Project final report: Clothing classification using U-Net with DCE and PL loss

Chang Liu
Michigan State University
liucha39@msu.edu

Jun Guo
Michigan State University
guojun2@msu.edu

Xingchi Zhou
Michigan State University
zhouxin6@msu.edu

Yiwen Hu
Michigan State University
huyiwen3@msu.edu

Abstract

The rapid development of online shopping has brought tremendous convenience to our daily life. People can buy food, electronics, and clothing on a shopping website. Hence, the number of product images is becoming more and more, and the categories of product images are becoming more and more complex. How to classify the images into different categories with high accuracy in a short time is important because it is related to the interests of merchants and the shopping experience of customers. In this project, we focus on visual clothing analysis, which has attracted lots of interests in computer vision in recent years. we learnt that Convolutional neural networks (CNNs) have been widely used for image classification with high accuracies. However, some research work also shown that CNN can be easily fooled by some adversarial examples and not robust enough for pattern classification. So we leverage the availability of fashion datasets and do the clothing image classification by implementing a convolutional prototype learning framework using U-Net with Distance based cross entropy loss (i.e., DCE) function and prototype loss (i.e., PL) function which can improve the robustness. Specifically, We first processed the data, then read the papers about U-Net and some loss functions. After that we wrote the code of U-Net with the loss function (i.e., DCE and PL) to do the clothing classification. Based on the result evaluation, we also discussed the advantages and limitations.

1. Introduction

Electronic commerce, the activity of buying or selling products over the Internet, brings great convenience to our lives. People can buy food, electronics, clothing, luggage, furniture on shopping websites. Especially, in the current special period of COVID-19, it is not convenient for people to go out frequently to buy food and clothing. Online shopping has become an indispensable part of many people's lives. The customers only need to click shopping websites (e.g., Amazon [2], Costco [5], ASOS [3]) on the computer or smartphone, and then customers can choose the products (e.g., clothes) they need under the specified catalog. In addition, the websites usually recommend several similar products based on customers' browsing products so that they can easily compare different products to find the most suitable one.

Specifically, imagining if we are shopping on a clothing website without any classification, which means the products on the websites are mixed together. It will take a very long time for a lady who wants a dress to find what she wants. Even worse, if the website always recommends pairs of men's pants to a lady who wants to buy a dress, it undoubtedly will be a very unpleasant shopping experience for that lady, therefore the website will probably lose their customers. Definitely, the online clothing shopping websites can hire some employees to manually classify product images into different categories (e.g., T-shirt, dress, hoodie). The e-commerce companies can also hire employees to serve online customers who are browsing the website and recommend similar products to them. However, this will be a big investment for time and money costs especially while they need to maintain 24-hour manual service.

In this project, we focus on clothing image analysis and classification, which has attracted lots of interests in computer vision in recent years. Benefited from the availability of high-performance hardware support and large-scale fashion datasets, deep learning based models for clothing item retrieval and fashion image classification gained astonishing success in this area [16]. We studied several prior arts and learnt that Convolutional neural networks (CNNs) is popular for image classification with high accuracies. However,

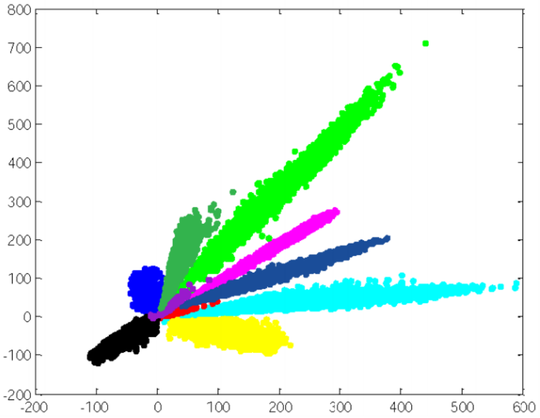

Figure 1. The inter-class variation is even smaller than the intra-class variation by traditional CNN model. (Figure credited by Yang *et al.* 2018

CNN has a limitation that when the small noises are added to the initial samples, CNN will give different predictions for these samples with high confidence. That means it can be easily fooled by some adversarial examples and it is not robust enough in real and complicated environments. As we can see in Figure 1, Inspired by prior arts, we implement a convolutional prototype learning framework combining U-Net with Distance based cross entropy loss (i.e., DCE) function and prototype loss (i.e., PL) function to do the clothing classification. We processed the data, wrote the code of U-Net with the loss function (i.e., DCE and PL) to do the clothing classification. We can see the whole framework can be trained efficiently and effectively, and can achieve 0.72 acc. Based on the result evaluation, we also discussed the advantages and limitations.

