INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Audiovisual Correspondence Learning in Humans And Machines

Venkat Krishnamohan[1], *Akshara Soman*[1], *Anshul Gupta*[2], *Sriram Ganapathy*[1]

[1]Learning and Extraction of Acoustic Patterns (LEAP) lab, Indian Institute of Science, Bangalore.
[2]Mercedes-Benz Research and Development, Bangalore, India.

sriramg@iisc.ac.in

Abstract

Audiovisual correspondence learning is the task of acquiring the association between images and its corresponding audio. In this paper, we propose a novel experimental paradigm in which unfamiliar pseudo images and pseudowords in audio form are introduced to both humans and machine systems. The task is to learn the association between the pairs of image and audio which is later evaluated with a retrieval task. The machine system used in the study is pretrained with the ImageNet corpus along with the corresponding audio labels. This model is transfer learned for the new image-audio pairs. Using the proposed paradigm, we perform a direct comparison of one-shot, two-shot and three-shot learning performance for humans and machine systems. The human behavioral experiment confirms that the majority of the correspondence learning happens in the first exposure of the audio-visual pair. This paper proposes a machine model which performs on par with the humans in audio-visual correspondence learning. But compared to the machine model, humans exhibited better generalization ability for new input samples with a single exposure.

Index Terms: Audiovisual correspondence learning, few-shot learning, multimodal learning, transfer learning, human-machine comparison

1. Introduction

The fusion of multimodal signals, i.e., signals measured in multiple domains, has been an area of considerable interest for both humans [1] and machines [2]. In this paper, we consider a type of fusion problem that pertains to learning how different modalities correspond to each other. The modalities under consideration are the audio and image domains.

In human cognition, cross-modal correspondence learning can be defined as the learning of the mapping that observers expect to exist between two or more features or dimensions from different sensory modalities (such as the shape of the visual object and associated speech phonemes) [3]. Once the learning is achieved, stimulation in one modality can elicit experiences in the other sensory modality (for example, sound-color associations [4]) which can also extend to behavioral changes and cross-modal retrieval [5]. However, many questions exist on how efficient humans are in learning object-audio associations for previously unknown shapes and sounds. In this paper, we attempt to highlight the behavioral performance for human subjects in learning cross-modal audio-visual correspondences for pseudo images and pseudo-words and the ability of humans to generalize to different orientations and color changes or speaker changes in the speech data. Past studies have shown that humans require only a very small number of instances to learn meanings of new words [6].

This work was supported in part by the grants from Department of Atomic Energy project 34/20/12/2018-BRNS/34088

It is therefore of significant interest to question whether machines can achieve human-level efficiency with limited data. In machine learning, cross-modal correspondence modeling and retrieval has received significant attention in recent years, in particular between domains of text and image [7]. For audio-visual cross-modal modeling, early work by Zhang et al. [8] focused on correlation-based modeling. Recent work using deep learning by Arandjelovic et al. [2] attempted to learn the association between audio and images in an unsupervised manner using video data. Chrupala et al. [9] performed joint semantic modeling of speech utterances and images and showed that semantic and form-related information of speech is encoded across the model hierarchy. Kamper et al. [10] trained a speech network using soft tags from images for better qualitative performance than a supervised model trained on transcriptions in a semantic speech retrieval task. Harwath and Glass [11] showed that a network trained jointly on images and spoken captions learns to associate segments of the audio and the corresponding semantically relevant regions in the image. Other studies [12] have also looked at the development of algorithms for cross-modal data generation. Our work is related to Eloff et al. [13] who approached multimodal one-shot learning through unimodal comparisons of the query and matching set with a support set. In our previous work [14], we explored a rapid language learning task where human subjects attempt to learn a set of words from a new language with image supervision. A machine comparison for this task revealed that machines can achieve human-level performance for this task on previously known image classes.

