Wavelet Scattering Transforms vs Residual Network : A Case Study on Fashion-MNIST

Samuel Cahyawijaya, Etsuko Ishii, Ziwei Ji, Ye Jin Bang
{scahyawijaya, eishii, zjiad, yjbang}@connect.ust.hk

MATH6380P : Mini Project 1. Feature Extraction and Transfer Learning

Introduction

Wavelet Scattering Transform or simply called Scattering is a wavelet-based feature extraction technique that produces a translation-invariant feature set from a given signal. In this experiment, the effectiveness of the Scattering approach is compared to a well-known deep neural network architecture called Residual Network (ResNet). As a case study, the two approaches will be compared in the same standard image classification dataset, Fashion-MNIST. The two approaches are compared with two different analyses, firstly by comparing the feature visualization of both approached to show the separability of the features across different classes, and secondly by comparing the evaluation performance of different models which are built from the corresponding feature set. For the visualization, several visualization approaches are incorporated, including PCA and t-SNE. For the modeling part, from different of models are considered, which are k-Nearest Neighbour (kNN), Support Vector Machine (SVM), Gradient Boosting Machine (GBM). The models are compared via four different metrics: accuracy, recall, precision, and F1 score. Further ablation study is also provided to show a more comprehensive analysis of the experiments.

Fashion-MNIST

Fashion-MNIST [4] is a dataset based on front look thumbnail images of assortment on Zalando's website, which is designed to be a direct drop-in replacement for the original MNIST dataset. The dataset consists of 28 × 28 grayscale images of 70,000 fashion products from 10 categories. There are 7,000 images for each category. Since the official datasets are divided into the training set of 60,000 images and the test set of 10,000 images, we split the training set into two parts, namely, 90% for training and the rest 10% for validation.

0	1	2	3	4	5	6	7	8	9
T-shirt	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Boot

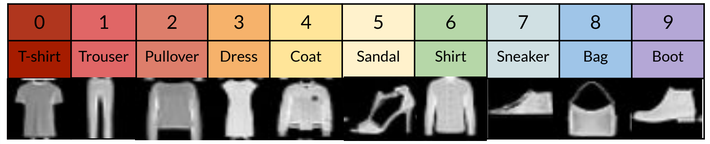

Table1: 10 classes of Fashion-MNIST

Methodology

For feature extraction methods, we adopt two deep convolutional architectures: the **Scattering** and **ResNet18**.

Scattering: The scattering net is a non-trainable architecture composed of convolutional layers whose filters are designed to be wavelet and lowpass filters with additional nonlinearity [1] to obtain an appropriate representation of signal data. We used python library kymatio [5] for implementation with J=1, 2, 4, where J denotes the maximum log-scale of the scattering transform. We call Scattering J1 if we use J=1.

Pre-trained Deep Neural Networks: Today, deep convolutional networks trained on large image datasets are commonly used as a feature extraction method on downstream tasks. In this report, we adopt pretrained ResNet18 [2] on ImageNet provided by PyTorch. We extracted 512-dimensional vectors right before the final fully-connected layer as features. In addition, we try to get label predictions by replacing the final fully-connected (FC) layer to be 2 FC layers with an output size of 64 in the first layer and a ReLU layer followed. We first trained these last 2 FC layers with Adam optimizer [3] with lr=1e-4 while other weights remain fixed for 10 epochs. Then, we fine-tuned all the weights including weights in ResNet18 with Adam optimizer lr=2.5e-5 for 10 epochs.

Feature Visualization

We visualize two different features sets, one from Scattering with J=1 and pretrained ResNet18 fine-tuned in Fashion-MNIST dataset. We visualize the features with 3 different dimensionality reduction methods, PCA and t-SNE. Since PCA does not visualize the distinctions among features from different classes, we display visualization with t-SNE only in this poster. The visualization with PCA can be found in github link.

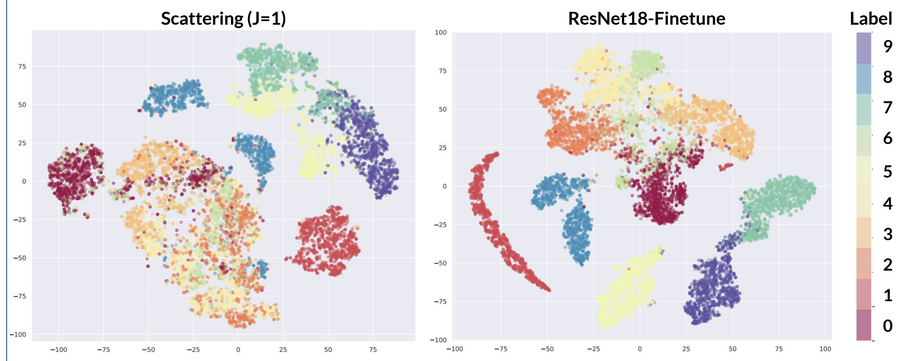

Fig1. Visualization of Scattering (J=1) (left) and ResNet18-finetune (right)

Based on the visualization, scattering can perform similarly compared to pretrained ResNet18, but further fine-tune of ResNet18 on Fashion-MNIST dataset provides a better separation feature set across different class. This fine-tuning approach provides a major advantage of Deep Neural Network over Wavelet Scattering. Further analysis is covered under ablation study.

