Action Recognition using Micro-Doppler Signatures and a Recurrent Neural Network

Jeff Craley*†, Thomas S. Murray*, Daniel R. Mendat*†, Andreas G. Andreou*†
*Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA
†Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland 21218, USA
Email: jcraley@gmail.com, andreou@jhu.edu

***Abstract*—This paper explores the long short-term memory (LSTM) recurrent neural network for human action recognition from micro-Doppler signatures. The recurrent neural network model is evaluated using the Johns Hopkins MultiModal Action (JHUMMA) dataset. In testing we use only the active acoustic micro-Doppler signatures. We compare classification performed using hidden Markov model (HMM) systems trained on both micro-Doppler sensor and Kinect data with LSTM classification trained only on the micro-Doppler signatures. For HMM systems we evaluate the performance of product of expert based systems and systems trained on concatenated sensor data. By testing with leave one user out (LOUO) cross-validation we verify the ability of these systems to generalize to new users. We find that LSTM systems trained only on micro-Doppler signatures outperform the other models evaluated.**

I. Introduction

Human actions occur in three-dimensional space and evolve over time. Most actions of note involve complicated sequences of simple motions. Classifying these actions requires systems capable of learning the time dependencies between these simpler motions in a high dimensional setting. In this paper, we investigate LSTM and HMM based action recognition systems using micro-Doppler signals. These experiments rely on data from the Johns Hopkins University MultiModal Action (JHUMMA) dataset [1]. The dataset contains both Microsoft Kinect and micro-Doppler recordings of a set of 21 actions performed by different actors. In both the HMM and LSTM classification systems only the micro-Doppler recordings were used for classification. However, the HMM system relied on the Microsoft Kinect sensor data for training while the LSTM system did not. We find that the LSTM based systems outperform the HMM based classifiers.

II. Dataset

The shift in frequency observed when either the source or observer of a sound is moving is known as the Doppler effect [2]. If the object itself contains moving parts, each part contributes its own Doppler shift proportional to the object's radial velocity component with respect to the receiver. All of the scattered waves are additive, and the resulting modulation is a superposition of the individual components known as the *micro-Doppler* effect [3]. The acoustic micro-Doppler effect was independently reported in 2007 by Zhang et. al. [4]. The micro-Doppler effect results in a reflected signal that is a combination of frequency, amplitude and phase modulation; by applying a short term Fourier transform (STFT), the changes in frequency are more readily apparent. Figure 1 shows an example taken from the JHUMMA of a micro-Doppler spectrogram of an human actor walking and pivoting.

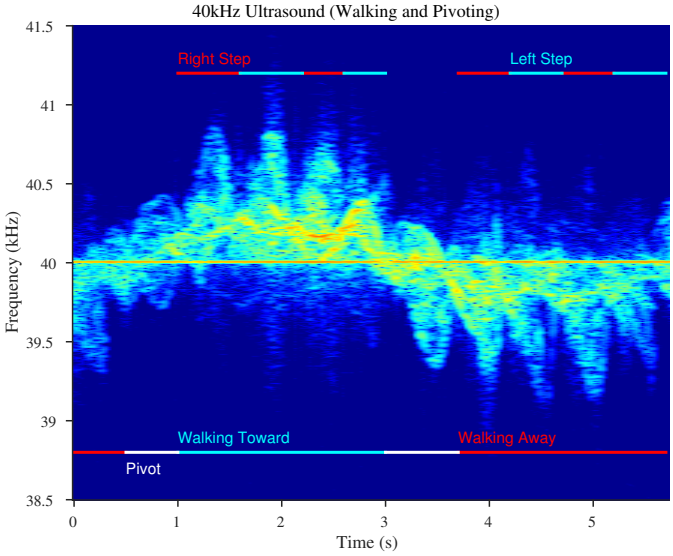

Fig. 1. Annotated spectrogram representation of Doppler modulations for a human walking toward an ultrasound sensor, pivoting, and walking back away from it.

