Multimodal Deep Learning

Jiquan Ngiam[1] JNGIAM@CS.STANFORD.EDU
Aditya Khosla[1] ADITYA86@CS.STANFORD.EDU
Mingyu Kim[1] MINKYU89@CS.STANFORD.EDU
Juhan Nam[1] JUHAN@CCRMA.STANFORD.EDU
Honglak Lee[2] HONGLAK@EECS.UMICH.EDU
Andrew Y. Ng[1] ANG@CS.STANFORD.EDU

[1] Computer Science Department, Stanford University, Stanford, CA 94305, USA
[2] Computer Science and Engineering Division, University of Michigan, Ann Arbor, MI 48109, USA

Abstract

Deep networks have been successfully applied to unsupervised feature learning for single modalities (e.g., text, images or audio). In this work, we propose a novel application of deep networks to learn features over multiple modalities. We present a series of tasks for multimodal learning and show how to train deep networks that learn features to address these tasks. In particular, we demonstrate cross modality feature learning, where better features for one modality (e.g., video) can be learned if multiple modalities (e.g., audio and video) are present at feature learning time. Furthermore, we show how to learn a shared representation between modalities and evaluate it on a unique task, where the classifier is trained with audio-only data but tested with video-only data and vice-versa. Our models are validated on the CUAVE and AVLetters datasets on audio-visual speech classification, demonstrating best published visual speech classification on AVLetters and effective shared representation learning.

1. Introduction

In speech recognition, humans are known to integrate audio-visual information in order to understand speech. This was first exemplified in the McGurk effect (McGurk & MacDonald, 1976) where a visual */ga/* with a voiced */ba/* is perceived as */da/* by most subjects. In particular, the visual modality provides information on the place of articulation and muscle movements (Summerfield, 1992) which can often help to disambiguate between speech with similar acoustics (e.g., the unvoiced consonants */p/* and */k/*).

Multimodal learning involves relating information from multiple sources. For example, images and 3-d depth scans are correlated at first-order as depth discontinuities often manifest as strong edges in images. Conversely, audio and visual data for speech recognition have correlations at a "mid-level", as phonemes and visemes (lip pose and motions); it can be difficult to relate raw pixels to audio waveforms or spectrograms.

In this paper, we are interested in modeling "mid-level" relationships, thus we choose to use audio-visual speech classification to validate our methods. In particular, we focus on learning representations for speech audio which are coupled with videos of the lips.

We will consider the learning settings shown in Figure 1. The overall task can be divided into three phases – feature learning, supervised training, and testing. A simple linear classifier is used for supervised training and testing to examine different feature learning models with multimodal data. In particular, we consider three learning settings – multimodal fusion, cross modality learning, and shared representation learning.

In the multimodal fusion setting, data from all modalities is available at all phases; this represents the typical setting considered in most prior work in audio-visual speech recognition (Potamianos et al., 2004). In cross modality learning, data from multiple modalities is available only during feature learning; during the supervised training and testing phase, only data from a single modality is provided. For this setting, the aim is to learn better single modality representations given unlabeled data from multiple modalities. Last, we con-

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

sider a shared representation learning setting, which is unique in that different modalities are presented for supervised training and testing. This setting allows us to evaluate if the feature representations can capture correlations across different modalities. Specifically, studying this setting allows us to assess whether the learned representations are modality-invariant.

In the following sections, we first describe the building blocks of our model. We then present different multimodal learning models leading to a deep network that is able to perform the various multimodal learning tasks. Finally, we report experimental results and conclude.

	Feature Learning	Supervised Training	Testing
Classic Deep Learning	Audio	Audio	Audio
	Video	Video	Video
Multimodal Fusion	A + V	A + V	A + V
Cross Modality Learning	A + V	Video	Video
	A + V	Audio	Audio
Shared Representation Learning	A + V	Audio	Video
	A + V	Video	Audio

Figure 1: Multimodal Learning settings where A+V refers to Audio and Video.