In this paper, we explore a learning paradigm where novel objects and novel sounds are associated. This novelty is critical for ensuring a direct human-machine comparison as the novelty guarantees that we are not merely tapping into knowledge acquired prior to the experiment in humans.

2. Human Experiment

2.1. Stimuli

The objects and labels used are primarily from the Novel Object and Unusual Name (NOUN) Database [15]. The NOUN dataset contains 60 novel objects and 173 pseudowords. We used the 60 objects as our novel classes and paired it with a pseudoword label. The label is a trisyllabic pseudoword generated by combining the original NOUN pseudowords. We conjecture that spoken pseudowords with one or two syllables are too short and lack adequate information. The original images from the NOUN dataset were augmented by performing RGB permutations, horizontal and vertical flipping, and rotations. For each object class, 10 such augmented variants of the original image were added to the stimuli set. The examples of the stimuli used are given in Fig. 1. We generated 12 audio samples per label from their synthesized speech, using Google [16], IBM [17],

http://dx.doi.org/10.21437/Interspeech.2020-2674

Figure 1: *Examples of the stimuli used in the experiment. Columns 3,4,5 show augmented variants of the image corresponding to the audio label.*

and Microsoft [18] Text-to-Speech (TTS) systems. These audio samples had variations as the TTS models with different gender and accents (American English, British English and Australian English) were used. An informal listening test of the files showed no major inconsistencies in pronunciation across the models.

2.2. Subjects

The participants were Indian nationals with self-reported normal hearing and vision. Twenty four adults participated in this study (mean age = 24.08, age span = 22-31). All subjects provided written informed consent to take part in the experiment and received monetary compensation. The Institute Human Ethical Committee of Indian Institute of Science, Bangalore approved all procedures of the experiment.

2.3. Experimental Setup

The stimuli were randomly divided into 6 blocks with 10 novel classes in each block. During the experiment, these blocks appeared in a different random order for different subjects. Each block was a n-shot learning paradigm of 10 novel objects ($n = 1, 2$ or 3). An n-shot learning task means the participant is exposed to n augmented variants of the stimulus before the evaluation. There were 2 blocks for each n-shot learning. The experiment starts with a practice session to familiarize participants with the experiment setup. The practice session used 3 objects as stimuli, which are not part of the main experiment. They were from the Fribbles stimuli set [19], and labels were sampled from the trisyllabic word pool. Each block in the main experiment had two phases: learning phase and testing phase.

Learning Phase: In the learning phase, the subjects were presented with image-audio pairs for each class in the session. Classes within a session were randomly shuffled. For the two-shot and three-shot sessions, the 10 classes were presented in cycles, in a random order within each cycle. No limit was imposed on the learning time per sample, or the number of times the audio clip could be heard. For each subject, we recorded the total time spent and the number of audio play button clicks on each sample.

Testing Phase: During the testing phase, the subjects were asked to perform two tasks: an image retrieval (IR) and an audio retrieval (AR) task. In IR, the subject had to pick the matching image for the given audio sample from a set of 10 image choices. In the AR, the subject had to pick the corresponding audio clip for the given image sample from a set of 10 audio choices. For each session, both tasks included 10 test cases (as each block had 10 novel objects to learn). The image and audio samples used in the test phase were augmented versions of their training counterparts. Further, one-shot learning blocks had an additional test case for each of the 10 objects showed in the training phase. In this testing case, the query sample was the same image/audio sample that appeared during the learning phase. Hence, the subject had already seen this exact sample before the evaluation. But the multiple-choice answer options include augmented versions of samples from the other modality. This strategy enables us to compare the generalisation capability of humans and machines on the one-shot task.