The JHUMMA dataset [1] used in this study contains micro-Doppler signatures from three ultrasound sensors originally discussed in [5] as well as Microsoft Kinect RGB-D data [6], [7]. From the Kinect data, only the skeletal data component was used. A schematic of the individual sensor locations is shown in Figure 2. All actions were performed within a demarcated bounding box referenced to a virtual north. This area was located on a large open stage to minimize uninteresting reflections from nearby objects. The three ultrasound sensors emitted frequencies at 25kHz, 33kHz, and 40kHz, and were located to the east, west, and north of the bounding box, respectively. The Kinect sensor was placed directly on top of the 40kHz sensor. The short term Fourier transform (STFT) was applied to waveforms recorded from each sensor. The STFT representation was band limited to 1.5kHz above and below the carrier frequency,

978-1-5090-4780-2/17/$31.00 ©2017 IEEE

resulting in 327, 328, and 328 spectrogram frequency bins respectively. The 21 actions recorded in the JHUMMA dataset are enumerated in Table II along with their orientations in the bounding box. These actions were recorded by 13 actors. However, 30 repetitions are missing resulting in a total of 2700 recorded actions.

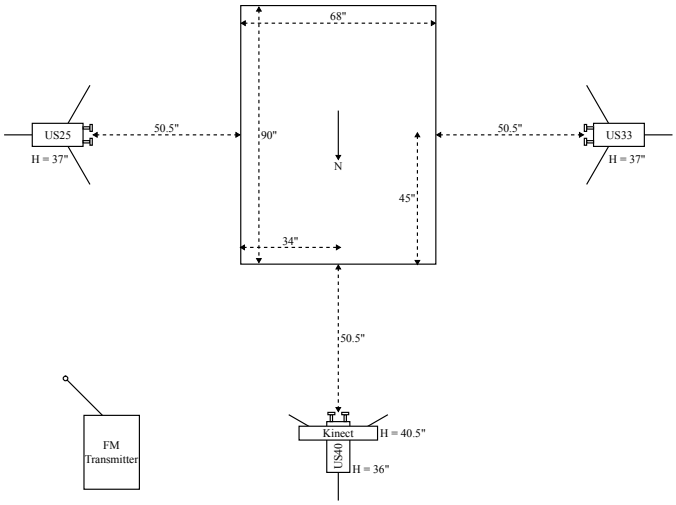

Fig. 2. Schematic of the sensor locations for the JHUMMA data collection

Action	Orientation	Repetitions / Duration
Lunges	N	10 Per Leg (Alternating)
Lunges	NE	10 Per Leg (Alternating)
Lunges	NW	10 Per Leg (Alternating)
Left Leg Steps	N	10
Right Leg Steps	N	10
Left Arm Raise (Forward)	N	10
Left Arm Raise (Sideways)	N	10
Right Arm Raise (Forward)	N	10
Right Arm Raise (Sideways)	N	10
Walk in Place	N	20 Steps
Walk Facing Forward	N-S	10 cycles
Walk Facing Sideways	W-E	10 cycles
Walk and Pivot	NE-SW	10 cycles
Walk and Pivot	NW-SE	10 cycles
Jumping Jacks	N	10
Jump Rope	N	10
Body Squats	N	10
Jump Forward then Backward	N-S	10 sets
Jump Forward then Backward	NE-SW	10 sets
Jump Forward then Backward	NW-SE	10 sets
Punch Forward	N	10 Per Arm (Alternating)

TABLE I
SCRIPT OF ACTIONS NOMINALLY DEMONSTRATED BY EACH ACTOR.

III. METHODS

We trained four classifiers in total, two HMM based and two LSTM based. The HMM classifiers were trained using both spectrogram data from the ultrasound sensors and skeletal data from the Kinect sensor. The LSTM classifiers were trained using spectrograms alone. The classifiers were evaluated using two different cross validation schemes. The first divides the data set into 5 folds, where two repetitions of each action are set aside as test data for each user. The second cross validation scheme uses a leave one user out approach (LOUO) to ensure that classification generalizes to new users.

A. Hidden Markov Model

In our experiments we use two HMM [8] based models. In each model, the ultrasound spectrogram recordings are regarded as visible emissions and the skeletal positions recorded by the Kinect are regarded as latent states. Both the spectrograms and the skeletal positions were clustered using k-means. Thus the emissions and latent states are both discrete.

