ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Training Deep Nets with Imbalanced and Unlabeled Data

Jeff Berry[1,3], *Ian Fasel*[2], *Luciano Fadiga*[3,4], *Diana Archangeli*[1]

[1]Department of Linguistics, University of Arizona, Tucson, AZ, USA
[2]School of Information Science, Technology, and Arts, University of Arizona, Tucson, AZ, USA
[3]Robotics, Brain and Cognitive Sciences, Istituto Italiano di Tecnologia, Genova, Italy
[4]Section of Human Physiology, University of Ferrara, Ferrara, Italy

jjberry@email.arizona.edu, ianfasel@sista.arizona.edu, fdl@unife.it, dba@u.arizona.edu

10.21437/Interspeech.2012-479

Abstract

Training deep belief networks (DBNs) is normally done with large data sets. Our goal is to predict *traces* of the surface of the tongue in ultrasound images of human speech. Hand-tracing is labor-intensive; the dataset is highly imbalanced since many images are extremely similar. We propose a bootstrapping method which handles this imbalance by iteratively selecting a small subset of images to be hand-traced (thereby reducing human labor time), then (re)training the DBN, making use of an entropy-based diversity measure for the initial selection, thereby achieving over a two-fold reduction in human time required for tracing with human-level accuracy.

Index Terms: deep belief networks, ultrasound imaging, tongue imaging, speech processing, bootstrapping, class imbalance problem

1. Introduction

Deep belief networks (DBNs) [1] have proven to be useful for a variety of machine learning tasks. DBNs are often trained on large pre-labeled data sets, such as `MNIST` [2] (50,000 image training set). For new applications, pre-labeled data set may not exist, and labeling new data for training can be prohibitively expensive.

We explore this problem in the context of a novel speech processing task, to extract the contour of the surface of the tongue from ultrasound images. This application of DBNs is unique in that the desired labels are not categorical, but are structured contours. Although no two contours are exactly alike, there is considerable redundancy in the dataset, making the data highly imbalanced. Imbalanced data may be especially problematic for DBNs, as pre-training has been shown to introduce a strong regularization bias to DBNs which remains even if pre-training is followed by supervised fine-tuning with enormous amounts of labeled data [3].

To address these problems, we use a method which decreases both the amount of time required to hand-trace data *and* to train the DBN. Our method is to (i) compute an entropy-based measure on ultrasound images; (ii) select a small, diverse initial training data set; (iii) train a translational-DBN (tDBN) [16] on this dataset; (iv) bootstrap this initial dataset and retrain the tDBN on the training set expanded with fixed mistakes. This method achieves a significant reduction in human labeling effort.

2. Background

Unlike widely-used benchmark data sets such as `MNIST`, real-world applications often face the *class imbalance problem*, where some classes within the data are over-represented when compared to others. This imbalance can seriously degrade a classifier's performance on the under-represented classes [5]. This has led to an increase in research on solutions to this problem in recent years (see [6] for a review).

With language data, imbalance is especially difficult to avoid. [7] argue that Zipf's law [8], which states that a word's frequency decays as a power-law of its rank, is a consequence of a fundamental organizing principle in language. Therefore we should expect any sample of language or speech data to naturally exhibit class imbalance.

A popular solution is *informed undersampling* [9, 10]: a subset of the data is selected for training according to some set of principles. We require a novel approach because, unlike the typical class imbalance problem, our labels are gradient structural labels (not categorical). No two labels are exactly the same, many are very similar, but a few are very different and unique.

The data set used in this study contains 3,209 ultrasound images taken from 76 short video clips of Italian words uttered in isolation by a single female speaker. In most, the tongue surface is clear, as in Figure 1. Labels consist of a trace of the surface of the tongue, shown in figure 1. The tongue surface is at the bottom edge of the bright white band in the ultrasound image[11, 12]. Linguists often use this type of tongue surface trace to investigate variability in sounds across languages [13], and between speakers of the same language [14].

