Understanding Scattering Transform and CNN for the Task of Medical Image Classification

Jialing Xu (jx1047), Xiao Li (xl998)
Center for Data Science, New York University
Advisor: Professor Carlos Fernandez-Granda

***Abstract*—End-to-end Convolutional Neural networks (CNN) training can be time consuming and data hungry, in [1], the authors claim that a scattering transform for the inputs may serve as the early layers of CNN and achieve competitive results. The authors also claim that by such way, a hybrid network with scattering transform as early layers and a shallow CNN after could also works on datasets that are short on quantity. We notice that the authors' success is mainly on natural image classification, hence in this work, we use two datasets to investigate if the results could be applied on medical image classification jobs. However, the result we found seems the other way around: on simple tasks of medical image classification where a CNN model don't need to go too deep to get a good result, adding a scattering transform to serve as early layers doesn't give equally good performances compared to plain end-to-end CNN no matter the sample sizes.**

I. Introduction

Scattering transform is mathematically proved to be a non-linear representation that builds invariance to geometric transformations without loss of class discriminability[2]. Because of these properties, it achieves great success on texture classification. Combined with a CNN that doesn't need to be very deep, it could also achieve state of the art solution in natural image classification tasks [1]. Despite its success in natural images, it seems to us that the class difference in medical images is more subtle than in natural images since they are the same objects, some only with mild degree of pathological changes, some are even not discriminable to professionals. Hence in this project, we try to test the efficacy of scattering representation for medical image classification in order to better understand this technique and also in an attempt to find better representation to tack the hard problem of automated medical image diagnosis. We wonder whether the removal of geometric information is actually important to discriminate benign and malignant images. Additionally, in the situation of few training samples as in the case of medical images, the classifier may benefit from using scattering priors as input of a CNN to avoid learning the initial filters[1]. If we don't need to learn the previous layers to be as powerful, this could potentially make deep learning models less data hungry.

II. State of the art

Scattering transform is first defined in [2] as a complex-valued convolutional neural network. In math, consider a signal with finite energy $x \in L^2(R^2)$. Let ϕ_J be a local averaging filter. A family of wavelets $\{\psi_{j,\theta}(u) = \frac{1}{2^{2j}}\psi_{r_{-\theta}\frac{u}{2^j}}\}$is obtained by dilating and rotating a complex mother wavelet, where $-\theta$ is rotation by $-\theta$, $j \geq 0$ is a scaling factor of the wavelet. At each layer, the signal goes through wavelet transform $\{x * \psi_{j_1,\theta_1}(2^{j_1}u)\}$, which builds sparse representation and preserves energy. But it is not translation invariant. In order to build non-trivial invariants across space, a point-wise complex modulus is first applied on the wavelet coefficients then followed by averaging and downsampling. The averaging results in loss

of high frequency resolution and thus a second wavelet transform is applied before that. Therefore for scale $\{j_1, j_2\}$ and rotation $\{\theta_1, \theta_2\}$,the second order scattering coefficient is

$$S^2(x)(j_1, j_2, \theta_1, \theta_2) = |x * \psi_{j_1,\theta_1}| * \psi_{j_2,\theta_2}| * \phi_J(2^J u)$$

The second cascade of wavelet transform recover discrimability by scattering the information of previous wavelet coefficients along another wavelet path. Throughout this work, we only use second order scattering coefficients because the first order coefficient doesn't have high frequency resolution due to averaging, and the energy of higher order scattering coefficients is negligible[1].

This representation as geometric priors can linearize small deformations of images. It is non-expansive and almost complete.[2] This representation followed by a simple classifier like SVM has yielded excellent results on tasks like digit[3] and texture classification[4] where the intra-class variance is well understood to be deformations and translations. But scattering alone is inadequate compared with CNN to capture more complex structure in for example natural images. That's why in [1], a hybrid architecture of scattering transform as input of CNN is proposed. It is a natural combination because in the case of AlexNet, wavelets are often observed in the initial layers[5] which could mean we don't need to learn these filters. This architecture has achieved competitive results with end-to-end learned CNNs on small natural image datasets. Now we want to see if it is still a powerful representation for medical images.

