ON EVALUATING CNN REPRESENTATIONS FOR LOW RESOURCE MEDICAL IMAGE CLASSIFICATION

Taruna Agrawal, Rahul Gupta, Shrikanth Narayanan

Signal Analysis and Interpretation Lab, University of Southern California

ABSTRACT

Convolutional Neural Networks (CNNs) have revolutionized performances in several machine learning tasks such as image classification, object tracking, and keyword spotting. However, given that they contain a large number of parameters, their direct applicability into low resource tasks is not straightforward. In this work, we experiment with an application of CNN models to gastrointestinal landmark classification with only a few thousands of training samples through transfer learning. As in a standard transfer learning approach, we train CNNs on a large external corpus, followed by representation extraction for the medical images. Finally, a classifier is trained on these CNN representations. However, given that several variants of CNNs exist, the choice of CNN is not obvious. To address this, we develop a novel metric that can be used to predict test performances, given CNN representations on the training set. Not only we demonstrate the superiority of the CNN based transfer learning approach against an assembly of knowledge driven features, but the proposed metric also carries an 87% correlation with the test set performances as obtained using various CNN representations.

Index Terms: Convolutional Neural Networks, medical imaging, transfer learning

1. INTRODUCTION

Recent advances in the design of Convolutional Neural Networks (CNNs) has led to state of the art performances in several tasks, including image classification [1], object detection [2] and object tracking [3]. CNNs can be viewed as models that extract features from raw images using convolution and pooling operations, followed by a classification using fully connected layers [1]. However, training these models typically requires large amount of samples, given the large number of trainable parameters. Transfer learning is a promising approach in such cases, wherein CNN models are pre-trained on larger unrelated corpora, followed by a fine-tuning on the task of interest. The task of interest in this paper is the classification of gastro-intestinal (GI) tract images, given a few hundred training samples per class. A vanilla classification approach in this case would be extraction of a selected set of features, followed by learning a classifier. First, we establish the superiority of transfer learning approach using CNN models learnt on larger unrelated corpora against the vanilla classification approach. However, given that several variants of CNN architectures exist, it is not evident which CNN representation will yield the best performance. To address this, we also propose a metric that can inform the choice of CNN architecture.

Previous work: CNNs have revolutionized research in several fields such as image classification/detection [1], automatic speech recognition [4] and natural language understanding [5]. CNNs typically perform a set of convolution and pooling operations, variants of which have been proposed by several researchers [6, 7]. These variants are typically developed taking into consideration the task at hand. A few examples with custom CNN design include acoustic modeling for low resource languages [8], object and action classification [9] and remote sensing [10]. On the other hand, medical image classification [11] requires assignment of medical images (drawn from real world patients) to a medical landmark, phenomenon or a disease and often, obtaining large amounts of training data can be challenging. A few approaches for medical image classification include the use of decision trees [12], k-nearest-neighbors [13] and support vector machines [14]. Researchers have also applied CNNs for medical image classification using training from scratch [15] as well as transfer learning [16]. In their work, Tajbakhsh et al. [17] address questions regarding the choice between full training versus fine tuning based on empirical performance evaluation. Shin et al. [16] also simplify existing CNN architectures to reduce the number of parameters for training on medical imaging dataset. Recently the MediaEval challenge [18] garnered further interest in medical image classification, with proposals to use CNN based classifiers [19, 20]. All the above papers report performances using a CNN, however they fall short of describing a process that can inform selection of a CNN variant appropriate for the task at hand. First, we obtain performances on the KVASIR dataset using transfer learning based approach. Using these results as an empirical testbed, we propose a metric that can predict performances on the test set. The goal of this metric is to inform the decisions regarding the choice of CNN architecture for transfer learning.

For the task of GI landmark classification, we first establish the performances using two kinds of approaches (i) a kitchen sink feature extraction and classifier training and, (ii) extracting mid-level CNN representations followed by classification layer fine-tuning. We observe that the CNN based transfer learning based approach obtains significantly better test performances on the dataset of interest (described in Section 2) for a majority of CNN variants. However, in real world, choosing a CNN representation based on test performances is not feasible. To address this issue, we propose a metric that can be computed using the training set to predict performance on the test set. We aim for a one shot metric estimation that is robust to the absence of large training sets. We propose a metric whose computation entails projecting the training data-points into a lower dimension, followed by estimation of class confusions in the projected space. Given the various feature representations, the trends predicted by the proposed metric carries a correlation coefficient of 0.87 with the actual test accuracies.

