Boosting pathology detection in infants by deep transfer learning from adult speech

Anonymous Author(s)
Affiliation
Address
email

Abstract

Can knowledge extracted from adult speech help improve the performance models developed for infant cry? This work investigates this question in the context of pathology detection in newborns. The analysis of infant crying patterns to detect pathology is of interest as it opens the possibility of more accessible diagnostic tools in resource-constrained settings. Classical machine learning approaches leveraging features extracted as Mel frequency cepstral coefficients, have supported the viability of the infant cry as a diagnostic input, but performance is not yet at a level of clinical utility. The application of deep learning models has been limited due to the unavailability of large infant cry databases which are costly to acquire. This work argues that the transfer of useful knowledge from adult speech is possible because it is driven by the same underlying physiologic process as that of infants. Our experiments demonstrate that on the task of predicting perinatal asphyxia from infant cry, such transfer learning provides an overall improvement of 13.5% in F1 score over a model trained from random initialization.

1 Introduction

This work develops a deep convolutional neural network model for cry-based pathology detection in infants using transfer learning from adult speech. The accurate diagnosis of conditions affecting newborns based solely on their cry is important for many practical reasons. It could give rise to low-cost, early and accessible diagnostic tools in resource-constrained settings. We tackle the specific case of predicting perinatal asphyxia, the inability of a newborn to establish regular breathing in the period after birth. Clinical researchers have shown that respiratory conditions like asphyxia alters infant crying patterns since both speech and breathing are coordinated by the same regions of the brain, and newborns do not have voluntary control of either function (1).

Transfer learning (or knowledge transfer) (2) (3) enables the use of knowledge learned from a task in one domain (source) to solve a task in a different domain (target). This is especially useful when the cost of acquiring labelled data in the target domain is very expensive or if the data distribution is constantly changing. Transfer learning relaxes the typical machine learning assumption that the data distribution in the training and test data must be independently and identically distributed. Transfer learning, especially based on deep learning models, has seen prominent success in image, and natural language processing.

Infant cry databases are scarce, and the acquisition/annotation of cries is costly, as it typically must be done in the context of a clinical study. The use of adult speech for transfer learning is thus interesting since large corpuses are freely available. It may not seem intuitive upon initial consideration, but we believe that there is common ground for knowledge transfer from adult to newborn speech. In contrast to newborns, adults have voluntary control over their speech. Furthermore, the way they exercise this control, in terms of accent and speaker mannerisms, has been influenced over several years by the environment. Despite this, the same underlying physiologic mechanism for sound production remains between the brain, vocal chords and respiratory system in both adults and infants. One can argue then that there exists a latent representation of adult speech which captures this mechanism. If our models can learn this, then it is reasonable to expect that knowledge can be transferred.

To test this hypothesis on our target task of predicting perinatal asphyxia based on infant cry, we select the problem of keyword spotting as source task. Keyword spotting (4) involves the identification of commands such as "go", "come", "up", "down", etc from spoken audio. Models for keyword spotting depend on the fact that each word is characterized by a different time-frequency patterns (or signature). Incidentally, this is also the fundamental assumption of how the cries of asphyxiating babies differ from normal ones (1) (5). Furthermore, requirements for keyword spotting models align well with those of cry-based pathology detection models, namely: small computational footprint, dependence on short audio segments, and speaker independence (4).

In this work, we take a state-of-the-art residual network for keyword spotting. We train it for that source task and attempt to transfer the knowledge gained to serve as priors for our target task of detecting perinatal asphyxia. We find that our knowledge transfer model performs significantly better - 18.9% higher sensitivity (or recall), 1.9% higher specificity and 10% higher precision - than the same residual network trained without transfer learning. The overall improvement in F1 score was 13.5%.

2 Tasks

2.1 Source task

The domain of adult speech is large with many corpuses tailored to different tasks. We set 3 main constraints for a desirable source task: 1) it has to be analogous in some form to our ultimate task of pathology detection, 2) there should be freely available corpuses for the task to facilitate pre-training, and 3) there should be openly verifiable benchmarks to enable contextual evaluation pre-trained of models. We select the problem of keyword spotting as our source task. This task is analogous to our task of predicting the presence or not of asphyxia in the sense that it relies on learning the differences in the time-frequency signature of each word of interest. One would thus expect a neural model to focus on extracting similar kinds of high level features. There also exists a reasonably large dataset for this task - Google speech commands (6) for which strong benchmarks (6) (7) have been set.

