Multimodal Transfer Deep Learning with Applications in Audio-Visual Recognition

Seungwhan Moon[*], **Suyoun Kim**[†], **Haohan Wang**[‡]
[*]Language Technologies Institute, [†]Electrical and Computer Engineering, [‡]Machine Learning Department
Carnegie Mellon University
{seungwhm@cs | suyoung1@ece | haohanw@cs}.cmu.edu

Abstract

We propose a transfer deep learning (TDL) framework that can transfer the knowledge obtained from a single-modal neural network to a network with a different modality. Specifically, we show that we can leverage speech data to fine-tune the network trained for video recognition, given an initial set of audio-video parallel dataset within the same semantics. Our approach first learns the analogy-preserving embeddings between the abstract representations learned from intermediate layers of each network, allowing for semantics-level transfer between the source and target modalities. We then apply our neural network operation that fine-tunes the target network with the additional knowledge transferred from the source network, while keeping the topology of the target network unchanged. While we present an audio-visual recognition task as an application of our approach, our framework is flexible and thus can work with any multimodal dataset, or with any already-existing deep networks that share the common underlying semantics. In this work in progress report, we aim to provide comprehensive results of different configurations of the proposed approach on two widely used audio-visual datasets, and we discuss potential applications of the proposed approach.

1 Introduction

Multimodal deep networks have been recently proposed to leverage the features learned from multiple modalities to predict patterns of single or multiple modalities ([1, 2]). While the main focus of this line of work has been to construct a shared representation that best combines multiple modalities, most of the work assume the existence of the parallel multimodal dataset where the shared multimodal representation can be learned.

In reality, however, acquiring a parallel multimodal dataset is extremely resource consuming, and thus there is often an imbalance in the amount of the labeled data among different modalities. For example, while the labeled audio speech data is readily abundant, the labeled data for lip reading videos is much more scarce, consequently making the unparallel portion of the audio data obsolete for multimodal learning. In this paper, we address the data imbalance among different modalities via a transfer learning approach.

The transfer learning relaxes the imbalance of the available data in source tasks and target tasks via selective instance selections or feature transformation ([3, 4]). However, while transfer learning works well when source tasks and target tasks are moderately in the same domain, direct knowledge transfer between different modalities is often intractable often due to their drastically different statistical properties in concept space ([5]). In this regard, several approaches such as DCCA ([6]), DeViSE ([7]), HHTL ([8]), and DCCAE ([9]) have proposed to utilize deep neural networks to obtain abstract representation of data (*e.g.* at the top-most layer of the network), thus allowing for more viable knowledge transfer. Most of these approaches aim at learning a single robust mapping

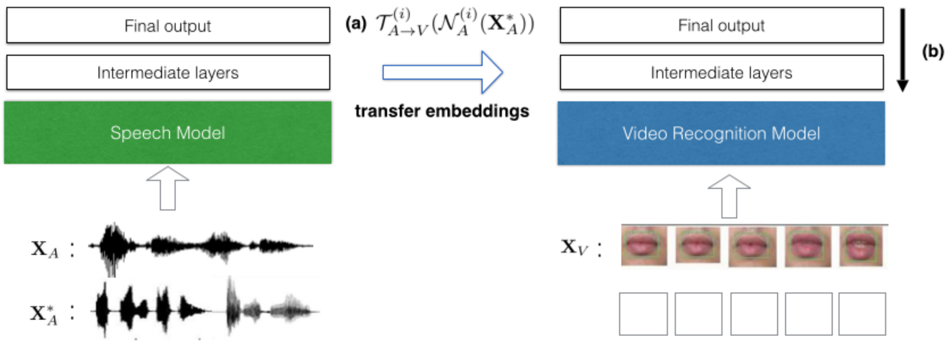

Figure 1: An illustration of our approach with an application in audio and lip-reading video recognition tasks. Notations are defined in Section 2. (a) We learn the embeddings between abstract representations of audio and video data using the parallel dataset ($\mathbf{X}_A$ and $\mathbf{X}_V$), from which we can transfer the new audio input $\mathbf{X}^*_A$, unforeseen in the label space of $\mathbf{X}_V$. (b) We can then fine-tune the video network $\mathcal{N}_V$ with the transferred audio data $\mathcal{T}_{A\to V}(\mathcal{N}^{(i)}_A(\mathbf{X}^*_A))$.

between the two modalities, and then use the newly learned embeddings to learn a new model for a shared task or a target task.

