Yelp Restaurant Photo Classification

Capstone Project

CSC 591 - Algorithms for Data Guided Business Intelligence

Team Members:
Omkar Acharya (oachary)
Amit Watve (awatve)
Akshay Arlikatti (aarlika)

Introduction

- In an age of food selfies and photo-centric social storytelling, it may be no surprise to hear that Yelp's users upload an enormous amount of photos every day alongside their written reviews.
- Currently, restaurant labels (like classy ambience, outdoor seating, etc.), are manually selected by Yelp users when they submit a review.
- Selecting the labels is optional, leaving some restaurants un- or only partially-categorized.
- We aim to solve this problem by providing automatic classification of restaurants using Yelp’s photo dataset.

Dataset

- At Yelp, there are lots of photos and lots of users uploading photos.
- Training Set: 234,842 photos of 2,000 restaurants
- Testing Set: 1,190,225 photos of 10,000 restaurants

Problem Statement

- Given photos that belong to a restaurant, train a model that can predict different restaurant attributes.
- The 9 different attributes in this problem are:
  - 0: good_for_lunch
  - 1: good_for_dinner
  - 2: takes_reservations
  - 3: outdoor_seating
  - 4: restaurant_is_expensive
  - 5: has_alcohol
  - 6: has_table_service
  - 7: ambience_is_classy
  - 8: good_for_kids
- These labels are annotated by the Yelp community.

Business Value

- These photos provide rich local business information across categories.
- Rather than rely on a user who may not manually fill in all the labels at review time, the goal is to utilize user uploaded photos to automatically label restaurants with descriptive attributes for the benefit and convenience of not only users but also restaurants.
- Capturing this information can be used for recommendation systems, restaurant search filtering, ratings, etc.

Literature Survey

- Wei et al. [2] proposes a model to predict multiple labels for images using a CNN pretrained on a large-scale single-label dataset such as ImageNet, aggregating predictions using max pooling.
- Razavian et al. [3] suggests the use of transfer learning on features extracted from deep learning with convolutional nets to use as image representations.
- Kraus et al. [4] experimented with utilizing global pooling layers with fully connected layers in a convolutional neural network to learn relationships between classes using an adaptive Noisy-AND pooling function.