2. DATASET

We use the KVASIR dataset [21] in our experiments. The dataset consists of 8000 images, equally drawn from eight different GI

Table 1. Brief description of features used as a baseline.

Feature : Description
Joint Composite Descriptor: Carries color and texture information in a compressed format
Tamura: Features corresponding to human visual perception: coarseness, directionality, line-likeness, regularity, and roughness
ColorLayour: Spatial distribution of color in the image
EdgeHistogram: Capture edge distribution in the image
AutoColorCorrelogram: Capture color correlation information in the image
Pyramid Histogram of Oriented Gradients: Quantifies spatial layout and local shapes within the image

anatomical landmarks: (i) esophagitis, (ii) normal z-line, (iii) ulcerative-colitis, (iv) normal-pylorus, (v) polyps, (vi) dyed-lifted-polyps, (vii) dyed-resection-margins and, (viii) normal-cesum. The size of these images ranges between 720x576 to 1920x1072, each annotated by a professional endoscopist. In order to perform experiments, we use a training and testing set partition suggested in [18], with 4000 instances in each partition. The objective behind this dataset collection is to aid early discovery of lesions, that can prevent cancer progression. More information regarding the dataset can be obtained from [18, 21].

3. CLASSIFICATION METHODOLOGY

We obtain a representation for each GI image in the KVASIR dataset using two strategies: (i) a baseline kitchen sink feature extraction strategy and, (ii) feature representations obtained using CNNs trained on external corpora. These representations are then used to train a classifier on the available training data. We describe the feature extraction below, followed by the classification setup.

3.1. Baseline: kitchen sink feature extraction strategy

In this strategy, we use an assembly of knowledge driven features (as opposed to the data driven feature representations extracted in CNNs). We use a set of baseline features, as shown in Table 1. These features were motivated by Pogorelov et al. [18, 21] for application to the KVASIR dataset. These features are global descriptors of the images and are designed in a knowledge driven manner to capture a specific property of the images. The dimensionality of baseline feature representation is 1179.

3.2. CNN based feature extraction

In this strategy, we initially train CNN models on an external unrelated dataset, the ImageNet dataset [1]. We then obtain image representations yielded by these networks in the penultimate layer (layer right before the output layer) for the KVASIR dataset. We scale each image in the KVASIR dataset to a size of 224×224, equal to the size of images in the ImageNet dataset. These scaled images are then fed to the CNNs and we apply global average pooling to the outputs of last convolutional layer in each of the CNNs. We test five CNN architectures in our experiments, as described in Table 2.

3.3. Classification setup

After obtaining feature representations for an image, we train a multi-class Support Vector Machine (SVM) classifier for classifying image to one of the eight GI landmark labels. In case of feature representations obtained from CNNs, this training can also be seen as pre-training the CNNs on large datasets and fine tuning the final classification layer using a hinge loss on the KVASIR dataset. The hyper-parameters for the SVM classifier (kernel and box-constraint) are tuned using a five fold inner cross validation on the training set. We present the classification results in the next section.

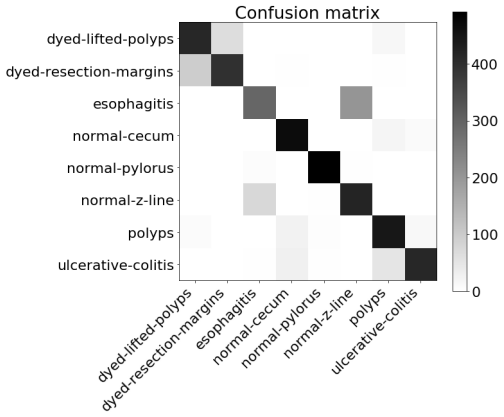

Fig. 1. Confusion matrix for a system trained on representation combined from each source. Bar on the right indicates absolute count in each class.