In this paper, we propose a new approach that can benefit the target task in the case of source-target data imbalance scenario. Specifically, we learn the semantic mappings between intermediate layers of each neural network, and leverage learned embeddings to fine-tune the target network. Our approach has several practical benefits in that we keep the topology of the target network unchanged, thus not requiring to build a separate model with a shared representation, etc. Instead, we directly modify the target network by tuning its hyper-parameters, and thus the same network can be used for the target task.

As an application of our framework, we choose to study the transfer between speech and lip-reading video data. However, our framework is unobtrusive and flexible, and thus can be used for transfer between any modalities with any two already-built deep networks, given the small initial parallel modality corpora that correspond to the same semantics.

2 Problem Formulation

We formulate the proposed framework with application in the audio and video lip-reading multi-modal learning settings. Figure 1 illustrates the following formulation. We initially have two input data of different modalities: $\mathbf{X}_A = \{X^1_A, \cdots, X^N_A\}$ and $\mathbf{X}_V = \{X^1_V, \cdots, X^N_V\}$, which correspond to the parallel audio and video data, respectively. Both $\mathbf{X}_A$ and $\mathbf{X}_V$ map to the same ground-truth categorical labels $\mathbf{Z} = \{Z^1, \cdots, Z^N\}$. Each input audio instance X^n_A and video instance X^n_V lies in different concept spaces, thus $X^n_A \in \mathbb{R}^p$ and $X^n_V \in \mathbb{R}^q$. Two separate neural nets, $\mathcal{N}_A$ and $\mathcal{N}_V$, can be built from $\mathbf{X}_A$ and $\mathbf{X}_V$, respectively. We denote $\mathcal{N}_A : X_A \to Y$ and $\mathcal{N}_V : X_V \to Y$ where Y is a predicted label as a final output of the neural networks, given some $X_A \in \mathbf{X}_A$ or $X_V \in \mathbf{X}_V$.

Additionally, we also denote $\mathcal{N}^{(i)}_A : X_A \to H^{(i)}_A$ and $\mathcal{N}^{(i)}_V : X_V \to H^{(i)}_V$, where $H^{(i)}_A \in \mathbb{R}^{p_i}$ and $H^{(i)}_V \in \mathbb{R}^{q_i}$ are the output of the i-th layer of the two neural nets, respectively, $\forall i \in \{0, 1, \cdots, l{+}1\}$. Note that $H^{(0)} = X$, $H^{(l+1)} = Y$, and $H^{(i+1)} := g(H^{(i)}, \mathcal{W}^{(i\to i+1)})$, where $\mathcal{W}^{(i\to i+1)}$ is the weight between $H^{(i)}$ and $H^{(i+1)}$ for $i \in \{0, \cdots, l\}$. H_A and H_V thus represent the input X_A and X_V at different abstraction levels.

We then define a new set of labeled audio data, ($\mathbf{X}^*_A = \{X^{*1}_A, \cdots, X^{*M}_A\}, \mathbf{Z}^* = \{Z^{*1}, \cdots, Z^{*M}\}$), which is unparallel and unforeseen in the video dataset in label space ($\mathbf{X}^*_V$ is assumed to be not available at training phase). Our purpose is then to learn a transfer function $\mathcal{T}^{(i)}_{A\to V} : H^{(i)}_A \to H^{(i)}_V \ \forall i \in \{0, 1, \cdots, m\}$ using $\mathbf{X}_A$ and $\mathbf{X}_V$, from which we can fine-tune $\mathcal{N}_V$ with $\mathcal{T}_{A\to V}(H^*_A)$, where $H^*_A = \mathcal{N}_A(X^*_A)$ for $X^*_A \in \mathbf{X}^*_A$ (Section 3).