2.1.1 Corpus

Google speech commands dataset (6) is an audio corpus collected to aid the training and evaluation of keyword spotting models. It contains approximately 65,000 recordings of utterances of 30 keywords from thousands of persons. The recordings are 1-second long, 16-bit PCM-encoded WAV audio files at 16kHz sampling frequency. Typically, work with this dataset has focussed on classifying 10 keywords that are "useful as commands in IoT or robotics applications" (6) (7): "Yes", "No", "Up", "Down", "Left", "Right", "On", "Off", "Stop", and "Go". The remaining words are lumped together to make an *unknown* words category.

2.1.2 Model

We select a CNN architecture which has state of the art performance on the speech commands dataset. To this end, we adopt the 'res8' model from (7). The model takes as input MFCC representation of an input audio, transforms it through a collection of 6 residual blocks, flanked on either side by a convolutional layer, and outputs a 12-way softmax of either of the 10 keywords or *unknown* or *silence* (Figure 1). In the work of (7), 'res8' achieves an accuracy of 94%, which is 1% less than the best model 'res15', but uses only half the number of parameters and 30 times fewer multiplies.

2.2 Target task

Our target task is the detection of perinatal asphyxia from newborn cry. We develop and evaluate our models using the Chillanto infant cry database (5). The database contains 1089 1-second long audio recordings of normal and asphyxiated infants at sampling frequencies of between 8kHz to 16kHz. (5) experimented with audio representations as linear predictive coefficients (LPC) and mel

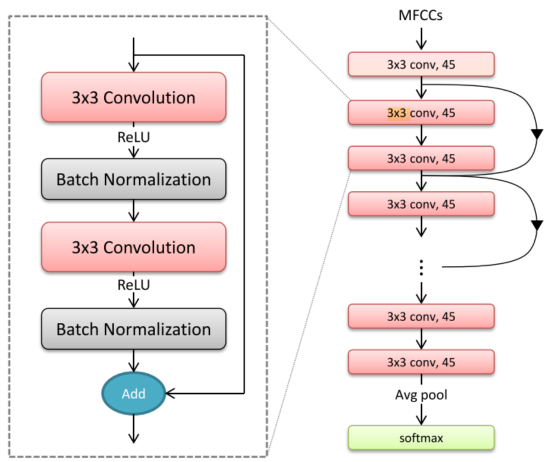

Figure 1: Architecture of convolutional neural network used to train transfer learning model. The 'res8' model employes 6 residual blocks (expanded section to left). Image source (7).

frequency cepstral coefficients (MFCC). Training a time delay neural network as classifier, they achieved best precision and recall in predicting asphyxia of 72.7% and 68% using MFCCs. Building on this work, (8) later showed that precision and recall could be improved to 73.4% and 85.3% by using support vector machines (SVM) which have more stable convergence properties on small but high-dimensional datasets.

3 Methods

3.1 Audio preprocessing

Audio samples are first downsampled to 8kHz for consistency in the input dimension. After downsampling, similar to (7), the audio is converted from 1D to 2D through a sequence of steps: spectrograms were computed for overlapping window sizes of 30ms with a 10ms shift, and across 40 Mel bands. Only frequency components between 20 and 4000 Hz are considered. The discrete cosine transform is then applied to the spectrogram output to compute the MFCCs. The resulting values in each frame is stacked in time to form a spatial, 2D representation of the MFCCs.

3.2 Model pre-training and transfer learning

In order to obtain a pre-trained model for transfer, we train *res8* on the Google speech commands dataset to classify the 10 typical keywords, unknown and silence, using the MFCC "images" as input. As the aim was to first obtain a model with equivalent accuracy as reported in (7), we built upon their open-source implementation (9) and employed the same set of training hyperparameters.

We further create a new instance of the res8 model, res8-transfer, for our task of classifying the presence or not of asphyxia, replacing only output layer to be a binary classifier. To transfer the learned model for our target task, we initialise res8-transfer using the weights from all but the fully-connected output layer of the pre-trained model. We then train the model from this start point. Considering that we had a high class imbalance of 3:1 (normal:asphyxia), we used a weighted sampling procedure to ensure that each batch of training data passed to the model contained roughly the same proportion of both classes. As we had just under a thousand training examples, we also applied data augmentation via time-shifting before transforming an audio into MFCCs.

3.3 Baselines

We implement and compare the performance of our transfer learning model with 2 baseline models for predicting asphyxia. One is a model which trains a radial basis function SVM on MFCC representations of the chillanto database, similar to (8). The other is a res8 model trained on MFCC representations of the chillanto database, but using random initialization of weight (res-no-transfer).

