Practical 3: Semi-Supervised Learning for Urban Sound Classification

Abdul Saleh, Dean Hathout, Brendan O'Leary
{abdelrhman_saleh, dhathout}@college.harvard.edu, boleary@g.harvard.edu
abdulsaleh, dhathout, boleary134

April 13, 2019

1 Technical Approach

Urban sound classification is the task of classifying audio sequences of environmental sounds into classes. Urban sound classification is employed in self-driving cars and numerous real world applications making it an important problem in machine learning. However, as is the case with many machine learning models, supervised models for classifying sounds are limited by the scarcity of labeled data.

In this practical, we explore two methods for learning from **both labeled and unlabeled data** to mitigate the issue of data scarcity. We hypothesize that leveraging both labeled and unlabeled data will help us find better classification boundaries leading to improved generalization. We also motivate our approaches from both *Bayesian* and *frequentist* perspectives. *Hypothesis*

Multiple approaches for learning from labeled and unlabeled data (or semi-supervised learning) have been proposed. We list below the two approaches we explore:

1. We apply k-means clustering on all the data and use the cluster assignments as input features to the supervised learning models similar to the analysis in Peikari et al. (2018). We refer to this as the **clustering approach**.
2. Following the analysis in Zhang and Schuller (2012), we first train a supervised model on the training data. We then infer the labels for the unlabeled data and use predictions made with $> 80\%$ confidence as true labels. We then retrain our supervised model on the expanded data set. We refer to this as the **label propagation approach**.

Note that for the clustering approach, the supervised models are still only trained on the originally labeled data points. However, we expect that the cluster assignments (derived from both labeled and unlabeled data points) will incorporate extra knowledge about the data distribution that is not obvious from considering the labeled data separately. This makes sense from a Bayesian perspective as the inherent structure of the features is assumed to generate the labels.

For the label propagation approach, we take a more frequentist perspective as we directly model $P(\text{label} \mid \text{data})$ and propagate the predictions learned by the supervised model to the unlabeled

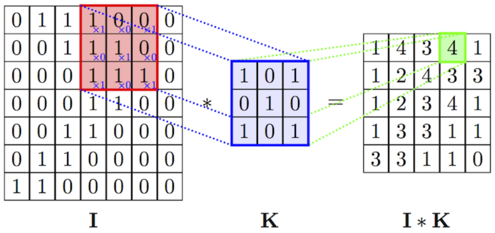

Figure 1: A $(3, 3)$ filter with stride 1 applied to $(32, 32, 3)$ input.

data. This is justified under a frequentist perspective as we are only concerned with learning the distribution of the label given the data rather than modelling the distribution of the data independently.

Preprocessing

The raw training data contains 6,374 audio clips represented as sampled amplitudes, only 2,307 of which are labeled. The labeled training samples come from ten classes of sounds: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music. We follow the preprocessing in Shu et al. (2018) and generate mel-scaled spectrograms from the amplitudes, which depict the sound frequencies of the signal over time. Since the sound waves are likely to repeat in a similar manner over time, we split each sample into multiple 50% overlapping segmentation windows 320ms in length. This increases the data sample size by 25 times. Our feature set includes both the mel-spectrogram frequency features as well as multiple width delta frequencies.

Our main model for this practical is the convolutional neural network (CNN) proposed in Shu et al. (2018). The network is composed of 8 layers defined as follows:

Layer	Output Dim.	Description
*	$(64, 14, 2)$	log-scaled mel-spectogram
1	$(64, 14, 32)$	$(2, 2, 2)$ convolution, 32 filters, ReLU
2	$(63, 13, 32)$	$(2, 2)$ convolution, 32 filters, ReLU
3	$(31, 6, 32)$	$(2, 2)$ max pool, 15% dropout
4	$(31, 6, 64)$	$(2, 2)$ convolution, 64 filters, ReLU
5	$(30, 5, 64)$	$(2, 2)$ convolution, 64 filters, ReLU
6	$(15, 2, 64)$	$(2, 2)$ max pool, 20% dropout
7	$(1, 256)$	Fully-connected, ReLU, 50% dropout
8	$(1, 10)$	Fully-connected, Softmax

Table 1: Table outlining our model architecture.

Figure 1 shows an example of a convolutional layer. In our network, the values of the filter (colored blue) are the learned weights. Our model has 500k trainable parameters. We minimize the cross-entropy loss defined by:

$$\mathcal{L}(\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = -\sum_{i} \sum_{j} y_{ij} \log p(\hat{y}_{ij} \,|\, \mathbf{x}_n) \tag{1}$$

When applying the unsupervised clustering approach, we append the cluster assignments as one-hot

Model	Labeled	Clustering	Propagation
LR	41.0	40.4	44.2
RF	45.2	44.2	54.3
CNN	56.0	**57.7**	56.6

Table 2: Test set accuracy of all trained models. Best result in bold. Largest improvement from semi-supervision underlined.

neurons at layer 7.