Table 1: *Machine model experiment: Details of the dataset used for pretraining. The number of samples per class is also given.*

Data	Train	Validation	Test
Image (ImageNet classes)	160 ImageNet train	16 ImageNet val	16 ImageNet val
Image (Other classes)	60 30 Google 30 Flickr	20 10 Google 10 Flickr	20 10 Google 10 Flickr
Speech (TTS voice) (English)	22 10 Google 1 IBM 11 Microsoft	5 2 Google 1 IBM 2 Microsoft	5 2 Google 1 IBM 2 Microsoft

3. Machine Experiments

3.1. Dataset

For pretraining, we use the 655 classes from [14]. These classes have labels of one-word length. Images are obtained from the ImageNet database [20], and the Flickr and Google image repository. The audio recordings of the labels are generated using the Google, Microsoft [18] and IBM [17] TTS systems.

For the novel object learning task, we use the same 60 novel classes as in the human experiment. The data split per class, as ($train$, $test$), is (n, 10-n) for images with a total of 10 images per class. For audio, it is (n, 12-n) with a total of 12 audio variants per class. Here, n = 1,2 or 3 for n-shot learning.

3.2. Audio-Visual Semantic Network

We use the joint audio-visual model [14] illustrated in Fig. 2. It consists of audio and image sub-networks which are jointly trained on the multimodal input.

Audio sub-network: The audio sub-network has two long short-term memory (LSTM) layers followed by fully connected (FC) layers. The audio is fed as 80-dimensional bottleneck features, from a deep neural network trained for automatic speech recognition (ASR) on the Switchboard and Fisher corpora [21]. After pretraining, only the final fully connected layer of 576 dimensions is trained on the novel audio labels.

Image sub-network: The image sub-network uses the Xception network [22] which is trained on the ImageNet classes. We use the 2048-dimensional pre-softmax layer from the Xception network as the input representation. This input is then mapped to a 576-dimensional latent space using a fully connected layer which is trained jointly with the audio sub-network.

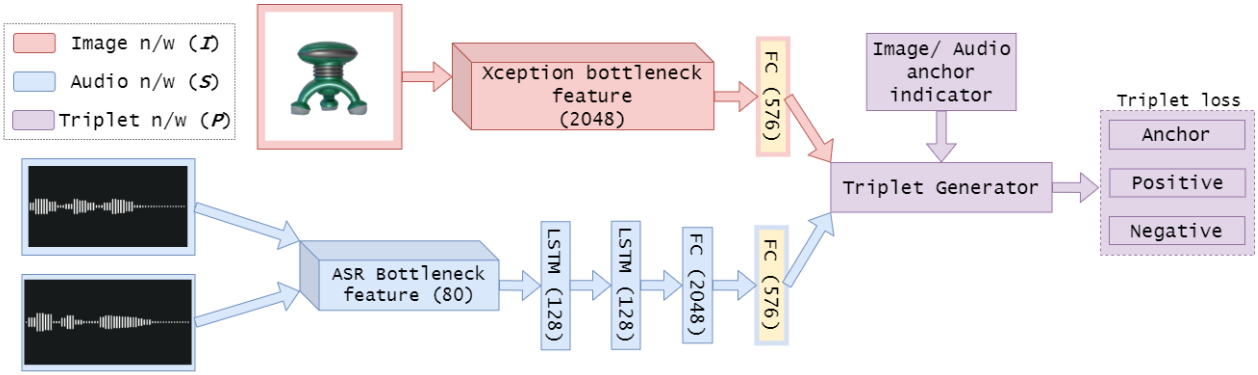

Figure 2: *Joint audio-visual semantic network with triplet loss (eg. with image as anchor). Yellow indicates the layers that are trained.*

Table 2: *Machine experiment: Top-k image and audio retrieval Accuracy over pretraining classes. Chance acc. - 0.15%*

Retrieval	Top-1	Top-5	Top-10
Image	72.40	84.17	87.10
Audio	70.46	84.10	87.13

3.3. Training

Pretraining: The 576-dimensional FC layers of the audio and image subnetworks are trained jointly on 655 pretraining classes. We use the modified proxy based approach of [23][14] to train using multimodal input. Similarity is maximized between the input representation and matching proxy vector, and minimized against the non-matching proxy vectors. In the first stage, the proxy matrix and FC layer of the image subnetwork are trained on the image data. The proxy matrix is then kept fixed. In the second stage, the FC layer of the audio subnetwork is trained using the audio data. We minimize the NCA loss for training [24] and train for 100 epochs with a learning rate of 0.001. We use the Adam optimizer with batch-norm and dropout. Results for retrieval on the pretraining classes are given in table 2.