Our Approach

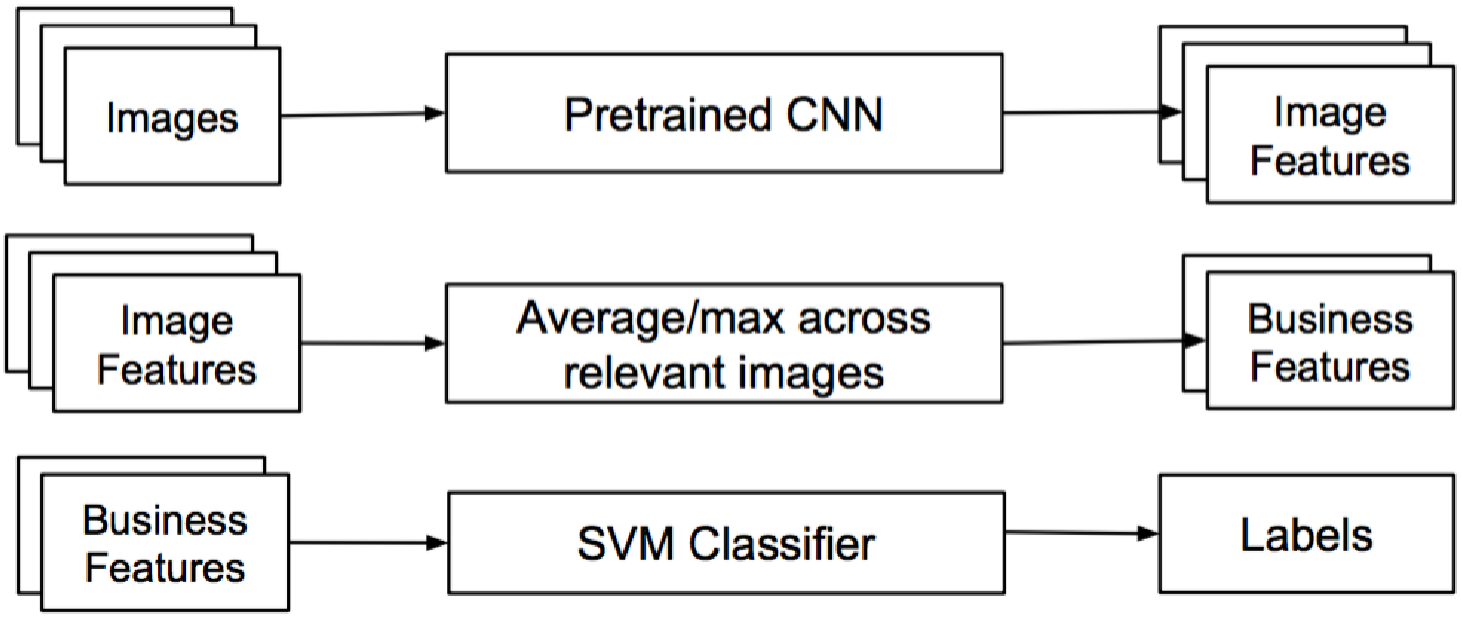

Fig. Three-stage transfer learning pipeline

Image Feature Extraction using CNN

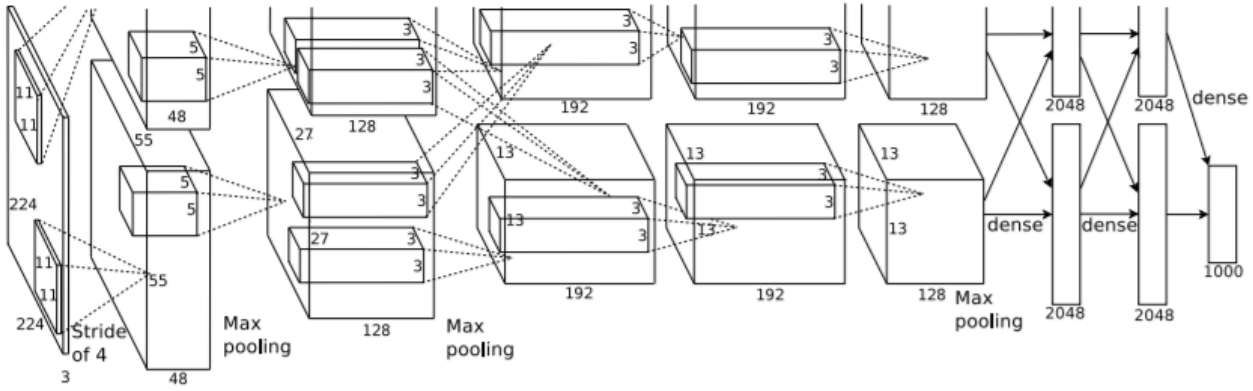

Fig. BAIR Reference CaffeNet

- Input Size: 224×224×3
- Convolutional Layer 1:
  - Number of Kernels: 96
  - Size: 11×11×3
  - Stride: 4 pixels
- Convolutional Layer 2:
  - Number of Kernels: 256
  - Size: 5 × 5 × 48
- Convolutional Layer 3:
  - Number of Kernels: 384
  - Size: 3 × 3 × 256
- Convolutional Layer 4:
  - Number of Kernels: 384
  - Size: 3 × 3 × 192
- Convolutional Layer 5:
  - Number of Kernels: 256
  - Size: 3 × 3 × 192
- Fully Connected Layer 6:
  - Output: 4096×1
- Fully Connected Layer 7:
  - Output: 4096×1

Restaurant Feature Extraction

- Having obtained these features for each image from CNN, relate the images to their associated network by merging the relevant image features together.
- Each restaurant is represented by its “average” image features.
- e.g. For a restaurant ‘r’, there are 50 images. ‘r’ can be represented by one 4096-vector which is the average of 50 feature vectors of size 4096.
- Having converted our image features into business features, it is now time to train an SVM classifier.

Support Vector Machine Classifier

- Train the SVM classifier to predict 9-dimensional vectors for each restaurant, where each of our nine labels receives a 1 or 0 score relating respectively to the presence or absence of a label.
- This is a case of “Multilabel Learning”, in comparison to multi-class learning where we predict one class out of multiple class options.
- To do this, we use a one-vs.-all multi-label strategy of training one classifier per label.
- For each label, SVM finds the separating hyperplane by using

$$f(x_i, W, b) = Wx_i + b$$

where, x_i is a given restaurant feature vector, W is a matrix of weights, and b is a bias vector.

Evaluation

- We used Mean F1-Score as the evaluation metric for our multi label classification

$$MFS = \frac{1}{N}\sum_{i=1}^{N} f_\beta(C_i)$$

where, N is row's size of train set, and $f_\beta(C_i)$ is the score of a single class label. In general, $f_\beta = 1$.

- **Validation results:**
  - Mean F1 Score: 0.797
  - Individual F1 Scores:

Class	0	1	2	3	4	5	6	7	8
F1 score	0.609	0.791	0.824	0.639	0.767	0.845	0.924	0.753	0.845

- **Test results:**
  - Mean F1 Score: 0.764

References

1. Huang, Jade. "Multiple Instance Multi-Label Learning for Yelp Restaurant Photo Classification."
2. Z. Zhou, M. Zhang, S. Huang, and Y. Li. MIML: A framework for learning with ambiguous objects. CoRR, abs/0808.3231, 2008.
3. A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. CoRR, abs/1403.6382, 2014.
4. O. Z. Kraus, L. J. Ba, and B. J. Frey. Classifying and segmenting microscopy images using convolutional multiple instance learning. CoRR, abs/1511.05286, 2015.