Features \ Models	SVM		KNN (k=7)		GBM		Fine-tune	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Resnet18-finetune	92.71%	92.72%	93.18%	93.17%	92.42%	92.42%	**93.22%**	**93.22%**
scattering (J=1)	89.02%	88.96%	86.56%	86.48%	**90.42%**	**90.36%**	-	-

▲ **Table2:** Performance comparison between fine-tuned ResNet18 with 2 FC-layers of classification head and Scattering(J=1) by using different classification methods.

ResNet18	Acc.	F1
Without finetune	92.14%	92.18%
Finetune-2-layers	**93.22%**	**93.22%**

◀ **Table3:** Performances of ResNet18 *without vs. with* fine-tuning 2 FC-layers of classification head.

Scattering	GBM	
	Acc.	F1
J=1	**90.42%**	**90.36%**
J=2	89.08%	89.01%
J=4	86.53%	86.48%

▶ **Table 4:** Performances of Scattering with different J classified by GBM

Evaluation Performance

Based on the Table 2, the performance of models trained on features extracted from *Scattering (J=1)* can perform as good as models trained on feature extracted from pretrained *ResNet18* - only **2.8%** difference in accuracy. ResNet18 features showed the best performance with accuracy of **93.22%** while Scattering shows accuracy of **90.42%**. Clearly, fine-tuned ResNet18 with 2 FC-layers of classification head performs better compared to other models which again suggests the effectiveness of deep trainable convolutional networks over non-trainable Scattering.

Looking at breakdown performances on each class, for both Scattering J1 and fine-tuned ResNet18, label 1 (Trouser) gets the highest F1 (97.93 & 98.74%) while label 6 (Shirt) gets the lowest scores (62.78 & 77.49%), label 2 (Pullover: 77.60 & 90.57%) and label4 (Coat: 77.17 & 88.75%) get relative low scores. This is because Shirt, Pullover, and Coat belong to the same category of "tops" and look relatively similar, yet Trouser is distinct from any other classes. This distinction of the samples is clearly shown in Fig1 as well.

Ablation Study

Clusters among different classes

From the visualization, the feature maps can be grouped into 4 different groups, which are Shoes, Bottoms, Tops, and Bags. **Shoes** group covers 3 classes: *Sandal, Sneaker*, and *Boot*. **Bottoms** group covers only *Trouser* class. **Tops** group covers 4 classes: *T-shirt, Pullover, Dress*, and *Coat*. **Bags** group covers only *Bag* class. The grouping of samples is shown in Fig2.

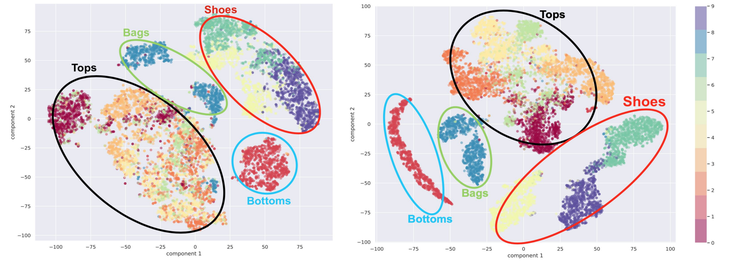

Fig2. Visualization of Scattering (J=1) (left) and ResNet18-finetune (right) with groups

From both feature visualizations, we can see that **Bags** class is separated into two different clusters. The visualizations of several samples on those two clusters are shown in Fig3. From the figure, we can conclude that the two clusters are formed due to the visualization of the bag strap, where in the one cluster the bag either has no strap or the strap is not hanging on the top, while on the other cluster, the bag definitely has a strap and its hanging on top of the bag.

Fig3. Examples of two types of **Bags**. Top row consists of samples from one cluster, while bottom row consist of samples from the other cluster.

Another interesting finding is analyzed from the **Tops** category, where the classes inside that category get the worst result. This is due to the some samples on those classes tends to be have very similar extracted features from one to another as depicted in the Fig2.

Without vs. With fine-tuning 2 FC-layers of classification head for ResNet18

From Table 1, we can conclude that Fine-tuned ResNet18 can consistently outperform Scattering. While compared to ResNet18 w/o finetuning, as shown in Table 3, ResNet18 with fine-tuning yields higher metric score with absolute 1.08% higher in accuracy and 1.04% in F1.

Scattering with different J = {1,2,4}

In Scattering, we experiment with different settings of maximum log-scale of the scattering transform as shown in Table 4. From our experiment, we can conclude that as J increases, the performances get worse. It loses 3.89% in accuracy as increase the value of J from 1 to 4.

Conclusion

In conclusion, the results suggest that Scattering with fixed weights works as good as the trainable deep convolutional network pretrained on large image datasets. Nevertheless, the fine-tuning capability of the pre-trained deep convolutional network allows the model to achieve a significantly better result on the specified task, as is able to get more complicated features than simple wavelet or lowpass filters.

Reference

1. S. Mallat. Group invariant scattering. *Comm. Pure Appl. Math., 65(10):1331–1398*, 2012.
2. K. He, X. Zhang, S. Ren and J. Sun. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV,* 2016.
3. D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*. 2014.
4. H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv:1708.07747*. 2014.
5. M. Andreux, T. Angles, G. Exarchakis, R. Leonarduzzi, G. Rochette, L. Thiry, J. Zarka, S. Mallat, J. Andén, E. Belilovsky, J. Bruna, V. Lostanlen, M. J. Hirn, E. Oyallon, S. Zhang, C. Cella, M. Eickenberg. Kymatio: Scattering Transforms in Python. arXiv:1812.11214v2. 2019.

Github: https://github.com/SamuelCahyawijaya/MATH6380P