Implementation

We implement our CNN using `Keras`[1] and use Adam optimizer (Kingma and Ba, 2014) for training with the default learning rate of 0.001. We use `sklearn`[2] for all other machine learning models. And we use `LibROSA`[3] for preprocessing the data.

2 Results

In addition to our CNN, we experiment with logistic regression (LR) and a random forest classifier (RF) for reference. The public test set results of our models are summarized in Table 2.

The CNN outperformed all other models across both the supervised and the semi-supervised experiments. The highest accuracy was achieved by training the CNN with semi-supervision using the pre-calculated cluster assignments as additional input features.

The results also showed that the effect of semi-supervision was inconsistent across different models. In some cases, leveraging unlabeled data resulted in significant performance improvements. For example, the RF classifier saw a 10% increase in accuracy through label propagation. However, in other cases the change in accuracy was negligible. Label propagation improved the CNN accuracy by a modest 0.6% while clustering semi-supervision hurt the performance of LR by 0.6%.

Overfitting

We also note that the training accuracy is much higher than test accuracy in our experiments. We get a training accuracy of $> 90\%$ and $> 80\%$ for the CNN and RF, respectively. We hypothesize this large gap between training and test performance is caused by the difference in class distribution between the test set and the training set. So here we have an issue of an unrepresentative training sample rather than an overly complex model.

3 Discussion

In this practical, we compared two semi-supervised learning methods for urban sound classification. Our results support our hypothesis and suggest that learning from both labeled and unlabeled data can be preferable to learning from labeled data alone. We found that our semi-supervised approaches sometimes produced significant gains of up to 10% in absolute accuracy when training

[1] https://keras.io/ [2] https://scikit-learn.org/ [3] https://librosa.github.io/

ID	Class	Train Freq. (%)	Train & Test Freq. (%)
0	Air Conditioner	19.33	13.61
1	Car Horn	4.07	2.77
2	Children Playing	19.12	13.23
3	Dog Bark	15.47	9.22
4	Drilling	0.87	10.98
5	Engine Idling	19.25	13.12
6	Gunshot	0.52	0.22
7	Jackhammer	0.87	10.96
8	Siren	0.87	12.24
9	Street Music	19.64	13.65

Table 3: Class frequencies in % for labeled training data and entire dataset (train and test).

an RF classifier with label propagation. However, these gains were not consistent across all models and incorporating unlabeled data often made negligible difference.

In retrospect, it is not surprising that semi-supervision resulted in limited improvements for most models. For the clustering case, using cluster assignments is not helpful if the labeled data is enough to learn generalizable decision boundaries. Using cluster assignments could also be unhelpful or problematic because of the difficulties clustering high dimensional audio data to begin with. Similarly, label propagation could be ineffective because the labels assigned to the unlabeled data come from predictions of a model that was only trained on the labeled examples. So label propagation could be interpreted as a data augmentation approach which supplements the labeled examples rather than mining for new patterns.

The influence of semi-supervision supervision could have been diminished by the fact that our preprocessing pipeline inflates the training data by factor of 25 increasing the size of the training set. This can be viewed as a data augmentation strategy similar to label propagation. However, if the training set size had remained small and uninflated, the effects of semi-supervision might have been more pronounced.

As mentioned above, we notice a significant disparity between training and test set accuracy in our models. We do not believe this is due to overfitting, as we scored our models on a separate validation split taken from the labeled training data, and achieved similar accuracy to the training set. Rather, we believe that this disparity is the result of significantly different class distributions in the training and test data. We calculate the distribution of sound classes in the training set, and compare to the distribution in the entire dataset (train & test) in Table 3.

References

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mohammad Peikari, Sherine Salama, Sharon Nofech-Mozes, and Anne L Martel. 2018. A cluster-then-label semi-supervised learning approach for pathology image classification. *Scientific reports*, 8(1):7193.

Haiyan Shu, Ying Song, and Huan Zhou. 2018. Time-frequency performance study on urban sound classification with convolutional neural network. In *TENCON 2018-2018 IEEE Region 10 Conference*, pages 1713–1717. IEEE.

Zixing Zhang and Björn Schuller. 2012. Semi-supervised learning helps in sound event classification. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 333–336. IEEE.