Transfer Learning: The data for the novel classes is divided into blocks of 10 classes. Our machine setup is trained on the data of one block at a time to emulate the human experiment. For two-shot and three-shot learning, the data is repeated in cycles. Since the blocks contain novel independent objects and labels, we train and test each block separately and not in an incremental fashion. Like the human experiment, the order of classes is randomized. We use a triplet based approach for training, and maximize the similarity between a matching image/audio pair while minimizing the similarity between non-matching image/audio pairs. Similarity is given by:

$$S_{j,k} = -||y_j - x_k||_2^2 \tag{1}$$

where y_j and x_k are the L2 normalized embedding for image j and audio k respectively. The triplet loss is given by,

$$C(\theta) = S_{a,n} - S_{a,p} + \alpha \tag{2}$$

where θ are the model parameters, a is the anchor point, p and n are the corresponding positive and negative points, and α is the margin which we set at 0.8. The model is trained to convergence on one class before moving to the next. In one-shot learning, the anchor and positive data samples are the image/audio samples used in the human experiment with negative data sampled from the pretraining classes. During the two-shot and three-shot learning phases, the anchor is the novel image/audio of the current cycle, positive points are sampled from the current and previous cycles, and negative points are sampled from the previous cycles and pretraining classes. We train on each novel class for 5 epochs with 4 batches per epoch of size 120, using a learning rate of 0.001. We use Adam optimizer with batch-norm and dropout.

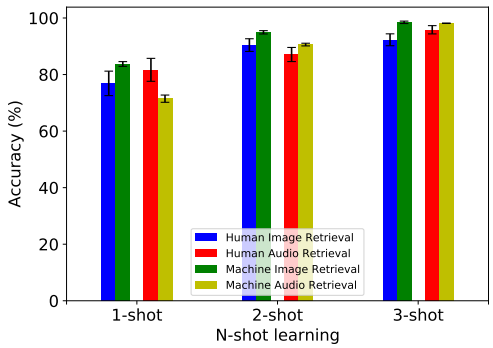

Figure 3: *Comparison of human and machine retrieval accuracy for 1, 2 and 3 shot blocks. Error bars indicate std. error.*

4. Results and Discussions

4.1. Human Experiment Results

4.1.1. Retrieval Scores

Fig. 3 shows the average score for each n-shot learning across all subjects as accuracy (%). For humans, both IR and AR scores in the 1, 2 and 3 shot blocks are in ascending order. The overall AR accuracy is higher than that of IR, with means of 88.194 and 86.528 respectively.

The subject-wise scores for each n-shot learning are shown in Fig. 4. The differences between IR and AR scores are more pronounced here for individual subjects. It is to be noted that the subjects are tested on the same set of objects in both retrieval tasks, with no feedback in between. Since the learnt association does not change after the learning phase, we expect the retrieval scores to be the same. The difference suggests that the modality of the query factors into the strength of the association. Another possibility is the elimination of options for harder test cases, which may vary across subjects depending on their memory preference of one modality over the other.