3.4. Classification Results

Given that the classes are balanced in the training and testing partitions, we use accuracy as our evaluation metric. Table 2 presents the results for each image representation. We also present results for a case where we combine features representations from all the sources. In almost all the cases, representations yielded by CNN outperform the baseline features (except ResNet). This indicates that data driven representations obtained on external corpora can outperform knowledge based features. Research has shown that CNN tends to learn filters sensitive to geometrical patterns observed in the training data [25]. Since the CNNs are initially trained on a large set of images, our results suggest that they learn to encode geometrical patterns, which yield better classification results over knowledge based features. We also observe that the combined model performs the best, indicating that the features from various sources are complementary. We also plot the confusion matrix for the system using combination of all features in Figure 1, as obtained on the testing partition. The confusion matrix is indicative of the classes that the classifier tends to confuse (e.g. we observe a confusion between the dyed-lifted-polyps and dyed-resection-margins class). We refer back to the confusion matrix for futher analysis in Section 4.2.

Although a majority of CNN representations yield better results than the baseline features, the choice amongst the CNN representations is not obvious during system design and training. Therefore, we need a mechanism to assess the discriminative capability of each representation during training. We address this by proposing a metric to estimate the accuracy yielded by a given feature representation in the next section.

4. ESTIMATING ACCURACY YIELDED BY FEATURE REPRESENTATIONS

Given a feature representation (baseline or extracted from a CNN), we propose a metric to estimate the accuracy yielded by a classifier trained on those features. We design the metric such that it could be computed based on the training set. Furthermore, we should be able

Table 2. Features representations used in our experiments. We also present the accuracies for the experiment presented in Section 3.3. Results for VGGNet, Inception-V3, XceptionNet and MobileNet are significantly better than the baseline (binomial proportions test, p-value $< 1\%$)

Features	**Description**	**Feature dimensionality**	**Accuracy**
Baseline	See Table 1	1179	71.6
VGGNet [22]	16 layer architecture, uses 3×3 convolution 2×2 pooling throughout the network.	512	80.1
ResNet50 [7]	50 layer networks with shortcut connections.	2048	61.1
Inception-V3 [23]	Performs convolution with filters of dimensionality 1×1, 2×2 and 3×3	2048	75.6
XceptionNet [24]	Extension of the Inception architecture with standard inception modules replaced by depthwise separable convolutions	2048	80.8
MobileNet [6]	Uses *depthwise separable convolution* to build light weight deep neural networks	1024	81.7
Combined			83.8

Given: Training data $d_n, n = 1, .., N$ with associated feature representations $\boldsymbol{x}_n \in R^D$ and label $y_n \in \{1, .., K\}$ (D is feature dimensionality and K is the number of classes);
Step 1: Obtain transformation $\boldsymbol{z}_i = f(\boldsymbol{x}_i)$, where f is an embedding function for x_i;
Step 2: Modeling class probabilities;
for $k = 1, .., K$ **do**
  Estimate PDF $P_k(\boldsymbol{z}|y = k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$;
  Where $\boldsymbol{\mu}_k = \text{Mean}(\boldsymbol{z}_n)\ \forall \text{d}_n$ with class k;
  $\boldsymbol{\Sigma}_k = \text{Co-variance}(\boldsymbol{z}_n)\ \forall \text{d}_n$ with class k ;
end
Step 3: Estimating accuracies based on the PDF $P_k(\boldsymbol{z}|y = k)$;
for $k = 1, .., K$ **do**
  Estimate accuracy for class k;
  $A_k = \int_{\boldsymbol{z}:P_k(\boldsymbol{z}) > P_{k'(\boldsymbol{z})} \forall k' \neq k} P_k(\boldsymbol{z}) \partial \boldsymbol{z}$
end
Step 4: Final accuracy estimate: $A = \text{mean}(A_1, .., A_K)$

Algorithm 1: Algorithm to estimate the metric A on training data. A is expected to inform the choice of a feature representation.

to obtain it using a one shot computation as opposed to estimation methods such as inner cross-validation on the training set. Such a method requires pre-selection of a classifier, hyper-parameter tuning and is computationally expensive. In particularly, on a small dataset as ours, results could vary from one cross-validation split to the other, leading to a noisy estimate. We outline the computation for the proposed metric in Algorithm 1.