III. Methodology

A. The datasets

1) Dataset 1: This dataset comes from 2 places, E_ophtha[6] and DIARETDB1[7]. In total, there are 460 fundus images, and 232 of them contain at least mild non-proliferative signs (Microaneurysms) of the diabetic retinopathy and the others don't have any lesions (no sighs of diabetic retinopathy). For the images that have Microaneurysms (MA), the dataset also contains corresponding images with annotations points out where the MA are, Fig. 1 and Fig. 9 (in the Appendix) shows an example of this, and Fig. 2 shows an healthy eye without MA. Note that MA are the nearly negligible red dots in the fundus image. Since the dataset size is pretty limited, we use a $40 * 40$ sliding window to cut non-MA (healthy) images uniformly. We also cut MA images using the same sliding window, but use the annotation points as centers and only cut the parts closely surround the annotation points. For cases that a annotation point is less than 20 pixels away from one of the edges, we adapt the edge and cut the part accordingly. In this way, we greatly increase our sample size and end up with 5000 training images and 766 test images with 4536 images without MA and 1230 images with MA in total.

Fig. 1. A fundus image with Microaneurysms

Fig. 2. A healthy fundus's image

2) Dataset 2: The second dataset has 1995 microscopic histopathological images of breast tumor tissue collected from 82 patients via excisional biopsy ($700 * 460$ pixels, 3-channel RGB), with 1370(68.6%) malignant cases. For our experiment we only used images with magnifying factor of 40X. Since for this dataset we don't have annotations to the position of malignant tissues, we experimented with downsampling rates and decided to resize all images to $(60, 60)$. Both benign and malignant cases have tumorous tissues and several subtypes, so we consider it a more difficult classification task than the previous one. Fig. 3 and Fig. 4 give examples of how malignant and benign tissues look like.

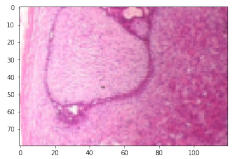

Fig. 3. Benign Tumor

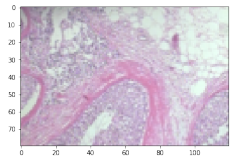

Fig. 4. Malignant Tumor

To compute the scattering coefficient, we used Kymatio[8] which supports fast implementation of scattering transform on GPU. We took its default behavior to only compute frequency increasing path second order coefficients. We construct an end-to-end CNN model (Model 1) and a hybrid model combining scattering coefficients and CNN (Model 2). We applied common image preprocessing steps like normalizing and went through data augmentation like random crop and flip before putting into these models. We ran the training process five times then computed the average and standard deviation on the disjoint test set. Since in [1], the authors claim that "the early layers of CNNs do not necessarily need to be learned, and can be replaced with a scattering network instead", we expect that with fewer layers in Model 1, it could achieve similarly competitive results compared to Model 2. And experiments in [1] also show that using scattering transform serving as early layers, the results could be superior than end-to-end CNN models in limited sample cases. Hence we also compare the result between two models on different sample sizes.

For evaluation metrics, most natural image classification jobs use accuracy as their first choice because it is straightforward; However, since we are doing medical image binary classification with imbalanced classes , accuracy alone is not convictive, hence we also look at recall and AUC as evaluation measure.