Hand labeling the tongue surface is expensive. For example, tracing 30 images (1 second of speech) can take 10 minutes or more (about 20 seconds per image) by

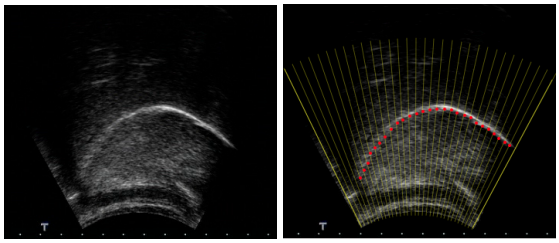

Figure 1: *Left:* Mid-sagittal view of the tongue using ultrasound, tongue tip to the right. The dark areas on the left and right are shadows caused by the hyoid bone and mandible, respectively. *Right:* The DBN is trained to automatically trace the position of the tongue surface.

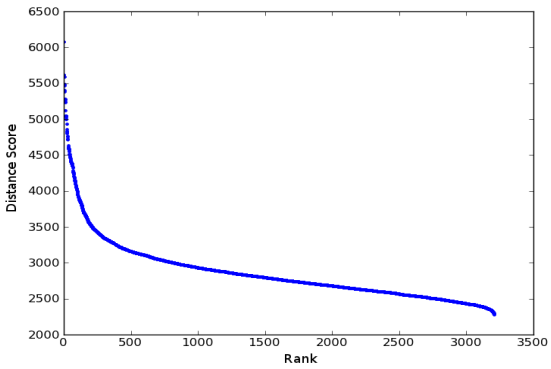

Figure 2: Distance scores ordered by rank.

a trained expert, even when using semi-automated systems such as that of [15]. Using the 'translational' DBN (tDBN) method of [16] reduces this to a fraction of the time (i.e. tracing can be done in real-time). However, training the tDBN requires thousands of pre-traced images: for instance [16] used over 8,600 pre-traced images to train. The method we present here allows us to significantly reduce the amount of training data needed to obtain results comparable to human-labelers when a large pre-labeled training set is unavailable.

The intuition behind using the highest entropy training data set is based on the observation of [3] that the unsupervised pre-training of a DBN acts like a regularizer: deep networks have many non-optimal local minima in which supervised gradient descent learning algorithms can easily become stuck. The problem is exacerbated if the training data is imbalanced: the standard pre-training will likely lead the network to a local minimum that ignores the minority cases. We therefore hypothesize that using a more balanced, high-entropy training set will give the network the best chance of finding a local minimum effective for all tongue shapes.

3. Entropy coding for undersampling

In order to choose a training set, we first score each image according to some criteria. For simplicity, we have chosen a simple distance metric. First, a region of interest is specified, here a 320×290 pixel region centered around the tongue image. An average image is calculated for the entire 3,209 image set by finding the average pixel intensity for each pixel in the region of interest. Finally, for each image, the absolute value of the differences in pixel intensities from the average image are summed to give a distance score for each image. The results of these measurements are shown in figure 2, arranged by rank.

The shape of the curve in figure 2 illustrates the uneven distribution of tongue shapes in the data, suggesting that Zipf's law holds at the level of tongue gestures. Over-representing any single shape will likely bias the DBN to perform well for the over-represented shape, but poorly for other shapes. For this reason, we select the highest-scoring images, which capture a large part of the variability in the tongue shapes. The curve in figure 2 begins to flatten out around 500 images, which take roughly 2.75 hours for a human expert to hand label (20 seconds per image × 500 images = 166 minutes).

Although taking the highest scoring images as the initial training set will represent many diverse tongue shapes, the most common tongue shapes, i.e. those with the lowest distance scores, will likely be underrepresented in the training set. For this reason, we select the 50 *lowest* scoring images to represent the more common tongue shapes. This adds an extra 17 minutes to the hand labeling time, for a total of 183 minutes to hand label the 550 image training set.