We regard the product of experts (POE) model detailed in [9] as a baseline upon which to compare our further experiments. In this system, an HMM for each action is trained for each sensor. At test time, the clustered spectrogram recordings are shown to each model and the log-likelihood is found. Denoting the sequence of spectrogram clusters as X, and denoting the log-likelihoods of a given action for the 25kHz, 33kHz, and 40kHz, as $\mathcal{L}_a^{25}(X)$, $\mathcal{L}_a^{33}(X)$, and $\mathcal{L}_a^{40}(X)$ respectively, the action is classified using a product of experts decision rule as shown in (1). Following [9], this system uses 200 skeletal clusters and 100 spectral clusters in each HMM model.

$$\hat{a} = \arg\max_a \left(\mathcal{L}_a^{40}(X) + \mathcal{L}_a^{33}(X) + \mathcal{L}_a^{25}(X) \right). \quad (1)$$

Noting that the POE baseline failed to discriminate between some diagonally oriented actions in previous experiments, an HMM system was trained using a concatenation of the spectrogram data for all sensors. It was hypothesized that this concatenation would enable the model to learn the coupling of forward-facing and lateral motion required to discriminate between diagonal directions. To accommodate the increased size of the concatenated spectrogram feature vector, 200 spectral k-means clusters were used in the stacked HMM system.

B. Long Short-Term Memory Model

The LSTM model [10]–[12] is a recurrent neural network architecture capable of learning complex dependencies across time. Equations 2-7 detail the procedure to update each node at a given timestep [13]. In these equations i_t, f_t, and o_t represent the value of the input, forget, and output gates respectively. $\tilde{C}_t$ represents the update to the hidden state and C_t represents the current hidden state. h_t is the output of a given node. Each node of the LSTM network maintains a hidden state that is updated at each timestep. In addition, each node contains an input, output, and forget gate, capable of controlling the behavior of the node depending on the current value of the hidden state. This architecture is thus capable of learning time dependencies in the data that a feedforward neural net is incapable of learning. A schematic diagram of the LSTM node architecture is shown in Figure 3. A diagram of an LSTM network with one hidden layer unrolled over time is shown in Figure 4.

$$i_t = \sigma\left(W_i X_t + U_i h_{t-1} + b_i\right) \tag{2}$$

$$f_t = \sigma\left(W_f X_t + U_f h_{t-1} + b_f\right) \tag{3}$$

$$o_t = \sigma\left(W_o X_t + U_o h_{t-1} + b_o\right) \tag{4}$$

$$\tilde{C_t} = \tanh\left(W_c X_t + U_c h_{t-1} b_c\right) \tag{5}$$

$$C_t = i_t \cdot \tilde{C_t} + f_t \cdot C_{t-1} \tag{6}$$

$$h_t = o_t \cdot \tanh\left(C_t\right) \tag{7}$$

In our experiments, we used a Keras [14] implementation of the LSTM running on a Theano backend [15]. In training, the normalized spectrogram was shown to the LSTM as well as the action label associated with the recording at each timestep. The skeletal pose data was not used. The training parameters used are shown in Table II. We used the Adam optimizer with a dropout of 0.5. Each of the networks were trained for 200 epochs. A categorical cross entropy loss was used. Two LSTM networks were evaluated. A one-layer architecture with 1200 hidden nodes and a two-layer architecture with two 800 hidden node hidden layers were used in our experiments.

Default Training Parameters	
Number of epochs	200
Seed	1337
Optimizer	Adam
Dropout	0.5
Max Sequence Length	404
Batch Size	100

TABLE II
TRAINING PARAMETERS FOR LSTM NETWORKS

C. Cross-Validation

Two cross-validation schemes were used in our experiment. The first cross validation scheme follows the previously reported results in [9] and divides the dataset into 5 folds. Whenever possible 2 examples of each action performed by each user were placed into all folds. Each of the folds was used for testing once, while the remaining 4 were used in training. Thus examples of each actor performing each action are present in both test and training sets. Noting the great degree of similarity in a given actor's set of performances of a single action, this potentially allows the classifiers to overfit to actor specific performances. In order to test the ability of the classifiers to generalize to new actors, a leave one user out cross validation scheme was also used. Each actor's data was used for testing once, while the remaining data was used for training. Given the 13 total actors, this resulted in 13 folds. In both cases results reported have been averaged across all folds.