2. Background

Recent work on deep learning (Hinton & Salakhutdinov, 2006; Salakhutdinov & Hinton, 2009) has examined how deep sigmoidal networks can be trained to produce useful representations for handwritten digits and text. The key idea is to use greedy layer-wise training with Restricted Boltzmann Machines (RBMs) followed by fine-tuning. We use an extension of RBMs with sparsity (Lee et al., 2007), which have been shown to learn meaningful features for digits and natural images. In the next section, we review the sparse RBM, which is used as a layer-wise building block for our models.

2.1. Sparse restricted Boltzmann machines

The RBM is an undirected graphical model with hidden variables ($\mathbf{h}$) and visible variables ($\mathbf{v}$) (Figure 2a). There are symmetric connections between the hidden and visible variables ($W_{i,j}$), but no connections within hidden variables or visible variables. The model defines a probability distribution over $\mathbf{h}$, $\mathbf{v}$ (Equation 1). This particular configuration makes it easy to compute the conditional probability distributions, when $\mathbf{v}$ or $\mathbf{h}$ is fixed (Equation 2).

$$-\log P(\mathbf{v},\mathbf{h}) \propto E(\mathbf{v},\mathbf{h}) = \frac{1}{2\sigma^2}\mathbf{v}^T\mathbf{v} - \frac{1}{\sigma^2}\left(\mathbf{c}^T\mathbf{v} + \mathbf{b}^T\mathbf{h} + \mathbf{h}^T W\mathbf{v}\right) \quad (1)$$

$$p(h_j|\mathbf{v}) = sigmoid(\frac{1}{\sigma^2}(b_j + \mathbf{w}_j^T\mathbf{v})) \quad (2)$$

This formulation models the visible variables as real-valued units and the hidden variables as binary units.[1] As it is intractable to compute the gradient of the log-likelihood term, we learn the parameters of the model $(w_{i,j}, b_j, c_i)$ using contrastive divergence (Hinton, 2002).

To regularize the model for sparsity (Lee et al., 2007), we encourage each hidden unit to have a predetermined expected activation using a regularization penalty of the form $\lambda\sum_j(\rho - \frac{1}{m}(\sum_{k=1}^m \mathbf{E}[h_j|\mathbf{v}^k]))^2$, where $\{\mathbf{v}^1, ..., \mathbf{v}^m\}$ is the training set and ρ determines the sparsity of the hidden unit activations.

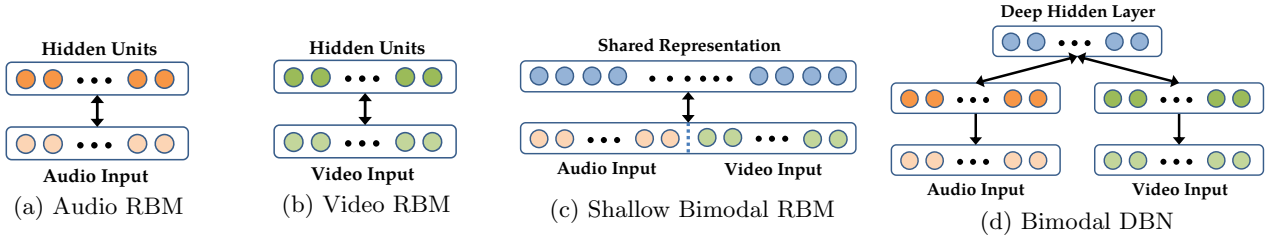
3. Learning architectures

In this section, we describe our models for the task of audio-visual bimodal feature learning, where the audio and visual input to the model are contiguous audio (spectrogram) and video frames. To motivate our deep autoencoder (Hinton & Salakhutdinov, 2006) model, we first describe several simple models and their drawbacks.

One of the most straightforward approaches to feature learning is to train a RBM model *separately* for audio and video (Figure 2a,b). After learning the RBM, the posteriors of the hidden variables given the visible variables (Equation 2) can then be used as a new representation for the data. We use this model as a baseline to compare the results of our multimodal models, as well as for pre-training the deep networks.