3 Method

We obtain abstract representations of the raw data using a standard deep belief network (DBN) with multiple RBM layers. [10] Output values at each intermediate layer of a DBN thus give abstract feature representation of the input data, allowing for a more tractable knowledge transfer between modalities. In our experiment, we build two DBNs ($\mathcal{N}_A$ and $\mathcal{N}_V$) for audio and video data that have the same number of intermediate layers, and we learn inter-modal embeddings for each layer at the same depth as detailed in Section 3.1. Using the learned mapping between the source and target modalities, we fine-tune the network with the transferred data as detailed in Section 3.2.

3.1 Learning the Embeddings

We learn the embedding function $\mathcal{T}_{A\rightarrow V}$ which maps two concept spaces $H_A^{(i)} \in \mathbb{R}^{p_i}$ and $H_V^{(i)} \in \mathbb{R}^{q_i}$. While any embedding method can be applied, we consider the following three embedding methods from literature.

Multivariate Support Vector Regression (SVR) Using Nonlinear Kernels: we formulate the mapping between two concept spaces as a multivariate regression problem. Specifically, we use Support Vector Regression (SVR) methods with nonlinear kernels ([11]), which can effectively learn the conditional expectation of the target space given the source space. For practical use, we kernelize the weights and add a soft-margin loss function for more flexible and high-dimensional mapping.

KNN-based Non-parametric Mapping: we can also obtain mappings of new audio input by first finding the K-closest audio samples in the training set, and then returning the average of the values of the corresponding video samples.

Normalized Canonical Correlation Analysis (NCCA): given a set of N audio and video pairs $\mathbf{H}_A \in \mathbb{R}^{N\times p'}$ and $\mathbf{H}_V \in \mathbb{R}^{N\times q'}$, NCCA obtains $\mathbf{U} \in \mathbb{R}^{p'\times c}$ and $\mathbf{V} \in \mathbb{R}^{q'\times c}$ which project audio and video into a common c-dimensional latent space by $\mathbf{H}_A\mathbf{U}$ and $\mathbf{H}_V\mathbf{V}$ ([12]). The objective can thus be formulated as:

$$\begin{aligned} &\max_{\mathbf{U},\mathbf{V}} \text{tr}(\mathbf{U}^T\mathbf{H}_A^T\mathbf{H}_V\mathbf{V}) \\ &\text{s.t. } \mathbf{U}^T\mathbf{H}_A^T\mathbf{H}_A\mathbf{U}=\mathbf{I}, \mathbf{V}^T\mathbf{H}_V^T\mathbf{H}_V\mathbf{V}=\mathbf{I}. \end{aligned} \tag{1}$$

which we can solve as a generalized eigenvalue problem. Once $\mathbf{U}$ and $\mathbf{V}$ are obtained, we can formulate the mapping function as $\mathcal{T}_{A\rightarrow V}(\mathbf{H}_A) = \mathbf{H}_A\mathbf{U}\mathbf{V}^T$ to obtain the estimated transfer $\mathbf{H}_V$.

3.2 Fine-tuning with the Transferred Data

We propose an algorithm to fine-tune the network with the transferred data (`TDLFT`) as described in Algorithm 1. `TDLFT(i)` obtains a transfer of the new audio input at i-th layer, $H_V^{(i)} := \mathcal{T}_{A\rightarrow V}^{(i)}(\mathcal{N}_A^{(i)}(X_A^*))$, treats it as an input to $\mathcal{N}_V$ at i-th layer, and computes $H_V^{(j+1)} := g(H_V^{(j)}, \mathcal{W}_V^{(j\rightarrow j+1)})$ for $j \in \{i, i+1, \cdots, l\}$, where g is an activation function. Finally, we fine-tune $\mathcal{W}_V^{(j\rightarrow j+1)}$ for $j \in \{i, i+1, \cdots, l\}$ via a standard backpropagation algorithm.