The Algorithm 1 first projects the feature representations from a high dimensional space into a lower dimension space using the projection function f. This is followed by obtaining a Probability Distribution Function (PDF) $P_k(\boldsymbol{z}|y = k)$ of the projected data-points ($\boldsymbol{z}_n = f(\boldsymbol{x}_n)$) based on the class k of their membership. Projecting the data-points on to the lower dimensional space is desirable, as with limited data, the parameters estimation for class specific PDF, $P_k(\boldsymbol{z}|y = k)$, is more robust. We chose $P_k(\boldsymbol{z}|y = k)$ to be a Gaussian distribution. In step 3, based on the estimated class distributions $P_k(\boldsymbol{z}|y = k)$, we compute the probability that a point $\boldsymbol{z}$ sampled from $P_k(\boldsymbol{z}|y = k)$ will yield highest PDF value from the same PDF. We term this estimate as A_k and it is integral of $P_k(\boldsymbol{z}|y = k)$ in the space spanned by $\boldsymbol{z}$ where $P_k(\boldsymbol{z}|y = k) > P'_k(\boldsymbol{z}|y = k'), \forall k \neq k'$. We average the A_k from each class to obtain the final estimate A (we chose averaging since the class distribution is uniform in the training set). We expect the metric A to be indicative of the accuracy obtained when using the feature representation $\boldsymbol{x}$.

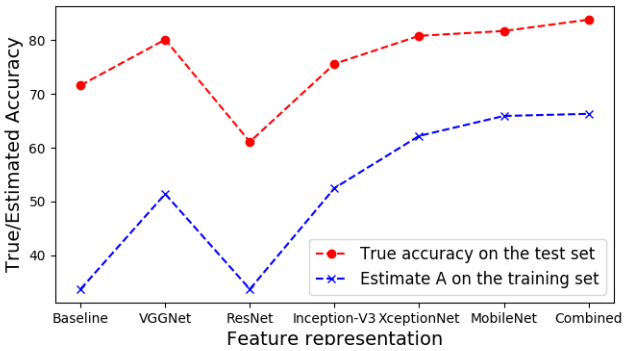

Fig. 2. Plot comparing the accuracies obtained on the test set against the estimates A obtained on each feature representation.

We considered multiple lower dimension projection techniques such as Principal Component Analysis, auto-encoders and t-SNE. Empirically, we observed that the t-SNE projections (in a 2-D space) yield good estimates for $P_k(\boldsymbol{z}|y = k)$ with different $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ values for each class k. $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ estimates using other methods tend to be close, implying a high degree of overlap between class specific distributions $P_k(\boldsymbol{z}|y = k)$ in the projected space. Next, we present our findings on the success of the proposed metric A in predicting the test accuracy.

4.1. Results

Figure 2 plots the accuracies obtained on the test set (as also presented in Table 2) against the estimate A for each feature representation. We note that although that the estimate is off by certain points, the metric A captures the accuracy trend on the testing set. We obtain a correlation of 87% between A and accuracies on the test set. We argue that despite an error in prediction, high correlation with test accuracy is useful as it can inform what feature representation is likely to yield the highest accuracy. We acknowledge that the absolute value of A itself may be off the actual accuracy estimate. Another point to note is that the algorithm to compute the metric A was not informed of the type of classifier (SVM) used in our experiments. Therefore the estimation is performed independent of the final classifier and the associated hyper-parameters.

4.2. Analysis on t-SNE projections

To further analyze the high correlation between the metric A and test performance, Figure 3 presents the t-SNE projections for each feature representation on the training set. The class-wise distribution