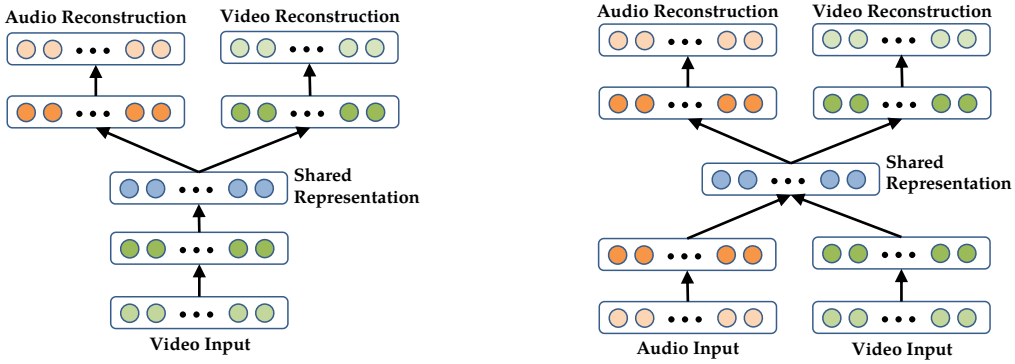
To train a multimodal model, a direct approach is to train a RBM over the concatenated audio and video data (Figure 2c). While this approach jointly models the distribution of the audio and video data, it is limited as a shallow model. In particular, since the correlations between the audio and video data are highly non-linear, it is hard for a RBM to learn these correlations and form multimodal representations. In practice, we found that learning a shallow bimodal RBM results in hidden units that have strong connections to variables from individual modality but few units that connect across the modalities.

[1] We use Gaussian visible units for the RBM that is connected to the input data. When training the deeper layers, we use binary visible units.



(a) Audio RBM (b) Video RBM (c) Shallow Bimodal RBM (d) Bimodal DBN

Figure 2: RBM Pretraining Models. We train RBMs for (a) audio and (b) video separately as a baseline. The shallow model (c) is limited and we find that this model is unable to capture correlations across the modalities. The bimodal deep belief network (DBN) model (d) is trained in a greedy layer-wise fashion by first training models (a) & (b). We later "unroll" the deep model (d) to train the deep autoencoder models presented in Figure 3.



(a) Video-Only Deep Autoencoder (b) Bimodal Deep Autoencoder

Figure 3: Deep Autoencoder Models. A "video-only" model is shown in (a) where the model learns to reconstruct both modalities given only video as the input. A similar model can be drawn for the "audio-only" setting. We train the (b) bimodal deep autoencoder in a denoising fashion, using an augmented dataset with examples that require the network to reconstruct both modalities given only one. Both models are pre-trained using sparse RBMs (Figure 2d). Since we use a sigmoid transfer function in the deep network, we can initialize the network using the conditional probability distributions $p(\mathbf{h}|\mathbf{v})$ and $p(\mathbf{v}|\mathbf{h})$ of the learned RBM.

Therefore, we consider greedily training a RBM over the pre-trained layers for each modality, as motivated by deep learning methods (Figure 2d).[2] In particular, the posteriors (Equation 2) of the first layer hidden variables are used as the training data for the new layer. By representing the data through learned first layer representations, it can be easier for the model to learn higher-order correlations across modalities. Informally, the first layer representations correspond to phonemes and visemes and the second layer models the relationships between them. Figure 4 shows visualizations of learned features from our models including examples of visual bases corresponding to visemes.

However, there are still two issues with the above multimodal models. First, there is no explicit objective for the models to discover correlations across the modalities; it is possible for the model to find representations such that some hidden units are tuned only for audio while others are tuned only for video. Second, the models are clumsy to use in a cross modality learning setting where only one modality is present during supervised training and testing. With only a single modality present, one would need to integrate out the unobserved visible variables to perform inference.