A random sample of images is likely to have low entropy (high similarity) in terms of tongue shapes. By selecting the highest scoring images together with some low scoring images, we are constructing a sample with high entropy in terms of tongue shapes, since no single tongue shape will be more represented than any other.

4. Experiments

We compared the performance of a tDBN trained on the 550 high entropy training set to that of a second tDBN trained on 550 randomly selected images. All tongue surface traces were hand labeled by the same human expert. Both tDBNs had layer sizes of 990×1662×1662×8310 nodes for the encoder, and 8310×1662×1662×672 nodes for the decoder. The first 990 node layer in the encoder is the input layer that accepts pixel values from the images (the 320×290 pixel region of interest is scaled to 33×30 = 990 pixels). During training the pixel values were concatenated with the tongue surface traces, which were converted into 32×21 = 672 pixel images. On the decoder network, the regenerated 990 pixels of the image

are discarded, leaving only the 672 node label.

To convert the image-like tDBN outputs into a list of $(\widehat{x}, \widehat{y})$ contour coordinates, for each jth column of the reconstructed contour image of height M, we set $\widehat{x}_j = j$, and let

$$\widehat{y}_j = \frac{1}{M} \frac{\sum_i^M i e^{\widehat{l}_{i,j}/\tau}}{\sum_k^M e^{\widehat{l}_{k,j}/\tau}} \quad (1)$$

where $\widehat{l}_{i,j}$ is the pixel at row i and column j of the reconstruction, and $\tau = 0.05$. The resulting set of coordinates are smoothed in the horizontal direction using local linear regression (Gaussian kernel, standard deviation 1 pixel), and the results are finally scaled up to original size.

Performance measure used the mean sum of distances (MSD) method of [15], shown in equation 2 where, given two curves $U = (u_1, ..., u_n)$ and $V = (v_1, ..., v_n)$

$$MSD(U,V) = \frac{1}{2n} \left(\sum_{i=1}^{n} \min_j |v_i - u_j| + \sum_{i=1}^{n} \min_j |u_i - v_j| \right) \quad (2)$$

4.1. Using a high entropy dataset

Each tDBN was used to trace the entire 3,209 image set; the output was compared to hand-traced labels for each image. Figure 3 shows the MSD scores using the high entropy training set were closer to zero than when images were randomly selected. A t-test showed that the difference between the scores was significant ($p < 0.001$). The high entropy training set had a median MSD score of 1.69 pixels, compared to 2.61 for the random training set.

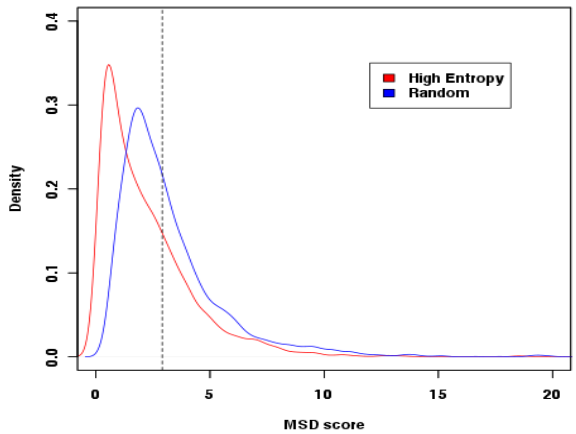

Figure 3: Density plots of the distributions of MSD scores for the entire 3209 image data set, for networks trained on 550 randomly selected images (blue) vs. 500 most diverse and 50 least diverse images (red). Human inter-labeler agreement is shown by the dashed line.

[15] report an average MSD score of 2.91 pixels when comparing labels drawn by different human experts. The high entropy training set gave 898 images with scores higher than 2.91, meaning that the traces needed to be hand-corrected. With the random training set, 1,374 images scored higher than 2.91, an increase of 88%.