2. Related Work

Visual clothing analysis has drawn lots of interests recent years due to its wide spectrum of human-related applications such as clothing classification and clothing recommendation. Earlier models [4, 8] have mostly relied on handcrafted features (e.g., SIFT, HOG) and seek for powerful clothing representations, such as graph models, contextual information, human parts, and semantic masks.

Convolutional neural networks [10] (CNNs) have achieved great success for this area [9, 7, 6] in recent years. Wang *et al.* [16] proposed a Bidirectional Convolutional Recurrent Neural Networks (BCRNNs) with two attention mechanisms (i.e., landmark-aware attention and category-driven attention) that incorporate both powerful learning capabilities of neural networks and high-level semantic relations for fashion landmark detection and clothing category classification. Liu *et al.* [12] introduced a large-scale clothes dataset DeepFashion over 800K images and proposed a deep model FashionNet to learn clothing features by jointly predicting clothing attributes and landmarks which are further employed to gate the learned features. Xiao *et al.* [18] proposed a framework to train CNNs with limited clean labels and some noisy labels. The relationships between images and labels are modeled with a probabilistic graphical model and integrated to an end-to-end deep learning system. Yang *et al.* [19] used CNN as a feature extraction tool to propose a prototype learning framework called Convolutional prototype learning (CPL) and design several modes of loss function for it to improve the rubust.

Some prior arts concentrate on improving existing loss function or proposing new loss functions to learn discriminative and robust representations. Sun *et al.* [15] combined the cross entropy loss and previous contrastive loss to train the CNN. The former loss function can increase the inter-personal variations; the later loss function can reduce the intra-personal variations, and both of them to guide the CNN to learn representations together. Liu *et al.* [11] used a generalized large-margin softmax loss to encourage intra-class compactness and inter-class separability between learned representations, which can make the representation more robust. Schroff *et al.* [14] proposed a triplet loss that can learn representations in a compact Euclidean space for CNN. They found it performs well even on several tasks including recognition, verification, and clustering. Wen *et al.* [17] proposed a center loss that work with cross entropy loss function together to train the CNN. For the centers, they used a mini-batch based update method and this method was proved to be useful especially for face recognition and verification.

3. Architecture of the framework

In this project, we implement a convolutional prototype learning framework combining U-Net with Distance based cross entropy loss (i.e., DCE) function and prototype loss (i.e., PL) function as shown in Figure 2.

Prototype learning is a classical and representative method in the area of patter recognition [19]. The earliest prototype learning method is k-nearest-neighbor (KNN), but it has the limitation of requiring the heavy burden of storages space and computation requirement. Different from the traditional CNN that use softmax layer to do the linear classification, here in this project, we multiple prototypes in convolutional feature space for each class. After that, CPL uses prototype matching for further decision making because convolutional prototype learning can do a comparable performance or even better classification accuracy than softmax-based CNN models. Here the prototypes are defined as m_{ij} where i $\in$ 1,2,3, ...C represents the index of the classes and j $\in$ 1,2,3, ...K represents the index of the prototypes in each class. The feature extractor f(x; θ) and the prototypes m_{ij} are both trained from data. The we clas-