Thus, we propose a deep autoencoder that resolves both issues. We first consider the cross modality learning setting where both modalities are present during feature learning but only a single modality is used for supervised training and testing. The deep autoencoder (Figure 3a) is trained to reconstruct both modalities when given only video data and thus discovers correlations across the modalities. Analogous to Hinton & Salakhutdinov (2006), we initialize the deep autoencoder with the bimodal DBN weights (Figure 2d) based on Equation 2, discarding any weights that are no longer present. The middle layer can be used as the

[2] It is possible to instead learn a large RBM as the first layer that connects to both modalities. However, since a single layer RBM tends to learn unimodal units, it is much more efficient to learn separate models for each modality.

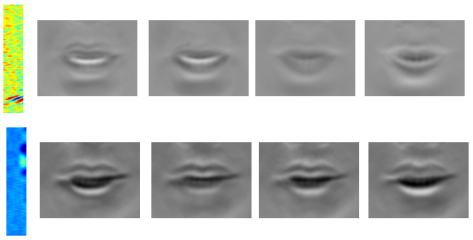


Figure 4: Visualization of learned representations. These figures correspond to two deep hidden units, where we visualize the most strongly connected first layer features. The units are presented in audio-visual pairs (we have found it generally difficult to interpret the connection between the pair). The visual bases captured lip motions and articulations, including different mouth articulations, opening and closing of the mouth, exposing teeth.

new feature representation. This model can be viewed as an instance of multitask learning (Caruana, 1997).

We use the deep autoencoder (Figure 3a) models in settings where only a single modality is present at supervised training and testing. On the other hand, when multiple modalities are available for the task (e.g., multimodal fusion), it is less clear how to use the model as one would need to train a deep autoencoder for each modality. One straightforward solution is to train the networks such that the decoding weights are tied. However, such an approach does not scale well – if we were to allow any combination of modalities to be present or absent at test time, we will need to train an exponential number of models.

Inspired by denoising autoencoders (Vincent et al., 2008), we propose training the bimodal deep autoencoder (Figure 3b) using an augmented but noisy dataset with additional examples that have only a single-modality as input. In practice, we add examples that have zero values for one of the input modalities (e.g., video) and original values for the other input modality (e.g., audio), but still require the network to reconstruct both modalities (audio and video). Thus, one-third of the training data has only video for input, while another one-third of the data has only audio, and the last one-third of the data has both audio and video.

Due to initialization using sparse RBMs, we find that the hidden units have low expected activation, even *after* the deep autoencoder training. Therefore, when one of the input modalities is set to zero, the first layer representations are also close to zero. In this case, we are essentially training a modality-specific deep autoencoder network (Figure 3a). Effectively, the method learns a model which is robust to inputs where a modality is absent.

4. Experiments and Results

We evaluate our methods on audio-visual speech classification of isolated letters and digits. The sparseness parameter ρ was chosen using cross-validation, while all other parameters (including hidden layer size and weight regularization) were kept fixed.[3]

4.1. Data Preprocessing

We represent the audio signal using its spectrogram[4] with temporal derivatives, resulting in a 483 dimension vector which was reduced to 100 dimensions with PCA whitening. 10 contiguous audio frames were used as the input to our models.

For the video, we preprocessed the frames so as to extract only the region-of-interest (ROI) encompassing the mouth.[5] Each mouth ROI was rescaled to 60×80 pixels and further reduced to 32 dimensions,[6] using PCA whitening. Temporal derivatives over the reduced vector were also used. We used 4 contiguous video frames for input since this had approximately the same duration as 10 audio frames.

For both modalities, we also performed feature mean normalization over time (Potamianos et al., 2004), akin to removing the DC component from each example. We also note that adding temporal derivatives to the representations has been widely used in the literature as it helps to model dynamic speech information (Potamianos et al., 2004; Zhao & Barnard, 2009). The temporal derivatives were computed using a normalized linear slope so that the dynamic range of the derivative features is comparable to the original signal.