Model	5 Fold	LOUO
POE Baseline	87.8889	67.963
Stacked HMM	93.6296	89.0
LSTM 800	98.1852	95.6296
LSTM 1200	97.5926	95.7037

TABLE III
CLASSIFICATION ACCURACY FOR ALL MODELS

IV. RESULTS

A. Five Fold Cross Validation

Confusion matrices for the classification systems are shown in Figure 5 and Figure 6. The product of experts baseline model achieved a classification accuracy of 87.8889%, suffering from confusion in the diagonally oriented walk and pivot actions. This confusion was resolved in the stacked data HMM classifier, which achieved a classification accuracy of 93.6296%. The LSTM classifiers further outperformed the HMM POE baseline, with the two layer LSTM 800 classifier achieving an accuracy of 98.1852%.

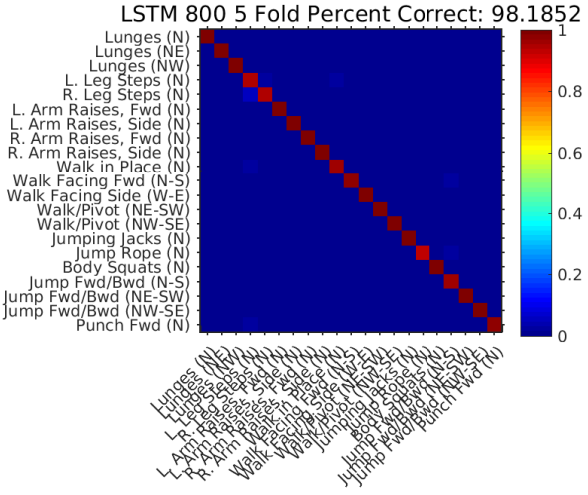

Fig. 5. Confusion matrix for LSTM with two 800 hidden node layers

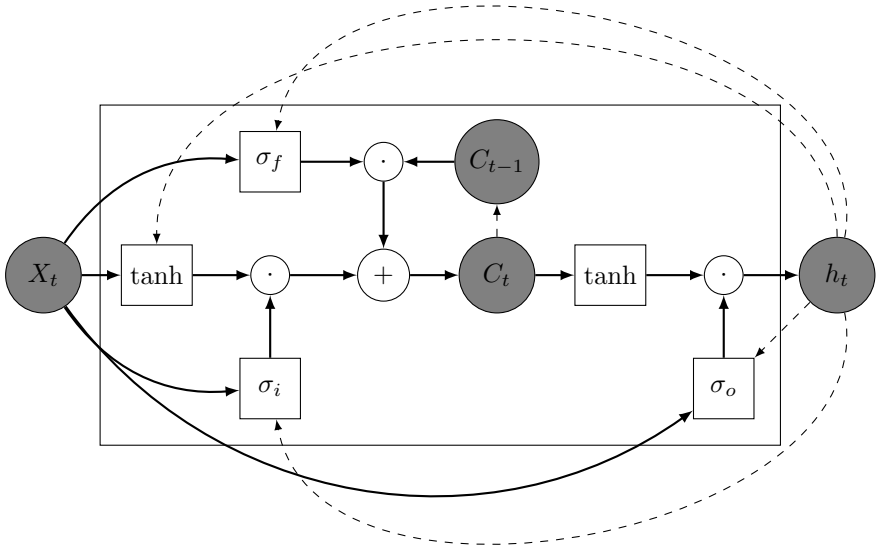

Fig. 3. Schematic of an LSTM node. Dashed lines indicate a recurrent connection.