Table 1: Performance of support vector machine (SVM), residual network without transfer learning (res8-no-transfer) and residual network with transfer learning (res8-transfer).

Model	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)
SVM	81.1	87.3	56.6	66.7
res8-no-transfer	73.6	89.6	59.0	65.5
res8-transfer	**92.5**	**91.5**	**69.0**	**79.0**

4 Experiments

4.1 Setup

There were a total of 1389 infant cry samples (1049 normal and 340 asphyxia) in the chillanto dataset. The samples were split proportionally into training, validation and test set in the ratio 60:20:20, ensuring that samples from the same patients were placed in the same set. We trained our models to select the point at which accuracy on the validation set was highest. In addition to accuracy, we tracked other relevant metrics: *recall* or *sensitivity*, the fraction of asphyxia samples correctly identified; *specificity*, the fraction of normal samples correctly identified; *precision*, the fraction of asphyxia predictions that were correct; and *F1 score*, the harmonic mean of precision and recall.

4.2 Results

We trained the 2 baseline models: an SVM trained on flattened MFCC features and a res8 model trained from scratch without any pre-training (res8-no-transfer). Both models were trained to identify asphyxia and normal samples from chillanto database. Results are shown in Table 1. The SVM[1] attains a sensitivity in detecting asphyxia of 81.1% which is 7.5% higher than that of res8-no-transfer. Though res8-no-transfer performs slightly better on the other metrics of specificity and precision, the SVM is the overall better model based on the F1 score of 66.7%.

To carry out our transfer learning experiment, we first pre-train a res8 model on the (downsampled) Google speech commands dataset. Our model achieved the same test set accuracy of 94.4%. Using the learned weights from this model, we initialized a new res8 instance and further trained it to classify normal and asphyxiated babies from the chillanto dataset. We tuned hyperparameters using our validation set. This model was trained for 30 epochs using a stochastic gradient descent (SGD) optimizer with start learning rate of 0.001, a schedule to a rate of 0.0001 after 15 epochs, a fixed momentum of 0.9, and batch size of 32. Hyperparameters were primarily derived from (6) and (7), but modified where necessary to suit chillanto dataset. We minimized the hinge loss as we found that it resulted to better accuracy on the validation set than cross-entropy loss. The performance of this model (res8-transfer) is summarized in table 1. It performs better than the others on all metrics, achieving sensitivity and specificity of 92.5% and 91.5%, respectively at a precision of 69%.

5 Discussion

We proposed and empirically tested the hypothesis that adult speech and infant cry are driven by the same physiologic mechanism whose elements can be captured in model. Such transfer learning, when compared to an equivalent deep model trained without, led to improvements of almost 20% in sensitivity, 2% in specificity and 10% in precision of detecting perinatal asphyxia.

The precision of 69%, despite the high sensitivity, indicates that there remains a fair number of false positives. While future models should aim to address this, the false identification of a baby as having asphyxia is not clinically dangerous since it implies that the infant may receive additional care such as increase oxygen support and neuro-protective strategies.

[1]The SVM in this work performed slightly worse than in referenced work (5) (8). We believe that the authors have performed biased split in which samples from the same subject occur across training, validation and test sets. A model developed in such way is an overestimate of actual performance.

References

[1] K. Michelsson, P. Sirviö, and O. Wasz-Höckert, "Pain cry in full-term asphyxiated newborn infants correlated with late findings," vol. 66, pp. 611–6, 10 1977.

[2] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, oct 2010.

[3] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A Survey on Deep Transfer Learning," Tech. Rep., 2018.

[4] T. N. Sainath and C. Parada, "Convolutional Neural Networks for Small-footprint Keyword Spotting," Tech. Rep., 2015.

[5] O. F. Reyes-Galaviz and C. A. Reyes-Garcia, "A System for the Processing of Infant Cry to Recognize Pathologies in Recently Born Babies with Neural Networks," in *SPECOM'2004: 9th Conference on Speech and Computer*, St Petersburg, Russia, 2004.

[6] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," Tech. Rep., 2018.

[7] R. Tang and J. Lin, "Deep Residual Learning for Small-footprint Keyword Spotting," Tech. Rep., 2018.

[8] C. C. Onu, "Harnessing infant cry for swift, cost-effective diagnosis of perinatal asphyxia in low-resource settings," in *Humanitarian Technology Conference-(IHTC), 2014 IEEE Canada International*. IEEE, 2014, pp. 1–4.

[9] R. Tang and J. Lin, "Honk: A PyTorch Reimplementation of Convolutional Neural Networks for Keyword Spotting," Tech. Rep., 2018.