4.2. Datasets and Task

Since only unlabeled data was required for unsupervised feature learning, we combined diverse datasets (as listed below) to learn features. AVLetters and CUAVE were further used for supervised classification. We ensured that no test data was used for unsupervised feature learning. All deep autoencoder models were trained with all available unlabeled audio and video data.

[3] We cross-validated ρ over $\{0.01, 0.03, 0.05, 0.07\}$. The first layer features were 4x overcomplete for video (1536 units) and 1.5x overcomplete for audio (1500 units). The second layer was 1.5x the size of the combined first layers (4554 units).

[4] Each spectrogram frame (161 frequency bins) had a 20ms window with 10ms overlaps.

[5] We used an off-the-shelf object detector (Dalal & Triggs, 2005) with median filtering over time to extract the mouth regions.

[6] Similar to (Duchnowski et al., 1994) we found that 32 dimensions were sufficient and performed well.

CUAVE (Patterson et al., 2002). 36 speakers saying the digits *0* to *9*. We used the *normal* portion of the dataset which contained frontal facing speakers saying each digit 5 times. We evaluated digit classification on the CUAVE dataset in a speaker independent setting. As there has not been a fixed protocol for evaluation on this dataset, we chose to use odd-numbered speakers for the test set and even-numbered speakers for the training set.

AVLetters (Matthews et al., 2002). 10 speakers saying the letters A to Z, three times each. The dataset provided pre-extracted lip regions of 60×80 pixels. As the raw audio was not available for this dataset, we used it for evaluation on a visual-only lipreading task (Section 4.3). We report results on the *third-test* settings used by Zhao & Barnard (2009) and Matthews et al. (2002) for comparisons.

AVLetters2 (Cox et al., 2008). 5 speakers saying the letters A to Z, seven times each. This is a new high-definition version of the AVLetters dataset. We used this dataset for unsupervised training only.

Stanford Dataset. 23 volunteers spoke the digits *0* to *9*, letters A to Z and selected sentences from the TIMIT dataset. We collected this data in a similar fashion to the CUAVE dataset and used it for unsupervised training only.

TIMIT (Fisher et al., 1986). We used this dataset for unsupervised audio feature pre-training.

We note that in all datasets there is variability in the lips in terms of appearance, orientation and size. For each audio-video clip, features were extracted from overlapping sequences of frames. Since examples had varying durations, we divided each example into S equal slices and performed average-pooling over each slice. The features from all slices were subsequently concatenated together. Specifically, we combined features using $S = 1$ and $S = 3$ to form our final feature representation for classification with a linear SVM.

4.3. Cross Modality Learning

In the cross modality learning experiments, we evaluate if we can learn better representations for one modality (e.g., video) when given multiple modalities (e.g., audio and video) during feature learning.

On the AVLetters dataset (Table 1a), our deep autoencoder models show a significant improvement over hand-engineered features from prior work. The video-only deep autoencoder performed the best on the dataset, obtaining a classification accuracy of 64.4%, outperforming the best previous published results.

On the CUAVE dataset (Table 1b), there is an improvement by learning video features with both video and audio compared to learning features with only video data (although not performing as well as state-of-the-art). In our models, we chose to use a very simple front-end that only extracts bounding boxes, without any correction for orientation or perspective changes. In contrast, recent AAM models (Papandreou et al., 2009) are trained to accurately track the speaker's face and further register the face with a mean face template, canceling shape deformations. Combining these sophisticated visual front-ends with our features has the potential to do even better.

Table 1: Classification performance for visual speech classification on (a) AVLetters and (b) CUAVE. Deep autoencoders perform the best and show effective cross modality learning. Where indicated, the error bars show the variation (± 2 s.d.) due to random initialization. §Results are on continuous speech recognition performance, though we note that the *normal* portion of CUAVE has speakers saying isolated digits. †These models use a visual front-end system that is significantly more complicated than ours and a different train/test split.