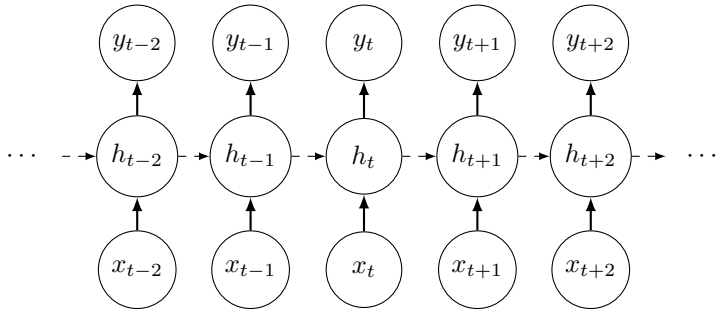

Fig. 4. Schematic of an LSTM network unrolled over time. Dashed lines indicate a recurrent connection.

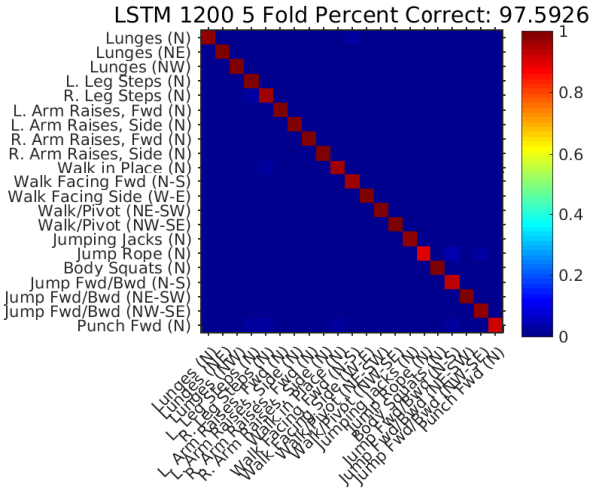

Fig. 6. Confusion matrix for LSTM with one 1200 node hidden layer

B. LOUO Cross Validation

Confusion matrices for the classification systems are shown in Figure 7 and Figure 8. The LOUO cross validation scheme resulted in lower overall classification accuracies. The baseline POE HMM model performed substantially worse, achieving a classification accuracy of 67.963%. However the stacked HMM model still performed relatively well achieving an accuracy of 89.0%. Again the LSTM classifiers outperformed the HMM classifiers, with the one layer LSTM 1200 model achieving a slightly higher accuracy of 95.7037%.