IV. Results

We provide our results: test accuracy, auc score and recall score on two datasets for both models to compare the performances as we increase the number of training samples. We test both models on two settings: $2+1$ (2 full convolution layers and 1 fully connected layer) and $1+1$ (1 full convolution layer and 1 fully connected layer), here a full convolution layer is defined as:

Conv2d layer	kernel size: (3 * 3), stride: 1
Relu layer	
Pooling layer	kernel size: (2 * 2), stride: 2
Batch normalization	

Fig. 5 shows the results for Dataset 1, we could see that nearly in every case, the hybrid model performs worse than the end-to-end model. In the $1+1$ structure comparision, this

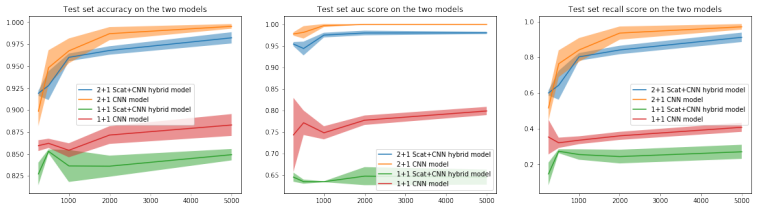

Fig. 5. Dataset 1 results

Since the Dataset 1 is relatively simple, there is not much room of improvement for both models. Hence we look at another dataset that is slightly more complicated. Fig. 6 shows the results for the $2+1$ setting on Dataset 2, we could see that the hybrid model only has a slight advantage on the recall score when the sample size is small, but for all the other cases, end-to-end CNN performs better.

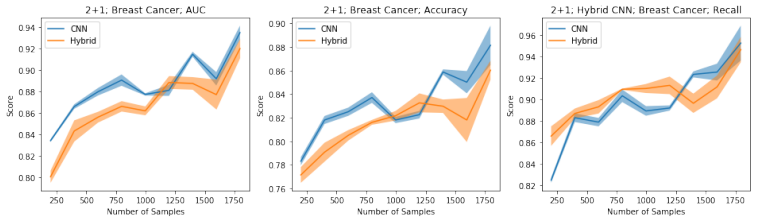

Fig. 6. 2+1 Preformance on breast cancer

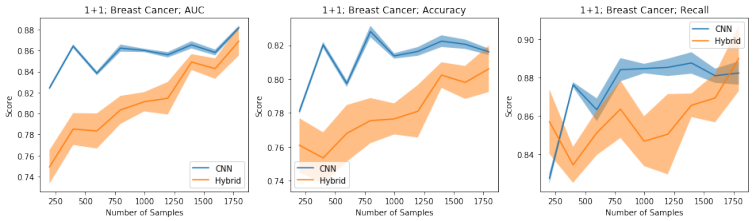

Fig. 7. 1+1 Preformance on breast cancer

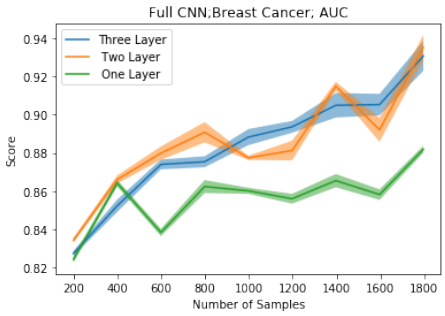

Fig. 8. Adding layers to CNN

In the 1+1 setting, using scattering representation in small sample regime yields high variance on test set but high training accuracy, which indicates overfitting. The results didn't make huge difference when we experimented with choices of hyperparameters. Also in one layer, we noticed that full convolution neural network maintains better performance and smaller variance. Therefore, based on these results, we failed to reproduce the phenomena described in [1].