Feature Representation	**Accuracy**
Baseline Preprocessed Video	46.2%
RBM Video (Figure 2b)	54.2%±3.3%
Video-Only Deep Autoencoder (Figure 3a)	**64.4%**±2.4%
Bimodal Deep Autoencoder (Figure 3b)	59.2%
Multiscale Spatial Analysis (Matthews et al., 2002)	44.6%
Local Binary Pattern (Zhao & Barnard, 2009)	58.85%

(a) AVLetters

Feature Representation	**Accuracy**
Baseline Preprocessed Video	58.5%
RBM Video (Figure 2b)	65.4%±0.6%
Video-Only Deep Autoencoder (Figure 3a)	**68.7%**±1.8%
Bimodal Deep Autoencoder (Figure 3b)	66.7%
Discrete Cosine Transform (Gurban & Thiran, 2009)	64% †§
Active Appearence Model (Papandreou et al., 2007)	75.7% †
Active Appearence Model (Pitsikalis et al., 2006)	68.7% †
Fused Holistic+Patch (Lucey & Sridharan, 2006)	77.08% †
Visemic AAM (Papandreou et al., 2009)	83% †§

(b) CUAVE Video

Table 2: Digit classification performance for bimodal speech classification on CUAVE, under clean and noisy conditions. We added white Gaussian noise to the original audio signal at 0 dB SNR. The error bars reflect the variation (± 2 s.d.) of the results due to the random noise added to the audio data. We compare performance of the Bimodal Deep Autoencoder model with the best audio features (Audio RBM) and the best video features (Video-only Deep Autoencoder).

Feature Representation	**Accuracy (Clean Audio)**	**Accuracy (Noisy Audio)**
(a) Audio RBM (Figure 2a)	**95.8%**	75.8% ± 2.0%
(b) Video-only Deep Autoencoder (Figure 3a)	68.7%	68.7%
(c) Bimodal Deep Autoencoder (Figure 3b)	90.0%	77.3% ± 1.4%
(d) **Bimodal + Audio RBM**	94.4%	**82.2%** ± 1.2%
(e) Video-only Deep AE + Audio-RBM	87.0%	76.6% ± 0.8%

These video classification results show that the deep autoencoders achieve cross modality learning by discovering better video representations when given additional audio data. In particular, even though the AVLetters dataset did not have any audio data, we were able to improve performance by learning better video features using other additional unlabeled audio and video data.

However, the bimodal deep autoencoder did not perform as well as the video-only deep autoencoder: while the video-only autoencoder learns only video features (which are also good for audio reconstruction), the bimodal autoencoder learns audio-only, video-only and invariant features. As such, the feature set learned by the bimodal autoencoder might not be optimal when the task at hand has only visual input.

We also note that cross modality learning for audio did not improve classification results compared to using audio RBM features; audio features are highly discriminative for speech classification, adding video information can sometimes hurt performance.

4.4. Multimodal Fusion Results

While using audio information alone performs reasonably well for speech recognition, fusing audio and video information can substantially improve performance, especially when the audio is degraded with noise (Gurban & Thiran, 2009; Papandreou et al., 2007; Pitsikalis et al., 2006; Papandreou et al., 2009). In particular, it is common to find that audio features perform well on their own and concatenating video features can sometimes hurt performance. Hence, we evaluate our models in both clean and noisy audio settings.

The video modality complements the audio modality by providing information such as place of articulation, which can help distinguish between similar sounding speech. However, when one simply concatenates audio and visual features (Table 2e), it is often the case that performance is worse as compared to using only audio features (Table 2a). Since our models are able to learn multimodal features that go beyond simply concatenating the audio and visual features, we propose combining the audio features with our multimodal features (Table 2d). When the best audio features are concatenated with the bimodal features, it outperforms the other feature combinations. This shows that the learned multimodal features are better able to complement the audio features.