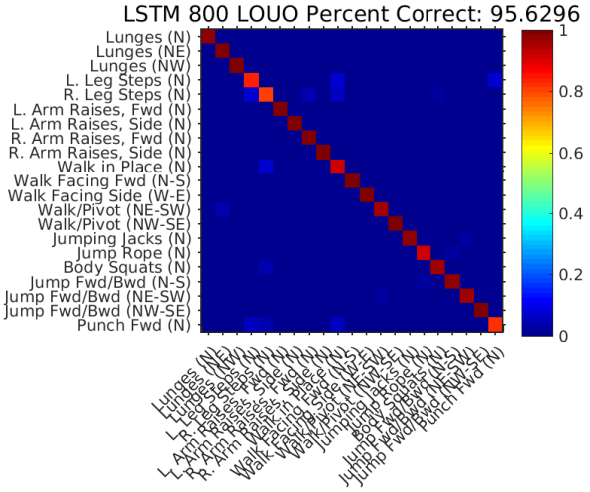

Fig. 7. Confusion matrix for LSTM with two 800 hidden node layers

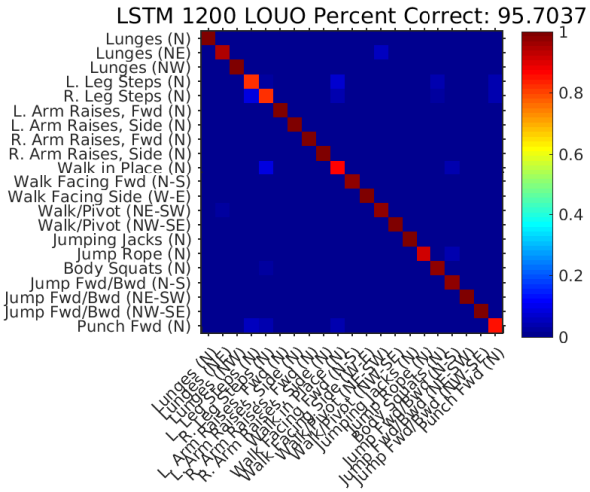

Fig. 8. Confusion matrix for LSTM with one 1200 node hidden layer

V. Conclusions

In this work we have shown the applicability of recurrent neural networks and specifically the long short-term memory model for action recognition using micro-Doppler signatures. The LSTM classifiers used in this work outperform the HMM based classifiers by substantial margins. Furthermore, the LSTM classifiers are trained without using the Kinect data required in the HMM models. In addition, we have found that a single classifier using a concatenation of data from each sensor outperforms a product of experts based model. We believe this gain in performance is due to the coupling of the lateral and forward-backward motion captured by the concatenated data. A product of experts model trained on separate sensors is unable to correctly capture the correlations between orthogonal dimensions present in diagonal motion.

Acknowledgment

This work was partially supported by the NSF grant INSPIRE SMA 1248056 through the Telluride Workshop on Neuromorphic Cognition Engineering, by the National Science Foundation grant IIS-1344772, SCH: INT: Mapping the Cardiac Acousteome: Biosensing and Computational Modeling Applied to Smart Diagnosis and Monitoring of Heart Conditions and by an ONR MURI N000141010278. Dan Mendat was supported by the Johns Hopkins University Applied Physics Laboratory Graduate Student Fellowship and Jeff Craley by a Northrop Grumman Graduate Fellowship. We also thank Jack Riddle of NG for his personal interest and support for this work.

References

[1] T. S. Murray, D. R. Mendat, P. O. Pouliquen, and A. G. Andreou, "The Johns Hopkins University Multimodal Dataset for Human Action Recognition," in *Proceedings of SPIE: Radar Sensor Technology XIX; and Active and Passive Signatures VI*, May 2015, pp. 79–94.

[2] C. Doppler, "Über das farbige Licht der Doppelsterne und einiger anderer Gestirne des Himmels (English Translation)," *Proceedings of the Royal Bohemian Society of Sciences*, vol. 2, pp. 465–482, 1842.

[3] V. C. Chen, F. Li, S. S. Ho, and H. Wechsler, "Micro-Doppler effect in radar: phenomenon, model, and simulation study," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 42, no. 1, pp. 2–21, 2006.

[4] Z. Zhang, P. O. Pouliquen, A. M. Waxman, and A. G. Andreou, "Acoustic micro-Doppler radar for human gait imaging," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. EL110–3, Mar. 2007.

[5] J. Georgiou, P. O. Pouliquen, A. S. Cassidy, G. Garreau, C. M. Andreou, G. Stuarts, C. d'Urbal, S. L. Denham, T. Wennekers, R. Mill, I. Winkler, T. M. Bohm, O. Szalardy, G. M. Klump, S. Jones, A. Bendixen, and A. G. Andreou, "A multimodal-corpus data collection system for cognitive acoustic scene analysis," in *Proceedings of the 45th Annual Conference on Information Sciences and Systems (CISS)*, Mar. 2011, pp. 1–6.

[6] G. Yahav, G. Iddan, and D. Mandelboum, "3D imaging camera for gaming application," in *Consumer Electronics 2007*, 2007, pp. 1–2.

[7] B. Freedman, A. Spunt, and Y. Ariell, "Distance-varying illumination and imaging technique for depth mapping," Patent, Jun., 2014.

[8] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[9] T. S. Murray, D. R. Mendat, K. Sanni, P. O. Pouliquen, and A. G. Andreou, "Bio-inspired Human Action Recognition With A micro-Doppler Sonar System -preprint- ," *IEEE Access*, pp. 1–16, 2016.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[11] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000.

[12] F. Gers, "Long Short-Term Memory in Recurrent Neural Networks," Ph.D. dissertation, Ph.D. Dissertation Ecole Polytechnique Federale de Lausanne, 2001.

[13] A. Graves, "Generating Sequences With Recurrent Neural Networks," *arXiv.org*, Aug. 2013.

[14] F. Chollet. (2017, Feb.) Keras: Deep Learning library for Theano and TensorFlow. [Online]. Available: https://keras.io

[15] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for scientific computing conference (SciPy)*, 2010.