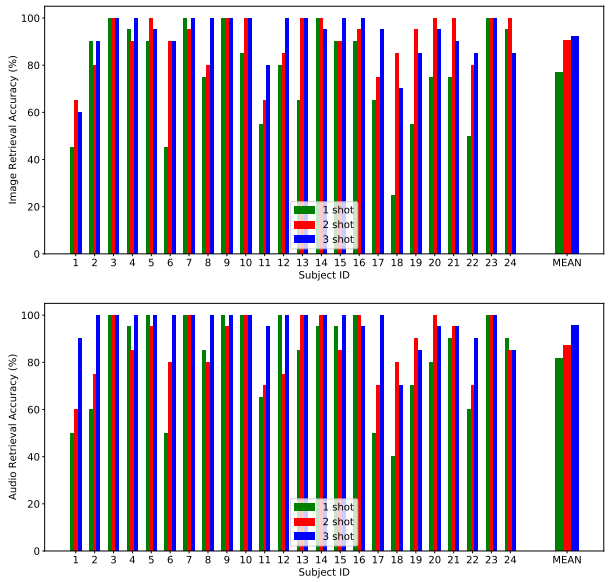

Figure 4: *Subject-wise accuracy for image (top) and audio (bottom) retrieval tasks.*

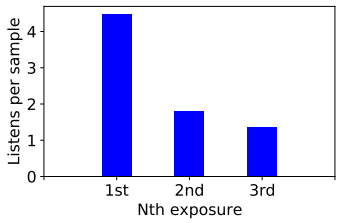

Figure 5: *Average # of listens for the n^{th} exposure of classes.*

4.1.2. Temporal Dynamics of Association Learning

For each exposure, we analyze the number of clicks for a sample termed as the number of listens per sample. The average number of listens per sample for every subject is shown in Fig. 5. From the figure, it is evident that the average number of listens per sample for the 1st exposure is significantly higher than that of the 2nd exposure for all subjects. This difference is observed between the 2nd and 3rd exposures as well. However, the difference in average listens per sample between the 2nd and 3rd exposures are fewer.

We can speculate that a major chunk of the association learning happens during the 1st exposure, which is reinforced in the memory during the 2nd exposure. The average number of listens for the 3rd exposure is 1.375 which suggests that, for most classes, the subjects are merely confirming the association by hearing the audio once.

4.1.3. Analysing Human's Correspondence Learning

Most of the human subjects participated in the experiment were able to learn audio-visual correspondence of unknown image and audio stimuli with more than 90% accuracy with just three time-separated exposures of the stimuli in 2 modalities. The retrieval accuracy is observed to increase significantly with two exposures than after the single exposure. The average accuracy is higher for three-shot learning than the two-shot learning blocks, but the relative change is less. The second exposure itself helped the humans to revise and recollect the correspondence they tried to learn in the first exposure. These results

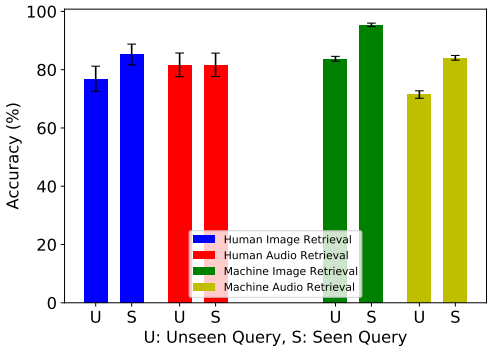

Figure 6: *Generalisation of humans and machines in one-shot learning. Error bars depict the standard errors.*

show that humans learn a major chunk of the association and cues of the multimodal input from a single exposure. The audio retrieval task shows a more gradual improvement in accuracy with an increase in exposures as compared to the image retrieval task. Image retrieval accuracy is more or less similar for two-shot and three-shot learning sessions.

4.2. Machine Results

The random sampling of negative points primarily from the large set of pretraining samples leads to variance in the model. We average the retrieval accuracy trained using 5 random seeds. The mean accuracy with std. error is shown in Fig. 3.

4.2.1. Machine vs Human performance

The model's retrieval accuracy is averaged over each n-shot as shown in Fig. 3. The model retrieval accuracy pattern is comparable to that of humans after n-shot learning. It follows a steady increasing trend in accuracy as n increases, with the IR accuracy consistently better than that of AR. It also outperforms humans in all cases, except for AR in one-shot learning.