4.5. McGurk effect

Table 3: McGurk Effect

Audio / Visual Setting	**Model prediction**		
	/ga/	*/ba/*	*/da/*
Visual */ga/*, Audio */ga/*	82.6%	2.2%	15.2%
Visual */ba/*, Audio */ba/*	4.4%	89.1%	6.5%
Visual */ga/*, Audio */ba/*	28.3%	13.0%	58.7%

The McGurk effect (McGurk & MacDonald, 1976) refers to an audio-visual perception phenomenon where a visual */ga/* with a audio */ba/* is perceived as */da/* by most subjects. Since our model learns a multimodal representation, it would be interesting to observe if the model is able to replicate a similar effect.

We obtained data from 23 volunteers speaking 5 repetitions of */ga/*, */ba/* and */da/*. The bimodal deep autoencoder features[7] were used to train a linear SVM on this 3-way classification task. The model was tested on three conditions that simulate the McGurk effect. When the visual and audio data matched at test time, the model was able to predict the correct class */ba/* and */ga/* with an accuracy of 82.6% and 89.1% respectively. On the other hand, when a visual */ga/* with a voiced */ba/* was mixed at test time, the model was most likely to predict */da/*, even though */da/* neither appears in the visual nor audio inputs, consistent with the McGurk effect on people. The same effect was not observed with the bimodal DBN (Figure 2d) or with concatenating audio and video RBM features.

[7] The */ga/*, */ba/* and */da/* data was not used for training the bimodal deep autoencoder.

4.6. Shared Representation Learning

Table 4: Shared representation learning on CUAVE. Chance performance is at 10%.

Train/Test	Method	Accuracy
Audio/Video	Raw-CCA	41.9%
	RBM-CCA Features	**57.3%**
	Bimodal Deep AE	30.7%
Video/Audio	Raw-CCA	42.9%
	RBM-CCA Features	**91.7%**
	Bimodal Deep AE	24.3%

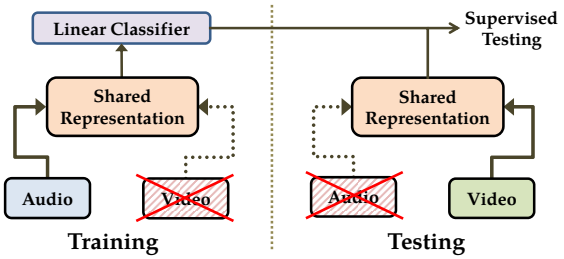


Figure 5: "Hearing to see" setting (train on audio, test on video) for evaluating shared representations.

In this experiment, we propose a novel setting which examines if a shared representation can be learned over audio and video speech data, During supervised training, the algorithm is provided data solely from one modality (e.g., audio) and later tested only on the other modality (e.g., video), as shown in Figure 5. In essence, we are telling the supervised learner how the digits "1", "2", etc. *sound*, while asking it to distinguish them based on how they are visually spoken – "hearing to see". If we are able to capture the correlations across the modalities in our shared representation, the model will perform this task well.

One approach to learning a shared representation is to find transformations for the modalities that maximize correlations. In particular, we suggest using canonical correlation analysis (CCA) (Hardoon et al., 2004), which finds linear transformations of audio and video data, to form a shared representation. [8] Learning a CCA shared representation on raw data results in surprisingly good performance (Table 4: Raw-CCA). However, learning the CCA representation on the first layer features (i.e., Audio RBM and Video RBM features) results in significantly better performance, comparable to using the original modalities for supervised classification (Table 4: RBM-CCA Features). This is particularly surprising since testing on audio performs better than testing on video, even when the model was trained on video data. These results show that capturing relationships across the modalities require at least a single non-linear stage to be successful. When good features have been learned from both modalities, a linear model can be well suited to capture the relationships. However, it is important to note that CCA, a linear transformation, does not help in other tasks such as cross-modality learning.

We further used this task to examine whether the features from the bimodal deep autoencoder captures correlations across the modalities. [9] While the bimodal deep autoencoder model does not perform as well as CCA, the results show that our learned representations are partially invariant to the input modality.

