Large-Scale Wi-Fi Hotspot Classification via Deep Learning

Chang Xu, Kuiyu Chang, Khee-Chin Chua, Meishan Hu, Zhenxiang Gao
LinkSure Inc.
{xuchang,kuiyu.chang,chuakheechin,humeishan,gaozhenxiang}@wifi.com

ABSTRACT

We describe the problem of classifying hundreds of millions of Wi-Fi hotspots using only connection and user count characteristics. We use a combination of deep learning and frequency analysis. Specifically, Convolution Neural Networks (CNN) capture the spatio-temporal relationship between adjacent connection/user counts across a $24hour \times 7day$ matrix, while FFT (Fast Fourier Transforms) extract user and connection frequencies. Our production system has been deployed to classify 239 million hotspots in 12 hours on a SPARK 2.0 cluster, achieving close to 80% F1-score for binary classification.

Keywords

Wi-Fi hotspot classification; deep learning; large-scale production

1. INTRODUCTION

WiFiMasterKey[1] (henceforth abbreviated as WMK) is the only non-BAT (Baidu, Alibaba, Tencent) affiliated App (mobile application) among the top 10 downloaded Chinese Apps in 2016. WMK enables users to freely connect to over 400m password protected WiFi Hotspots (HS) worldwide. HS passwords are contributed by WMK users and partners. As of June 2016, there are 520m monthly active users of WMK.

Since each HS is user-contributed, only its SSID (Service Set Identifier) and hardware MAC (Media Access Control) address are known (other than the password). If a HS can be identified by category, be it private/residential, office or retail, etc., targeted ads can be pushed to the WMK user. For example, if a user is connected to a restaurant HS on a Friday evening, promotions for nearby cinemas or cafes could be sent to the user. Likewise, private/residential and corporate HS can also be filtered to protect the privacy of individuals and corporations, respectively.

[1]`https://play.google.com/store/apps/details?id=com.halo.wifikey.wifilocating&hl=en`

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054239

The challenge of classifying individual HS lies in the dearth of HS features. The SSID (WiFi name) can sometimes reveal a HS's category, e.g., JacksRestaurant. However, only 8.9% of the hotspots in our data have SSID that matches our SSID dictionary. In this work, we seek a more general solution and propose an approach to extract HS features using time-domain (CNN Features) and frequency domain (FFT Features) analysis of the HS connection patterns. The two sets of features and their combinations are then fed into a CNN and Tree classifier.

Problem Definition: We aim to classify hotspots into 11 base classes: 1 Dining (restaurants), 2 Entertainment (ktv, amusement), 3 Living (convenience stores, barbers, etc.), 4 Hotel, 5 Residential, 6 Office, 7 Shopping, 8 Health (clinics, gyms, hospitals), 9 Tourist (parks, attractions), 10 Education (schools and colleges), 11 Transportation (railway stations and airports). To a user, a HS can be associated to a specific location that he/she may find interesting, and thus can serve as a point of interest (POI) to him/her. In this regard, our work complements existing studies on POI mining that focus on POI characterization and recommendation [1].

2. FEATURES AND MODELS

WMK logs the time-stamp of each user connection to a HS, which ranges from one to several thousands per day.

CNN Features - Weekly Connected Hours: Two 7×24 matrices were created for each HS by averaging the historical connection and user counts over 4 weeks. Figure 1 shows the normalized average connection count matrices for the 11 classes of the o2o133k dataset. We can see that the user connection distributions differ from class to class across the dimensions of both hours and days.

FFT Features - Connection Frequency: Assuming that the aggregated user and connection periodicities vary with different classes of hotspots, we compute the FFT or power spectrum analysis [2] of each HS's aggregated connection and user count time series, yielding the connection and user-count spectra.

For example, a restaurant HS can have many irregular users (the customers) connecting to it occasionally while a residential HS can have few regular users (the HS owners) connecting to it almost everyday. Power spectrum analysis allows us to capture such subtlety in the frequency domain.

Learning Models: We evaluated three learning approaches: (1) **CNN**: a 10-layer CNN trained on the CNN features ($2 \times 24 \times 7$ tensor). The architecture is inspired by [3], as shown in Figure 2. The layers are tuned to fit our task.

