Wide, Deep, and Recurrent: An End-to-End Neural Architecture for Automatic Fashion Product Labeling

Binbin Xiong, Minfa Wang
Stanford University

Abstract

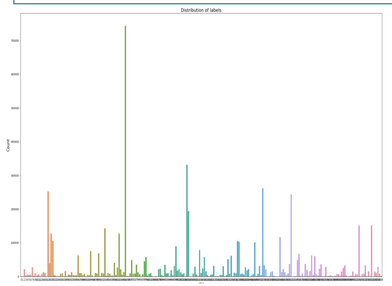
We propose a novel neural network architecture for the iMaterialist Challenge (Fashion) at FGVC5[1]. We leverage transfer-learned Xception[cite] as our base image feature extractor which we call it deep. On top of it, we propose two novel module components. One is to use recurrent network with pre-trained label embeddings to explicitly measure label dependency, the other is to combine our generalized deep model with shallow yet wide features such as color histogram and HOG as a way to enhance memorization. To deal with class imbalance, we also leveraged label sampling and per-class threshold selection techniques

Introduction

Automatic labeling and classification for product photos has been a constant effort for online merchandise companies and could potentially improve the shopping experience for shoppers drastically. However, it is not an easy task to tackle because of the complex settings: different lighting, angles, backgrounds, and levels of occlusion, etc.

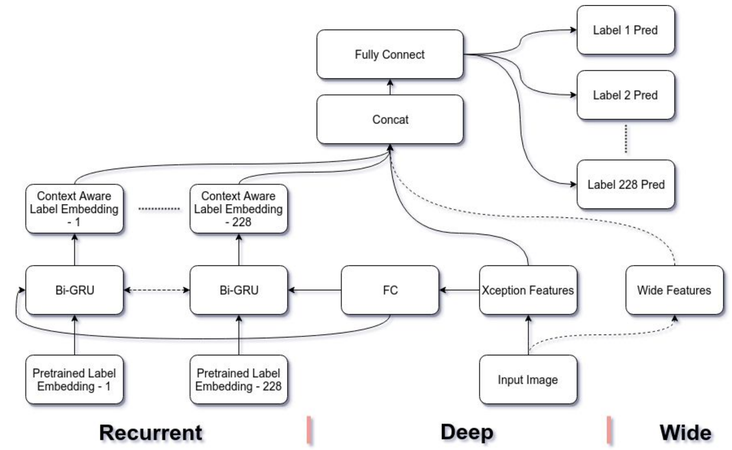
Problem Statement

The problem that we would like to solve is given a product image, output labels that the image belongs to. As required by the competition holder, we use F1 metric (micro-average) for our model evaluation.
We use dataset provided by FGVC5, which contains 1M training images, 40K test images and 10K validation images. There are a total 288 classes.
Figure below in the left shows label count distribution on training set, the figures on the right are from training set.



Model	Val F1	Test F1
Deep (fixed all)	0.511	0.512
Deep(fixed half)	0.520	0.516
Deep(tune)	0.610	0.609
Deep(tune)+Wide+Recurrent	0.620	0.616
Deep(tune)+Wide+Recurrent*	0.648	0.643
Ensembled*	0.661	0.656
Test @1st	-	0.725

Method



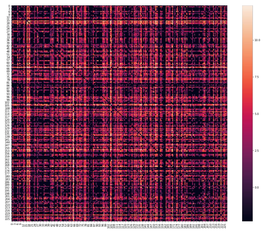
Our proposed architecture could be found in the figure above, where we combine three sub modules, Deep, Wide and Recurrent. For each of the sub module:

- **Deep**: This part is for image understanding. We use pretrained Xception[2] as the image feature extractor. We fine-tune all layers with weights from ImageNet.
- **Wide**: This part is for memorization. According to our observation, some labels are just colors. Inspired by Wide-And-Deep[3], we add shallow yet wide HOG features and Color Histogram features to enhance the memorization of our model.
- **Recurrent**: This part is for label dependency modeling. Labels are not IID, thus we would like to explicitly model the dependencies between labels. One technique that has been proven effective in modeling dependency is RNN. We use bi-directional GRU[4] as our RNN unit. We use matrix factorization on the label co-occurrence matrix to pre-train the label embeddings.