Note that `TDLFT(i)` does not fine-tune $\mathcal{W}_V^{(j\rightarrow j+1)}$ for $j \in \{0, 1, \cdots, i-1\}$. Therefore, only the top $(l-i+1)$ layers are fine-tuned to the new stimuli, which could be a drawback of this algorithm. However, we argue that this issue can be naturally mitigated through some good choice of i, at which a typical error backpropagation technique reaches the inevitable *vanishing errors* at bottom layers [13]. Assuming a perfect transfer function learned at this choice of i, we almost have the same effect as if we had a parallel video data X_V^* available for training.

It is interesting to note the expected trade-off behavior for different choices of i. When i is big ($i \simeq l$), we expect that the accuracy of transfer is more reliable (e.g. intuitively, $\mathcal{T}_{A\rightarrow V}^{(l+1)}$ should always be correct, because it is a mapping from and to the same label space), however we only get to fine-tune the top $(l-i+1)$ layer(s), leaving the rest of the bottom layers un-tuned. A smaller choice of i mitigates this issue, however we typically suffer from an unreliable transfer for $i \simeq 0$ (e.g. $\mathcal{T}_{A\rightarrow V}^{(0)}$ is almost intractable, given drastically different concept spaces of the two input types).

In section 4, we show our results with different combinations of embedding methods and application of the `TDLFT` at different depths of the network.

Algorithm 1 Transfer Deep Learning Fine-Tuning (TDLFT)

```
**Input**: $\mathcal{N}_A$ trained with $\mathbf{X}_A, \mathcal{N}_V$ trained with $\mathbf{X}_V$, an input parameter $i \in \{0, 1, \cdots, l\}$, $\mathcal{T}^{(j)}_{A \to V} : \mathbf{H}^{(j)}_A \to \mathbf{H}^{(j)}_V$ learned for $j := i$, a new unparallel data ($\mathbf{X}^*_A$, $\mathbf{Y}^*$).
**Output:** $\mathcal{N}_V$ fine-tuned with $\mathbf{X}^*_A$.
Obtain $\mathbf{H}^{(i)}_V := \mathcal{T}^{(i)}_{A \to V}(\mathcal{N}^{(i)}_A(X^*_A))$, and $\mathbf{H}^{(l+1)}_V := \mathbf{Y}^*$
**for** $j \in \{i, i+1, \cdots, l\}$ **do**
  $\mathbf{H}^{(j+1)}_V := g(\mathbf{H}^{(j)}_V, \mathcal{W}^{(j \to j+1)}_V)$
**end for**
Fine-tune $\mathcal{W}^{(j \to j+1)}_V$ for $j \in \{i, i+1, \cdots, l\}$ via a standard backpropagation algorithm.
```

Table 1: Overview of Datasets.

Dataset	Division	Labels	# Attribtues	# Instances
AV-Letters	$\mathbf{X}_A$	1-20	624	600
	$\mathbf{X}^*_A$	21-26		180
	$\mathbf{X}_V$	1-20	57,600	600
	$\mathbf{X}^*_V$	21-26		180
Stanford	$\mathbf{X}_A$	1-44	1,573	2,064
	$\mathbf{X}^*_A$	45-49		504
	$\mathbf{X}_V$	1-44	19,481	2,064
	$\mathbf{X}^*_V$	45-49		504

4 Empirical Evaluation

4.1 Dataset

We use two widely used audio-visual multimodal datasets for empirical evaluation, AV-Letters and Stanford. The AV-Letters dataset ([14]) consists of both audio and lip-reading video data of 10 speakers saying the letters A to Z (thus 26 labels), three times each. The dataset contains pre-extracted lip regions at 60 × 80 pixels. We represent each audio example as one feature vector by concatenating 24 contiguous audio frames, each with 26 mel-frequency cepstral coefficients (MFCC). Similarly, for each video example, we concatenate 12 contiguous video frames, each with 60×80 pixels. The Stanford dataset ([1]) consists of both audio and lip-reading video data of 23 speakers saying the digits 0 to 9, letters A to Z, and some selected sentences (total 49 labels).