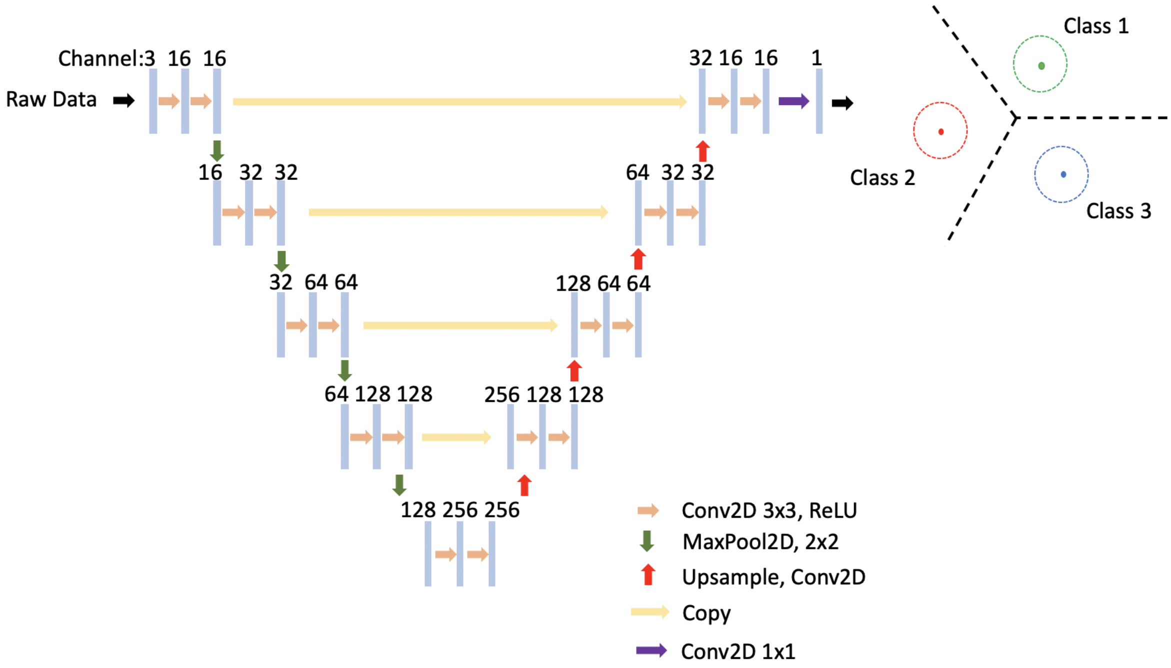

Figure 2. The overview of our network architecture for clothing classification

sify the objects by prototype matching and assign the class of this prototype to the particular object.

U-Net was proposed by Ronneberger *et al.* in 2015 which only requires small number of training examples and outperforms the prior best method (i.e., a sliding-window convolutional network) on the ISBI challenge for segmentation of neuronal structures [13]. Besides, U-Net is fast. It takes less than a second on a recent GPU for the segmentation of a 512x512 image. Hence, to show our respect to this great work, we use this in our model. The network model has two paths, the left side is a contracting path and the right side is an expansive path. This two paths make the artichecture of this network looks like a big U. The contracting path is a typical architecture of a convolutional network which consists of the repeated application of two 3x3 unpadded convolutions. Then for downsampling, it is each followed by a ReLU (i.e., rectified linear unit) and a max pooling operation (2x2) with stride 2. Then it double the number of the feature channels at each down-sampling step. For the expensive path on the right, it consists of an unsampling of the feature map which is followed by a 2x2 upconvolution to cut down the number of feature channels to a half. Others are a concatenation with cropped feature map from the contracting path correspondingly, and two convolutions which is 3x3 and followed by a rectified linear unit. Here, it is necessary to do the cropping because of the loss of those border pixels while executing each convolution. In total, the network we used has 5 convolutional layers.

$$l((x,y);\theta,M) = -log\underline{p(y|x)}.$$

$$p(y|x) = \sum_{j=1}^{K} \underline{p(x \in m_{yj}|x)}.$$

$$p(x \in m_{ij}|x) = \frac{e^{-\gamma d(f(x),m_{ij})}}{\sum_{k=1}^{C}\sum_{l=1}^{K} e^{-\gamma d(f(x),m_{kl})}}$$

Figure 3. The distance based cross entropy loss (DCE) (credited by Yang *et al.* 2018

Many recent methods concentrate on proposing or improving loss functions (e.g., Minimum classification error loss (MCE), Margin based classification loss (MCL)) to let the network learn discriminative and robust representations. In our project, we use Distance based cross entropy loss (DCE) and prototype loss (PL) as a regularization. Let's see the Distance based cross entropy loss (i.e., DCE) first. The probability that whether a sample belongs to the prototype m_{ij} or not can be measured by the distance between them because the distance can be used to measure the similarity between the samples and the prototypes. Besides, the probability should satisfy the non-negative and sum-to-one properties. So the Distance based cross entropy loss is shown in Figure 3.