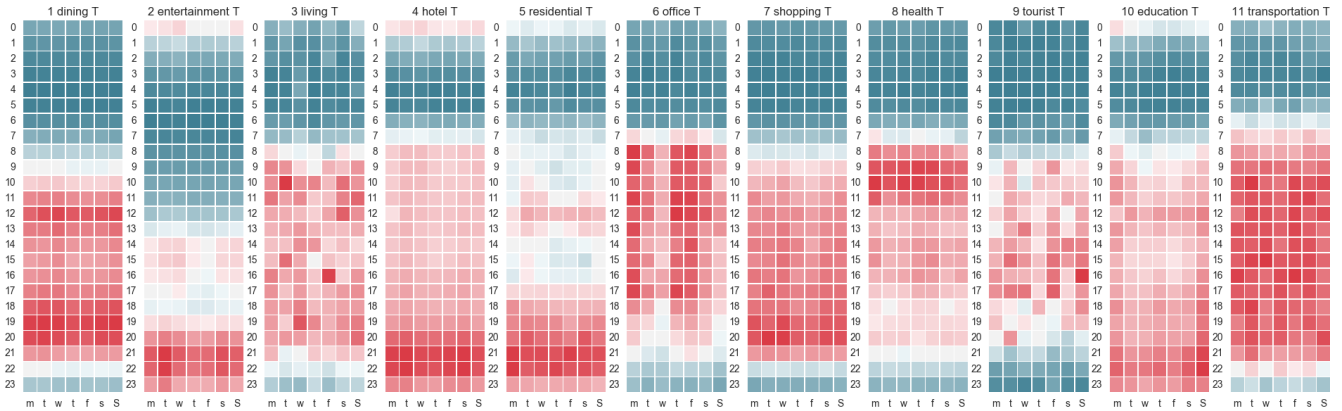

Figure 1: Average connection counts for 11 types of hotspots. Counts normalized from 0 (Dark Blue) to 1 (Dark Red). Vertical labels 0-23 denote 24 hour blocks, while horizontal labels "mtwtfsS" denote day-of-week.

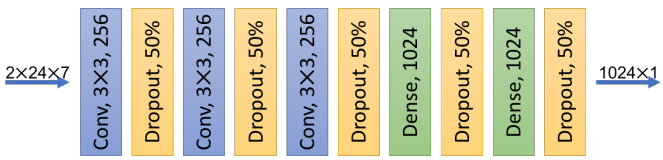

Figure 2: The architecture of our CNN model.

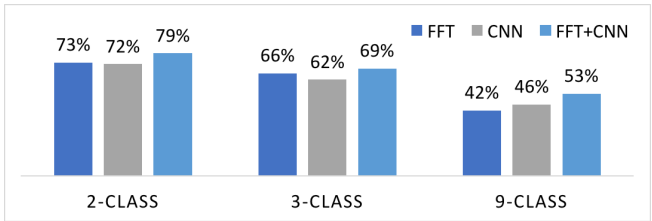

Figure 3: 5-Fold CV F1 scores comparing CNN, FFT, CNN+FFT on o2o133k dataset.

(2) **FFT**: a Gradient Boost Random Forest classifier (GBRF) [4] trained on the FFT features (142-D feature vector for the user and connection spectra).
(3) **FFT+CNN**: a GBRF trained on the extended vector comprising the 142-D FFT feature vector and the 336-D ($2 \times 24 \times 7$) flatten CNN tensor.

3. EXPERIMENT

Datasets: The following datasets were used for evaluation[2],
(1) **o2o133k** contains 133,508 HS, extracted via SSID keyword matching from a list of 19m user submitted HS data. This class-imbalanced dataset is used to tune and validate our learning models; class 3 (Living) and 9 (Tourist) are approximately 50 times smaller than the largest class (4 Hotel). Standard counter measures against imbalanced classes, e.g., over/under sampling, were employed during training.
(2) **production55k** contains 55,000 HS, extracted from the full list of 239m active HS using the same SSID keyword matching as o2o133k. This balanced dataset is used to train the finalized model for the production classifier. This dataset was not used for parameter tuning as it has not undergone human evaluation and validation.

The SSID keyword matching approach yields high-precision results, but covers less than 10% of the entire production data. As such, it is best used to extract high-quality labelled data for our learning models.

Various class aggregations were done to serve business needs: 2-class (commercial and non-commercial), 3-class (retail, private, corporate), and 9-class (exclude class 3 and 9).
Results and Deployment: Figure 3 shows the 5-fold cross-validated F1 scores of the 3 models on the o2o133k dataset for various class aggregations. We see that CNN performed better than FFT for the more complex 9-class problem, while FFT outperformed CNN for 2 and 3 class formulations. All in all, a simple combination of FFT and CNN features fed into a GBRF classifier outperformed classifiers using CNN and FFT alone, achieving close to 80% F1 score for binary classification. We trained the o2o133k tuned 10-layer CNN on the 11-class production55k data to obtain a production quality model[3]. This trained model is then deployed on a 4TB SPARK 2.0 cluster to classify over 239m HS, which took 12 hours. The class-distribution of 239m classified hotspots is 82.07% for Retail, 2.76% for Corporate, and 15.17% for Private.

4. CONCLUSIONS

We proposed a method to classify hotspots using only hotspot connection and user count information. By utilizing both time and frequency domain features, we achieved close to 80% F1 for two classes, and 53% F1 for 9 classes. We plan to incorporate additional features like the geo-information of POI to further improve classification accuracy.

5. REFERENCES

[1] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *SIGIR*. ACM, 2011.
[2] Qi He, Kuiyu Chang, and Ee-Peng Lim. Analyzing feature trajectories for event detection. In *SIGIR*. ACM, 2007.
[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
[4] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[2]The use of data in this work strictly adheres to WMK's User Terms & Privacy Policy.

[3]The FFT+CNN distributed production model is currently under development