We divide each dataset by its label space to simulate the data imbalance scenario. For instance, in the AV-Letters dataset, we assume that only the data with labels from 1 to 20 (denoted $\mathbf{X}_V$) are available as a parallel corpus during training of video data, and the video data with labels ranging from 21 to 26 (denoted $\mathbf{X}^*_V$) are assumed to be completely unforeseen during the training phase. On the other hand, we assume that the audio data has access to the entire label space (labels 1 to 26) during training ($\mathbf{X}_A \cup \mathbf{X}^*_A$). This is analogous to many real-world situations where the training data for multi-modal learning is imbalanced for one of the modalities, thus the test data of the target modality is radically different from the training data. Table 1 summarizes the datasets configuration.

4.2 Task

To evaluate how the transferred knowledge from the source modality improves classification performance of the target modality, we perform comparison studies on the following task with the AV-Letters and Stanford dataset (in Section 4.1). The task is to build a classifier that categorizes each lip-reading video into a label (1-26 for AV-Letters, and 1-49 for Stanford). We assume that only $\mathbf{X}_V$ and $\mathbf{X}_A$ are available as a parallel corpus during initial training.

Assuming that the audio data has extra labeled data ($\mathbf{X}^*_A$) unparallel in the video training set, we fine-tune the video model ($\mathcal{N}_V$) with the extra transferred audio data as described in Section 3.2. The classification task is performed in a cross validation way on the data spanning the entire labels ($\mathbf{X}_V \cup \mathbf{X}^*_V$). We then compare the result of our transfer deep learning approach (TDL) against the unimodal baseline where the model is trained only with the available video data ($\mathbf{X}_V$).

Table 2: 5-fold lip-reading video classification accuracy on the AV-Letters dataset (26 labels)

	Unimodal	**TDL**	Oracle
Train:	$\mathbf{X}_V$	$\mathbf{X}_V \cup \underline{\mathcal{T}_{A\to V}(\mathbf{H}^*_A)}$	$\mathbf{X}_V \cup \underline{\mathbf{X}^*_V}$
Test:	$\mathbf{X}_V \cup \mathbf{X}^*_V$	$\mathbf{X}_V \cup \mathbf{X}^*_V$	$\mathbf{X}_V \cup \mathbf{X}^*_V$
KNN + TDLFT(3)	51.1%	**55.3**%	61.7%
NCCA + TDLFT(3)		**53.1**%	
SVR + TDLFT(3)		**54.8**%	
KNN + TDLFT(0)	51.1%	34.4%	66.8%
NCCA + TDLFT(0)		32.1%	
SVR + TDLFT(0)		48.3%	

Table 3: 5-fold lip-reading video classification accuracy on the Stanford dataset (49 labels)

	Unimodal	**TDL**	Oracle
Train:	$\mathbf{X}_V$	$\mathbf{X}_V \cup \underline{\mathcal{T}_{A\to V}(\mathbf{H}^*_A)}$	$\mathbf{X}_V \cup \underline{\mathbf{X}^*_V}$
Test:	$\mathbf{X}_V \cup \mathbf{X}^*_V$	$\mathbf{X}_V \cup \mathbf{X}^*_V$	$\mathbf{X}_V \cup \mathbf{X}^*_V$
KNN + TDLFT(3)	54.9%	**61.3**%	68.2%
NCCA + TDLFT(3)		**58.2**%	
SVR + TDLFT(3)		**56.8**%	
KNN + TDLFT(0)	54.9%	49.8%	73.2%
NCCA + TDLFT(0)		52.4%	
SVR + TDLFT(0)		45.3%	

4.3 Results

In this section, we provide comprehensive empirical results on the task described in Section 4.2 with different combinations of embedding transfer methods and fine-tuning methods (in Section 3).