However, we further find although Distance based cross

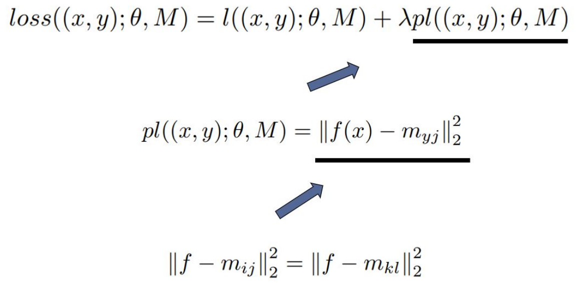
$loss((x,y);\theta,M) = l((x,y);\theta,M) + \lambda \underline{pl((x,y);\theta,M)}$
$pl((x,y);\theta,M) = \underline{\|f(x) - m_{yj}\|_2^2}$
$\|f - m_{ij}\|_2^2 = \|f - m_{kl}\|_2^2$

Figure 4. The prototype loss (PL) function (credited by Yang *et al.* 2018

entropy loss (DCE) measures the classification accuracy, We can train the model to classify the data correctly by minimizing these losses but it also leads to heavily over-fitting. Hence, we add a prototype loss (PL) as a regulation, which is used to overcome the problem that the network separates the whole feature space and the learned representation is still linearly separable so that it is not robust enough to reject the samples from unseen classes. The prototype loss (PL) function in Figure 4, which use a hyper-parameter to control the weigh of prototype loss.

4. Training

4.1. Data Source

The classification model requires a pretty labeled image dataset with both training and validation set. And The *Clothing Models* dataset [1] from kaggle.com is suitable for our experiment. The compressed dataset contains about 3000 images which are separated into 2400 training images and 600 test images.

The input images are 4 types of fashion clothing including: hoddies, long sleeves, shirt and sweatshirt and all images are compressed to a $3*128*128$ RGB tensors in order to feed the model. To minimize the data overhead, we use batch size 50 for all experiments. Also, to simplify the SGD loss function in our model, we use 0.9 momentum and $1e{-}4$ weight decay for all the experiments.

All the training process are done on the *dev-amd20-v100* nodes on HPCC (High Performance Computing Cluster) MSU with interactive batch jobs. The hyper-parameters we used is shown in Table 1.

Table 1. Hyper-parameters in training process

Learning rate	0.0005
Batch size	32
Regularization coefficient	0.5
Epoch	50
Momentum	0.99
Weight decay	0.0001

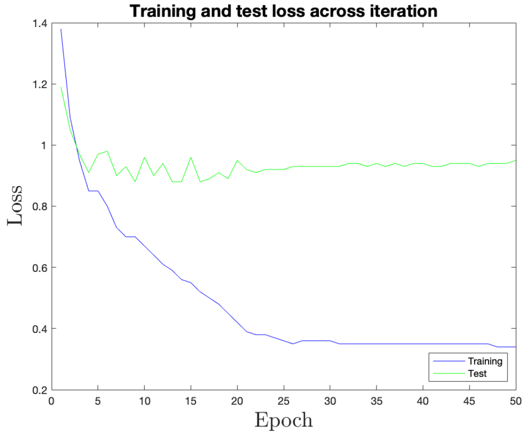

Figure 5. The training and test loss over iterations

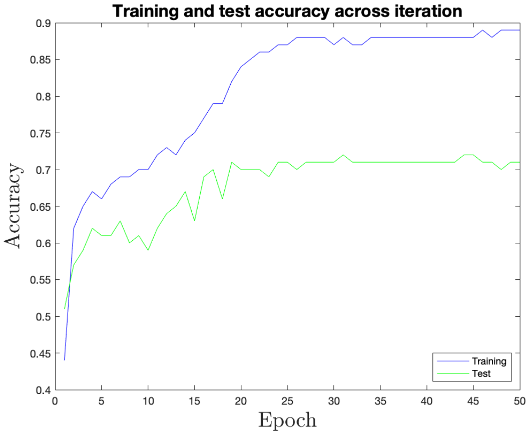

Figure 6. The training and test accuracy over iterations

5. Experiments and results

Our results are shown in Figure 5 to Figure 7. Figure 5 shows that the training and test loss decrease with the number of epoch in the training process. Figure 6 shows that the training and test accuracy increase with the number of epoch and finally we can achieve 0.72 accuracy on test dataset. As shown in Figure 7, we can successfully identify class 0, 2 and 3. However, class 1 overlaps with class 0, which we will discuss later.

6. Discussion

We analysed why the ACC can not be better.One main reason behind this is that this Zalando Clothing dataset has categories for male and female (e.g., hoodie and hoodie-female) and they are very similar as shown in Figure 8 and Figure 9. Even we human are hard to distinguish them for different categories. We will continue improving this as our future work.