Using the high entropy training set, the entire image set can be correctly labeled in less than 8 human hours (3 hours for the initial training set, 5 hours or less for hand correcting). In contrast, tracing all the images by hand would take nearly 18 hours of human labor. The small size of the initial training set reduces compute time for training the tDBN. Training the tDBN with prototype took about 1 hour on a mid-2010 8-core Mac Pro.

4.2. Selecting an optimum training set

We conducted a second experiment to examine the trade-off in human-time between labeling the initial training set and hand-correcting the machine-labeled images. We trained 8 new tDBNs in 100 image increments from the top 100 to 900 highest scoring images, together with the 50 lowest scoring images, which resulted in high-entropy training sets of 150, 250, . . . , 950 images with their hand-traced labels. The new tDBNs had the same number and sizes of layers as the tDBNs discussed above.

Figure 4 shows the results in terms of total human time spent to label the entire 3,209 images, assuming each label or correction takes on average 20 seconds, and labels with MSD score greater than 2.91 need correction.

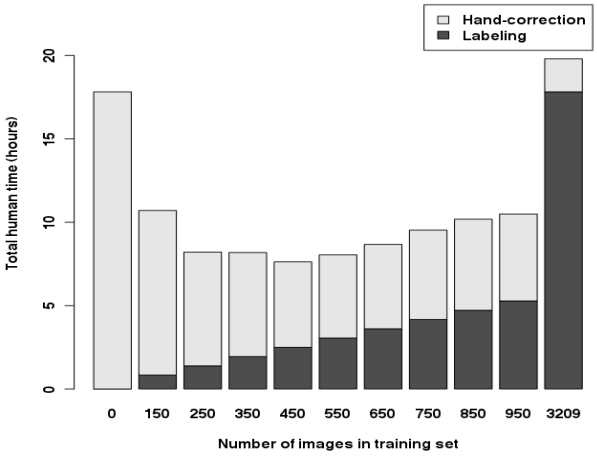

Figure 4: Total human time spent labeling data (initial training + hand-correction) as a function of the number of images labeled in the initial training set. This chart does not reflect any iterative bootstrapping, which further decreases labeling time.

These results show that the optimum number of im-

ages for the training set is 450, using 7.62 human hours (450 training labels + 922 hand-corrections). While training with larger training sets improved the accuracy of the tDBN, the gain in accuracy was not large enough to offset the additional time needed to create the training set.

4.3. Bootstrapping to reduce time

An iterative bootstrapping procedure further reduces the total human time. The newly-trained tDBN is used to label the remaining non-labeled images. These new labels are visually inspected, and the worst ones are hand corrected. These hand corrected labels are then added to the initial training set, and a new tDBN is trained on the supplemented data set.

We tested this method using the best performing tDBN from the previous section. Retraining with the new 550 image training set resulted in a 41% decrease in the number of images needing hand correction, i.e. 536 images with MSD scores above 2.91 for the 450 + 100 hand-corrected network compared to 922 images for the 450 only network. The total human time required for labeling the entire data set using this bootstrapping procedure was 6.03 hours (450 initial images + 100 hand-corrected after 1st training + 536 hand-corrected after 2nd training), compared to 7.62 hours using only the first network.

Iterating this procedure a second time did not decrease the total human time. We suspect that adding more images is causing the natural imbalance in the data to surface, so the network finds a non-optimal local minimum during training. Given the relatively small size of the entire data set, and the flattening of the curve in figure 2 around 400-500 images, it is unsurprising that the imbalance is present in a sample of 650 images.

5. Discussion

We have shown that using an entropy coding method based on informed undersampling for selecting data points for human labeling can dramatically reduce the human time needed to label an imbalanced data set for training a tDBN. The bootstrapping approach of retraining with corrected mistakes further improved accuracy and reduced overall required human labeling time.