4.2.2. One-shot generalisation

The test results for generalisation from a single exposure are shown in Fig. 6. Unseen query stands for the testing case where augmented variants of the query sample, different from those in the training, are used. Seen query stands for the testing case where the query sample is the same sample used in training. The machine model has a significant improvement in performance for IR and AR on the seen queries. For humans, the scores appear more consistent with a minor improvement between the seen and unseen queries. This suggests that humans have a better ability to generalize from a single exposure.

5. Conclusions

In this paper, we examine the audio-visual correspondence learning of novel classes for humans and machines. We note that with an increase in the number of exposures to the stimulus, humans and machines learn the audio-visual correspondence with better accuracy. For humans, the number of times a subject listens to the audio at a particular exposure confirms that the major chunk of association learning happens in the first exposure of the audio-visual pair. Our proposed machine model performs on par with the humans in audio-visual correspondence learning. However, humans have a better ability to generalize to new samples even from a single exposure.

6. References

[1] L. Shams and A. R. Seitz, "Benefits of multisensory learning," *Trends in cognitive sciences*, vol. 12, no. 11, pp. 411–417, 2008.

[2] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[3] B. Baier, A. Kleinschmidt, and N. G. Müller, "Cross-modal processing in early visual and auditory cortices depends on expected statistical relationship of multisensory information," *Journal of Neuroscience*, vol. 26, no. 47, pp. 12 260–12 265, 2006.

[4] J. Ward, B. Huckstep, and E. Tsakanikos, "Sound-colour synaesthesia: To what extent does it use cross-modal mechanisms common to us all?" *Cortex*, vol. 42, no. 2, pp. 264–280, 2006.

[5] M. O. Ernst, "Learning to integrate arbitrary signals from vision and touch," *Journal of Vision*, vol. 7, no. 5, pp. 7–7, 2007.

[6] A. Borovsky, J. L. Elman, and M. Kutas, "Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context," *Language Learning and Development*, vol. 8, no. 3, pp. 278–302, 2012.

[7] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 7–16.

[8] H. Zhang, Y. Zhuang, and F. Wu, "Cross-modal correlation learning for clustering on image-audio dataset," in *Proceedings of the 15th ACM international conference on Multimedia*, 2007, pp. 273–276.

[9] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 613–622.

[10] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2514–2517.

[11] D. Harwath and J. R. Glass, "Learning word-like units from joint audio-visual analysis," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

[12] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017, pp. 349–357.

[13] R. Eloff, H. A. Engelbrecht, and H. Kamper, "Multimodal one-shot learning of speech and images," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8623–8627.

[14] K. Praveen, A. Gupta, A. Soman, and S. Ganapathy, "Second language transfer learning in humans and machines using image supervision," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 1040–1047.

[15] J. S. Horst and M. C. Hout, "The novel object and unusual name (noun) database: A collection of novel images for use in experimental research," *Behavior research methods*, vol. 48, no. 4, pp. 1393–1409, 2016.

[16] "Google cloud text-to-speech," https://cloud.google.com/text-to-speech/, last accessed May 8, 2020.

[17] "IBM Watson Text-to-Speech," https://www.ibm.com/watson/services/text-to-speech/, last accessed May 8, 2020.

[18] "Microsoft Azure Text-to-Speech," https://azure.microsoft.com/en-in/services/cognitive-services/text-to-speech/, last accessed May 8, 2020.

[19] T. J. Barry, J. W. Griffith, S. De Rossi, and D. Hermans, "Meet the fribbles: novel stimuli for use within behavioural research," *Frontiers in Psychology*, vol. 5, p. 103, 2014.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[21] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. Greenberg, D. Reynolds, E. Singer, L. Mason, and J. Hernandez-Cordero, "The 2017 NIST language recognition evaluation," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 82–89.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[23] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 360–368.

[24] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in neural information processing systems*, 2005, pp. 513–520.