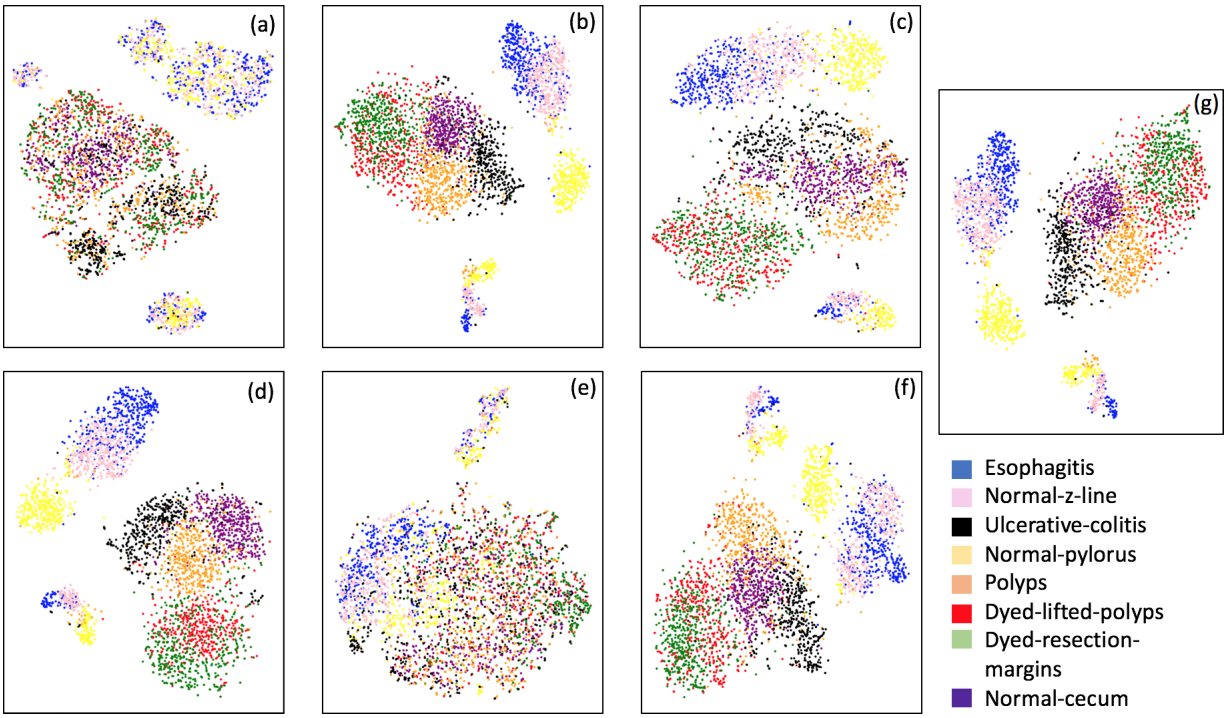

Fig. 3. t-SNE plots obtained using various image feature representations: (a) Baseline features, (b) Inception-V3, (c) VGG-16, (d) MobileNet, (e) ResNet, (f) XceptionNet, (g) All representations combined. We chose a 2-dimensional projection for the ease of visual inspection.

trends in Figure 3 closely associate with the classification performances on the testing set. We observe that in the case of baseline feature representations, different classes gets clustered together in the t-SNE projection plot. The plot suggests that the t-SNE method deems images from different classes to be similar to each other, based on the baseline features. This is coherent with the poor performance observed using the baseline features. On the other hand, class separation is evident in the case of Inception-V3, VGG-16, MobileNet and XceptionNet CNN architectures. The t-SNE representations using ResNet do not cluster as per the eight GI landmark classes, which is in line with the low performance observed using these representations. Overall, the visual trends observed in Figure 3 correspond to the actual test performances, explaining the success of metric A in predicting test accuracies.

Another question we aim to answer is if we can predict the class confusion amongst the eight classes using the t-SNE analysis. t-SNE plots in Figure 3 present promising trends with this regards as well. For instance in the confusion matrix (Figure 1), we observe that the class normal-pylorus has the least confusion with other classes. In the t-SNE plot in Figure 3(g), we observe that this class occurs as a separate cluster by itself. The clusters for three classes: ulcerative-colitis, normal-cecum and polyps, are close to each other, which does reflect as small amount of confusion amongst these three classes. A large confusion is observed between dyed-lifted-polyps and dyed-resection-margins and, esophagitis and normal-z-line classes. The clusters corresponding to these classes have fair amount of overlap in the t-SNE plots. We note that one could obtain a pairwise class confusion metric between two classes k and k' as $A_{k,k'} = \int_{\boldsymbol{z}:P_{k'}(\boldsymbol{z})>P_{k^*}(\boldsymbol{z});k^*\neq k'} P_k(\boldsymbol{z})\partial\boldsymbol{z}$ (we integrate $P_k(\boldsymbol{z})$ over the region where $P_{k'}(\boldsymbol{z})$ dominates). We observed that this metric obtains mediocre performances in predicting class confusions (a correlation between 20% - 50%, depending upon the feature representation). Since the development of this particular metric needs further research, we do not present the detailed results in this paper and consider this as an avenue for future research.

