: Normal-z-line	0	0	427	8	202	0	0	0
4: Normal:pylorus	4	0	5	480	5	0	0	0
5: Esophagitis	0	0	67	10	293	0	0	2
6: Dyed-resection margins	0	0	0	0	0	406	55	1
7: Dyed-lifted-polyps	2	0	0	0	0	94	421	0
8: Ulcerative-colitis	24	9	0	2	0	0	1	437

5 RESULTS

We present our results for each set of features in Table 1. The evaluation metric used in the challenge is a multi-class generalization of Matthew's Correlation Coefficient (R_k). From the results, we observe that the combination including all sets of features performs the best. Since the evaluation metric also takes into account the amount of training data used, we consistently use only 3200 samples out of the 4000 samples for training. We also provide the class-wise confusion matrix in Table 2. We observe that most of the confusion lies between the classes Normal z-line and Esophagitis. We aim to investigate this class confusion in future to reduce the error rate.

In order to further understand the complementarity of the three feature sets used in our experiments, we performed another cross-validation experiments. We used only one out of the baseline, Inception-V3 based and VGGNet based feature sets and evaluate their performance. We observed that the Inception-v3 and VGGNet based features outperform the baseline features, indicating that these CNN based features capture better representation in the images. This may be due to the fact that they are trained on a larger (albeit mismatched) corpus and can model a larger number of geometrical shapes in the images, as compared to the baseline features.

6 CONCLUSION

The Medico task at MediaEval-2017 challenge addresses the problem of detecting GI landscapes from images. The task focuses on training limited amount of dataset with fast evaluation. We address this problem by adopting a transfer learning method, borrowing pre-trained CNN architectures, successfully applied to other image detection and classification problems. We borrow features extracted from VGGNet and Inception-V3 models and train a supervised algorithm along with the provided baseline features. With only 3200 training samples, we obtain an MCC value of 0.826.

In the future, we aim to add more sources of transfer learning to this task. Recently further modifications have been proposed to deep CNN architectures such as GoogLeNet [18], ResNet [12] and generative adversarial networks [11]. We also aim to experiment with ensemble methods to fuse the prediction from these network based features, along with the simple feature fusion in this paper. Since each of the CNN networks carry a different methodology for convolution and pooling, we also aim to understand the discriminative power of each of these feature sets independently. Finally, we also aim to extend the proposed methodology to more image classification tasks within the medical domain.

REFERENCES

[1] Sabyasachee Baruah, Rahul Gupta, and Shrikanth S. Narayanan. 2017. A Knowledge Transfer and Boosting Approach to the Prediction of Affect in Movies. In *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*.

[2] Yoshua Bengio and others. 2009. Learning deep architectures for AI. *Foundations and trends® in Machine Learning* 2, 1 (2009), 1–127.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 248–255.

[4] George Konidaris and Andrew Barto. 2006. Autonomous shaping: Knowledge transfer in reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 489–496.

[5] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernández, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder. 2015. The visual object tracking vot2015 challenge results. In *Proceedings of the IEEE international conference on computer vision workshops*. 1–23.

[6] Sue Han Lee, Yang Loong Chang, Chee Seng Chan, and Paolo Remagnino. 2016. Plant Identification System based on a Convolutional Neural Network for the LifeClef 2016 Plant Classification Task.. In *CLEF (Working Notes)*. 502–510.

[7] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. 2016. Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3512–3520.

[8] Qinyi Luo, Rahul Gupta, and Shrikanth Narayanan. Transfer Learning between Concepts for Human Behavior Modeling: An Application to Sincerity and Deception Prediction.. In *Interspeech, 2017*.

[9] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2010), 1345–1359.

[10] 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, and others. 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.

[11] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[13] Michael Riegler, Konstantin Pogorelov, PÃěl Halvorsen, Carsten Griwodz, Thomas de Lange, Kristin Ranheim Randel, Sigrun Losada Eskeland, Duc-Tien Dang-Nguyen, Mathias Lux, and Concetto Spampinato. Multimedia for Medicine: The Medico Task at MediaEval 2017,. In *MediaEval, 13-15 September 2017, Dublin, Ireland*.

[14] Saman Sarraf, John Anderson, Ghassem Tofighi, and others. 2016. DeepAD: AlzheimerâĂŹ s Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI. *bioRxiv* (2016), 070441.

[15] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging* 35, 5 (2016), 1285–1298.

[16] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[17] Richard Socher, Brody Huval, Bharath Bath, Christopher D Manning, and Andrew Y Ng. 2012. Convolutional-recursive deep learning for 3d object classification. In *Advances in Neural Information Processing Systems*. 656–664.

[18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[20] Limin Wang, Sheng Guo, Weilin Huang, Yuanjun Xiong, and Yu Qiao. 2017. Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs. *IEEE Transactions on Image Processing* 26, 4 (2017), 2055–2068.

[21] Liang Zheng, Yali Zhao, Shengjin Wang, Jingdong Wang, and Qi Tian. 2016. Good practice in CNN feature transfer. *arXiv preprint arXiv:1604.00133* (2016).