- Other techniques:
  - Weighted sampling:
    - Labels are very skewed in our case. We thus leverage[5] to adjust weights for each sample.

$$\begin{cases} (w_{l_1} \cdot f_{l_1})/(w_{l_2} \cdot f_{l_2}) = (f_{l_1}/f_{l_2})^{\alpha} \quad \forall l_1, l_2 & (1) \\ \sum_l (w_l \cdot f_l) = \sum_l f_l & (2) \end{cases}$$

  - Per class thresholding:
    - We use a simple "cyclic optimization procedure"[6] on the validation set to choose thresholds. This in general gives 2-3% F1 gain.



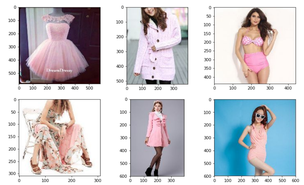


Figure above shows the label co-occurrence matrix. Figure below shows images from label 131, which is very likely to be Pink Color.

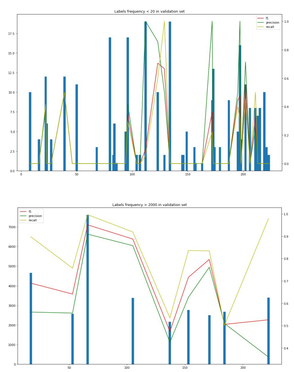
Experimental Results & Analysis

The model performance differs with label frequency. In the validation set, we observed that out of the 228 unique labels, a number of them were not predicted to be positive once.
In addition, the two bar-line plots below illustrated that:

- the model tends to be more aggressive on recalls for very frequent labels (count > 2000)
- the model tends to be more conservative and less robust on very rare labels (count < 20)

The saliency map illustrates what the model emphasises when making predictions.
Some highlights:

- the model could distinguish background noise from main figure very well.
- the model focuses on different regions of the body when making different predictions.



Conclusions and Future Works

We proposed and experimented a novel Wide, Deep and Recurrent architecture for automatic fashion product labeling on the FGVC5 fashion dataset with over one million images. Deep for input image understanding, Wide for image label memorization, and Recurrent for label dependency understanding. Our best single model achieves 0.648 F1 score on validation set which **ranks 6th** on the public leaderboard. (Note that most of the work was done after the competition deadline, thus on the public board our team don't appear on 6th.)

In the future, we would like to explore further with:

- more model combination: incorporating conditional random fields or mask-RNN.
- distributed training across different machines/GPUs.
- different ensemble techniques and hyper parameter tuning
- better model understanding: visualization of what the model learns for different labels.
- sophisticated data augmentation using generative models

Contact

Binbin Xiong
bxiong@stanford.edu

MInfa Wang
minfa@stanford.edu

References

1. Kaggle. imaterialist challenge (fashion) at fgvc5 dataset, 2018. https://www.kaggle.com/c/imaterialist-challenge-fashion-2018.
2. F. Chollet. Xception: Deep learning with depthwise separa-ble convolutions.arXiv preprint, 2016.
3. J. H. T. S. T. C. H. A. G. A. G. C. W. C. M. I. R. A. Z. H.L. H. V. J. X. L. H. S. Heng-Tze Cheng, Levent Koc. Widedeep learning for recommender systems.arXiv:1606.07792,2016.
4. K. Cho, B. Van Merri ̈enboer, C. Gulcehre, D. Bahdanau,F. Bougares, H. Schwenk, and Y. Bengio. Learning phraserepresentations using rnn encoder-decoder for statistical ma-chine translation.arXiv preprint arXiv:1406.1078, 2014.
5. R. S. Jeffrey Pennington and C. D. Manning. Glove: Globalvectors for word representation. InEmpirical Methods inNatural Language Processing (EMNLP), 2014.
6. R.-E. Fan and C.-J. Lin. A study on threshold selection formulti-label classification.Department of Computer Science,National Taiwan University, pages 1–23, 2007.