Tables 2 and 3 shows the comparison of 5-fold cross validation performance between the unimodal (video) model and the transfer deep learning model (TDL) where the network is additionally fine-tuned with the transferred audio data. KNN, NCCA, and SVR refer to the embedding methods that were applied to transfer the audio representations, as detailed in Section 3.1. TDLFT(i) refers to the fine-tuning method applied starting at the i-th layer of the network (Section 3.2). The Oracle baseline shows the semi-oracle bound of TDL which can be achieved if the perfect knowledge transfer $\mathcal{T}_{A\hat{\to}V}$ is obtainable. The underlined portion on the train sets denote the data that was used to further fine-tune the network via TDLFT. Bold denotes significant improvement of TDL over the baseline. The chance performance for AV-Letters and Stanford is 3.85% (=1/26) and 2.04% (=1/49), respectively.

TDL with TDLFT(3) outperforms unimodal for both datasets when audio data (source domain) was transferred with any of the proposed embedding methods, showing that the transferred modality enhances the performance of the target modality classification. This is promising because it indicates that we can improve the classification performance in the previously unknown classes of the target task without training samples for those classes. Specifically, our results show that the KNN-based transfer method (KNN) generally yields the best embeddings, followed by NCCA and SVR.

Note that the TDL performance at TDLFT(0) (where the transfer between modalities was at the raw feature representation level) does not show a significant improvement over the Unimodal baseline. This is because learning embeddings at the input level between the target and the source modalities is often intractable, leading to a poor transfer performance. The network is then fine-tuned with noised transferred data, thus negatively affecting the classification performance. Note that this approach is similar to the traditional feature-transformation-based transfer learning methods.

All of the TDL performances are upper-bounded by the Oracle performance which simulates a perfect transfer between modalities. This result indicates that we can improve the transfer deep learning performance with a better transfer embedding method between modalities. Note also that the Oracle performance with TDLFT(0) has the best overall performance. While the transfer of embeddings at the raw input level is intractable in reality, TDLFT(0) fully fine-tunes every layer in the network, which would thus improve the performance the most if embeddings are reliable.

5 Conclusions

We proposed a framework for performing transfer learning on neural networks (TDL) under multimodal learning settings. We proposed several embedding methods for transferring knowledge between the target and source modalities, and presented our results on two audio-visual datasets. Our results show that the transferred modality of an abstract representation obtained from intermediate layers of the source network can be effectively utilized to further fine-tune the target network. Specifically, our results indicate that our approach is especially applicable when the data in the target modality is much more scarce (*i.e.* in label space) than in the source modality.

Future Work: we note that the proposed framework can be extended to the idea of reconstructing the modality given the transferred modality via top-down inference [15]. This has many potential applications, for instance in automatically generating lip-motion videos given any audio input using the transferred audio input, which is an improvement over the conventional lip-motion generator that is heavily rule-based or human-engineered. We plan on addressing this problem in the future work.

References

[1] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," *International Conference on Machine Learning (ICML)*, 2011.

[2] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, pp. 2222–2230, 2012.

[3] M. E. Taylor and P. Stone, "Cross-domain transfer for reinforcement learning," in *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, (New York, NY, USA), pp. 879–886, ACM, 2007.

[4] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pp. 677–682, AAAI Press, 2008.

[5] C. Navarretta, "Transfer learning in multimodal corpora," in *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*, pp. 195–200, Dec 2013.

[6] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th International Conference on Machine Learning*, pp. 1247–1255, 2013.

[7] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, *et al.*, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems*, pp. 2121–2129, 2013.

[8] J. T. Zhou, S. J. Pan, I. W. Tsang, and Y. Yan, "Hybrid heterogeneous transfer learning through deep learning," 2014.

[9] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning multi-view representation for face recognition," *CoRR*, vol. abs/1406.6947, 2014.

[10] H. Wang and B. Raj, "A survey: Time travel in deep learning space: An introduction to deep learning models and how deep learning models evolved from the initial ideas," *arXiv preprint arXiv:1510.04781*, 2015.

[11] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, Aug. 2004.

[12] G. Kim, S. Moon, and L. Sigal, "Ranking and retrieval of image sequences from multiple paragraph queries," in *CVPR*, 2015.

[13] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[14] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 198–213, Feb. 2002.

[15] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, (New York, NY, USA), pp. 609–616, ACM, 2009.