The results of the experiments are especially encouraging for the growing community of speech researchers using ultrasound images of the tongue, since labeling images has proven to be a major bottleneck that inhibits researchers conducting larger studies [13, 14], and exploring innovative uses of ultrasound for speech research.

Further, this work contributes to our understanding of deep belief networks in real-world settings. Since DBNs seem to fall prey to biases due to imbalanced data when trained on partially labeled data, methods for overcoming this problem are an important advance in making DBNs practical for use in many real-world problems.

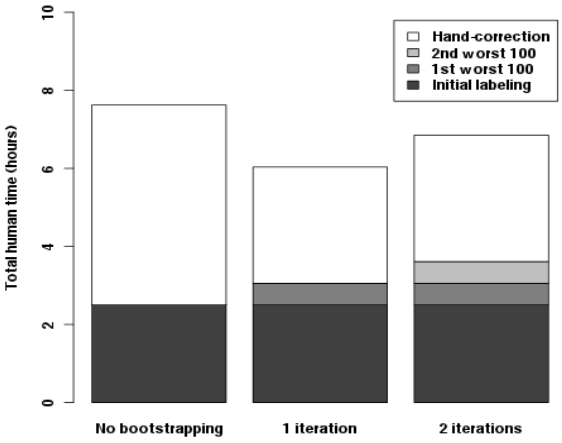

Figure 5: Total human time spent labeling data using bootstrapping procedure.

6. References

[1] Hinton, G.E., and Osindero, S. & Teh, Y.W., "A fast learning algorithm for deep belief nets", Neur. Comp. 7(18):1527–1554, 2006.

[2] Salakhutdinov, R. & Hinton, G. "Learning a nonlinear embedding by preserving class neighbourhood structure", AI & Stat, 3(5), 2007.

[3] Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P. and Bengio, S., "Why Does Unsupervised Pre-training Help Deep Learning?", J Mach Learn Res 11:625–660, 2010.

[5] Chawla, N.V., Japkowicz, N., & Kotcz, A., Editorial: special issue on learning from imbalanced data sets, ACM SIGKDD Explorations Newsletter 6(1):1–6, 2004.

[6] He, H. & Garcia, E.A., "Learning from imbalanced data", IEEE Trans Knowledge & Data Eng, 21(9), 1263–1284, 2009.

[7] Cancho, R.F. & Solé, R.V., Least effort and the origins of scaling in human language, Proc. Nat Acad Sci 100:788, 2003.

[8] Zipf, G., Psycho-biology of languages, Houghton-Mifflin, 1935.

[9] Liu, X.Y., Wu, J. & Zhou, Z.H., "Exploratory undersampling for class-imbalance learning", Systems, Man, and Cybernetics, B: Cyb, 39(2):539–550, 2009.

[10] Drummond, C. & Holte, R.C., "C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling", ICML, Workshop on Learning from Imbalanced Datasets, 2003.

[11] Stone, M., "A guide to analysing tongue motion from ultrasound images", Clin ling & phon, 19(6-7):455–501, 2005.

[12] Iskarous, K., "Detecting the edge of the tongue: A tutorial", Clin ling & phon, 19(6-7):555–565, 2005.

[13] Gick, B., Campbell, F., Oh, S., & Tamburri-Watt, L., "Toward universals in the gestural organization of syllables: A cross-linguistic study of liquids", J Phon 34(1):49–72, 2006.

[14] Mielke, J., Baker, A. & Archangeli, D., Variability and Homogeneity in American English /ɹ/ Allophony and /s/ Retraction, Lab Phon 10, 699–730, Berlin: Mouton de Gruyter, 2011.

[15] Li, M., Chandra, K. & Stone, M., "Automatic contour tracking in ultrasound images", Clin Ling & Phon, 19(6-7), 545–554, 2005.

[16] Fasel, I. & Berry, J., "Deep belief networks for real-time extraction of tongue contours from ultrasound during speech", Int'l Conf Pat Rec, 2010.