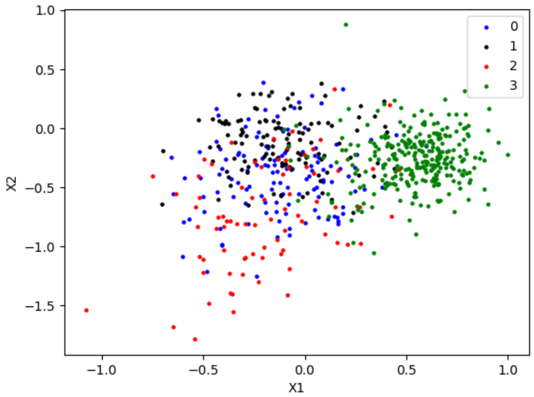

Figure 7. The learned representations on our dataset. Different colors represent different classes.

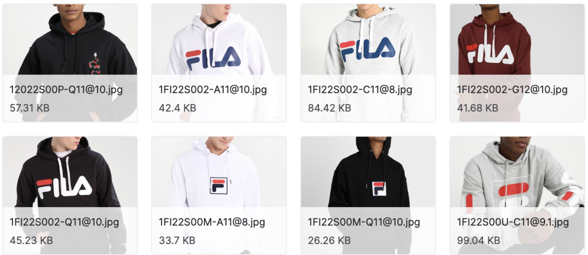

Figure 8. Images from Category: Hoodies from Zalando Clothing dataset.

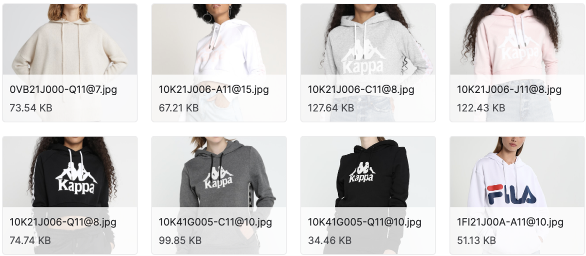

Figure 9. Images from Category: Hoodies-female from Zalando Clothing dataset.

7. Conclusion

In this project, we focus on visual clothing analysis. We learnt that Convolutional neural networks (CNNs) have been widely used for image classification with high accuracies but it exsits the limitation that can be easily fooled by some adversarial examples and not robust enough for pattern classification. To improve the robust, we implemented a convolutional prototype learning framework using U-Net with Distance based cross entropy loss (i.e., DCE) function and prototype loss (i.e., PL) function. Different from the softmax-based models, convolutional prototype learning can directly learn multiple prototypes for each class and then use prototype matching for decision making which will lead to a higher classification accuracy than softmax-based CNN models. Two loss functions work together to improve the robust of the model. Based on the result using fashion datasets, it can achieve 0.72 acc. Through this assignment, during the procedures of processing the data, reading the papers, and writing the code of U-Net with the loss function (i.e., DCE and PL), we have a deeper understanding of the network structure and the designs of loss functions.

8. Acknowledgments

Thank Dr. Xiaoming Liu for his advice! Thank TA for his help! Thank CSE803 classmates for their support during our presentations!

References

[1] Clothing models: A collection of clothing pieces, scraped from zalando.com. `https://www.kaggle.com/dqmonn/zalando-store-crawl`, 2019.

[2] Amazon. Amazon.com: Online shopping. `https://www.amazon.com/`, 2020.

[3] ASOS. Asos: Online shopping for the latest clothes fashion. `https://www.asos.com/us/`, 2020.

[4] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *European conference on computer vision*, pages 609–623. Springer, 2012.

[5] Costco. Costco wholesale. `https://www.costco.com/`, 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[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*, pages 770–778, 2016.

[8] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *European conference on computer vision*, pages 472–488. Springer, 2014.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[10] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[11] W. Liu, Y. Wen, Z. Yu, and M. Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016.

[12] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 1096–1104, 2016.

[13] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[14] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[15] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. *Advances in neural information processing systems*, 27:1988–1996, 2014.

[16] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4271–4280, 2018.

[17] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[18] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.

[19] H.-M. Yang, X.-Y. Zhang, F. Yin, and C.-L. Liu. Robust classification with convolutional prototype learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3474–3482, 2018.