5. CONCLUSION

Several variants of CNNs have been proposed in the past to address problems related to computer vision, speech recognition and natural language understanding. We test their application on a medical imaging problem involving identification of GI landmarks given an image. We use a set of baseline feature representations crafted to capture specific aspects of images as well as feature representations yielded by a set of five different CNN architectures. Classifier trained on four out of five CNN representations outperform the baseline features. Furthermore, we develop a novel metric to inform the choice of CNN architecture for obtaining these representations. We observe that we can foretell the relative performance on the test set by using the proposed metric obtained on the training set. We analyze that the success of the proposed metric stems from a robust lower dimension projection yielded by the t-SNE projections.

In the future, we aim to perform further investigations on the transfer learning approach with CNNs. As of now, we use representations as obtained from the penultimate layer. However, intermediate representations may contain further complementary information. One may also investigate additional low dimensional projection techniques to estimate performance on the testing set. Future work may include decision on classifier design itself based on the t-SNE plots. For instance, a mixture of experts [26] model can be used to distinguish classes using a specific set of features, which otherwise carry a significant overlap in other sets of feature representations.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[3] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernández, T. Vojir, G. Hager, G. Nebehay, and R. Pflugfelder, "The visual object tracking vot2015 challenge results," in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 1–23.

[4] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition." in *Interspeech*, 2013, pp. 3366–3370.

[5] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[6] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] T. Alumäe, S. Tsakalidis, and R. M. Schwartz, "Improved Multilingual Training of Stacked Neural Network Acoustic Models for Low Resource Languages." in *Interspeech*, 2016, pp. 3883–3887.

[9] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[10] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," *arXiv preprint arXiv:1510.00098*, 2015.

[11] M.-L. Antonie, O. R. Zaiane, and A. Coman, "Application of data mining techniques for medical image classification," in *Proceedings of the Second International Conference on Multimedia Data Mining*. Springer-Verlag, 2001, pp. 94–101.

[12] P. Rajendran and M. Madheswaran, "Hybrid medical image classification using association rule mining with decision tree algorithm," *arXiv preprint arXiv:1001.3503*, 2010.

[13] R. Ramteke and Y. K. Monali, "Automatic medical image classification and abnormality detection using K-Nearest Neighbour," *International Journal of Advanced Computer Research*, vol. 2, no. 4, pp. 190–196, 2012.

[14] Y. Fan, D. Shen, and C. Davatzikos, "Classification of structural images via high-dimensional image warping, robust feature extraction, and SVM," *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, pp. 1–8, 2005.

[15] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on*. IEEE, 2014, pp. 844–848.

[16] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[17] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[18] R. Michael, P. Konstantin, H. Paal, G. Carsten, L. Thomas, R. Kristin, E. Sigrun, D. Tien, D. Nguyen, L. Mathias *et al.*, "Multimedia for medicine: the medico task at mediaeval 2017," *MediaEval challenge, Dublin*, 2017.

[19] T. Agrawal, R. Gupta, S. Sahu, and C. E. Wilson, "SCL-UMD at the Medico Task-MediaEval 2017: Transfer learning based Classification of Medical Images," *MediaEval challenge, Dublin*, 2017.

[20] S. Petscharnig, K. Schöffmann, and M. Lux, "An Inception-like CNN Architecture for GI Disease and Anatomical Landmark Classification," *MediaEval challenge, Dublin*, 2017.

[21] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, ser. MMSys'17. New York, NY, USA: ACM, 2017, pp. 164–169. [Online]. Available: http://doi.acm.org/10.1145/3083187.3083212

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[24] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *arXiv preprint arXiv:1610.02357*, 2016.

[25] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196.

[26] R. Gupta, K. Audhkhasi, and S. Narayanan, "A mixture of experts approach towards intelligibility classification of pathological speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1986–1990.