We conclude that using scattering transform as early layers doesn't give equally promising results as end-to-end CNN. This could be due to the following reasons:

- Scattering transform could be seen as a scale-invariant feature extractor. This works great on natural image classification jobs due to obvious reasons: a cat, no matter how small it is shown in an image, it is still a cat. But things are different in medical images. For example in Dataset 1, Microaneurysms are the tiny red dots in the background, but another big red dot could be a normal structure on the blood vessel. Hence, when the input first goes through the scattering transform, it loses some scale relevant information. And when it goes to the pooling layer of CNN, the loss of scale information intensifies which may affect the prediction of the hybrid model.
- By using scattering transform, the number of features that we use as CNN input is significantly increased (for example from 34560 to 164025 on Dataset 2). This potentially adds to model complexity that makes a simple one layer CNN fails. As shown in 7, the hybrid model is not at all robust with less samples. Instead, it needs much more data to catch up with end-to-end CNN performance, which means the redundant representation is not relevant and necessary for this task.
- Though the high frequency resolution is partly recovered by the second cascade of wavelets, we are still losing information when the input goes through the final local averaging filter followed by a downsampling of scale $2^2 = 4$ [1] . Since we made our input image sizes relatively small in both datasets to be (40, 40) and (60, 60) for computational convenience, the averaging filter and downsampling may be too much for the inputs which make them hard to maintain enough information to be classified and hence hurt the prediction results.

V. Discussion

- For both data sets we choose, we found CNN does not need to go very deep to achieve high score. For example in8 if we only look at the AUC of end to end CNN as we add layers, the performance of only two layers is approximately as good as using three layers though not as stable. In fact for Microaneurysms detection task, the performance of SVM is just as competitive as neural networks. Therefore, the relative failure of hybrid net could be due to the ease of task. If the data is already quite separable by simple classifiers, scattering representation could be counterproductive.
- We suspect that scattering transform is a bad representation for medical images because the variability it

reduces may be interclass-wise. But we haven't find adversarial examples to showcase that due to lack of domain knowledge. Further exploration into this question may help us better understand what classification tasks are suitable for the use of this technique.

- We think there is a possibility that the local averaging filters applied when calculating scattering coefficients could cause loss of information and this loss of information may be essential to dataset with small-size images. At the same time, adding local average filter is also a way to achieve translation invariance, and we don't know whether translation invariance is a good thing to medical image classification. Hence in the future, we want to find a way to calculate the scattering coefficients without cascading local averaging filter and see if this approach would make any change to the prediction results.

VI. Appendix

Fig. 9. The corresponding annotation for Fig. 1

References

[1] Edouard Oyallon, Sergey Zagoruyko, Gabriel Huang, Nikos Komodakis, Simon Lacoste-Julien, Matthew Blaschko, Eugene Belilovsky. *Scattering Networks for Hybrid Representation Learning*. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 2018

[2] Stéphane Mallat. *Group Invariant Scattering*. Communications on Pure and Applied Mathematics, 2012

[3] Joan Bruna and Stéphane Mallat. *Invariant scattering convolution net- works*. IEEE transactions on pattern analysis and machine intelligence, 2013

[4] Joan Bruna and Stéphane Mallat. *Classification with scattering opera- tors*. Computer Vision and Pattern Recognition (CVPR), 2011

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. *ImageNet classification with deep convolutional neural networks*. Advances in neural information processing systems, 2012

[6] Decencire E, et al. TeleOphta *Machine learning and image processing methods for teleophthalmology.*. IRBM (2013), http://dx.doi.org/10.1016/j.irbm.2013.01.010

[7] Tomi Kauppi, Valentina Kalesnykiene, Joni-Kristian Kamarainen, Lasse Lensu, Iiris Sorri, Asta Raninen, Raija Voutilainen, Juhani Pietil, Heikki Klviinen, and Hannu Uusitalo. Machine Vision and Pattern Recognition Research Group Laboratory of Information Processing Lappeenranta University of Technology. http://www.it.lut.fi/project/imageret/diaretdb1/

[8] Andreux M., Angles T., Exarchakis G., Leonarduzzi R., Rochette G., Thiry L., Zarka J., Mallat S., Andn J., Belilovsky E., Bruna J., Lostanlen V., Hirn M. J., Oyallon E., Zhang S., Cella C., Eickenberg M. (2019). *Kymatio: Scattering Transforms in Python. arXiv preprint* arXiv:1812.11214.