4.7. Additional Control Experiments

The video-only deep autoencoder has audio as a training cue and multiple hidden layers (Figure 3a). We first considered removing audio as a cue by training a similar deep autoencoder that did not reconstruct audio data; the performance decreased by 7.7% on CUAVE and 14.3% on AVLetters. Next, we trained a video-only shallow autoencoder with a single hidden layer to reconstruct both audio and video[10]; the performance decreased by 2.1% on CUAVE and 5.0% on AVLetters. Hence, both audio as a cue and depth were important ingredients for the video-only deep autoencoder to perform well.

We also compared the performance of using the bimodal DBN without training it as an autoencoder. In cases where only one modality was present, we used the same approach as the bimodal deep autoencoder, setting the absent modality to zero.[11] The bimodal DBN performed worse in the cross-modality and shared representation tasks and did not show the McGurk effect. It performed comparably on the multimodal fusion task. [12]

[8] Given audio data $\mathbf{a}$ and video data $\mathbf{v}$, CCA finds matrices P and Q such that $P\mathbf{a}$ and $Q\mathbf{v}$ have maximum correlations.

[9] For the bimodal deep autoencoder, we set the value of the absent modality to zero when computing the shared representation, which is consistent with the feature learning phase.

[10] The single hidden layer takes video as input and reconstructs both audio and video.

[11] We also tried alternating Gibbs sampling to obtain the posterior, but the results were worse.

[12] For the video-only setting, the bimodal DBN performed 4.9% worse on the CUAVE dataset and 5.0% worse on the AVLetters dataset. It was at chance on the "hearing to see" task and obtained 28.1% on "seeing to hear".

5. Related Work

While we present special cases of neural networks for multimodal learning, we note that prior work on audio-visual speech recognition (Duchnowski et al., 1994; Yuhas et al., 1989; Meier et al., 1996; Bregler & Konig, 1994) has also explored the use of neural networks. Yuhas et al. (1989) trained a neural network to predict the auditory signal given the visual input. They showed improved performance in a noisy setting when they combined the predicted auditory signal (from the network using visual input) with a noisy auditory signal. Duchnowski et al. (1994) and Meier et al. (1996) trained separate networks to model phonemes and visemes and combined the predictions at a phonetic layer to predict the spoken phoneme.

In contrast to these approaches, we use the hidden units to build a new representation of the data. Furthermore, we do not model phonemes or visemes, which require expensive labeling efforts. Finally, we build deep bimodal representations by modeling the correlations across the learned shallow representations.

6. Discussion

Hand-engineering task-specific features is often difficult and time consuming. For example, it is not immediately clear what the appropriate features should be for lipreading (visual-only data). This difficulty is more pronounced with multimodal data as the features have to relate multiple data sources. In this work, we showed how deep learning can be applied to this challenging task for discovering multimodal features.

Acknowledgments

We thank Clemson University for providing the CUAVE dataset and University of Surrey for providing the AVLetters2 dataset. We also thank Quoc Le, Andrew Saxe, Andrew Maas, and Adam Coates for insightful discussions, and the anonymous reviewers for helpful comments. This work is supported by the DARPA Deep Learning program under contract number FA8650-10-C-7020.

References

Bregler, C. and Konig, Y. "Eigenlips" for robust speech recognition. In *ICASSP*, 1994.

Caruana, R. Multitask learning. *Machine Learning*, 28(1): 41–75, 1997.

Cox, S., Harvey, R., Lan, Y., and Newman, J. The challenge of multispeaker lip-reading. In *International Conference on Auditory-Visual Speech Processing*, 2008.

Dalal, N. and Triggs, B. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005.

Duchnowski, P., Meier, U., and Waibel, A. See me, hear me: Integrating automatic speech recognition and lip-reading. In *ICSLP*, pp. 547–550, 1994.

Fisher, W., Doddington, G., and Marshall, Goudie. The DARPA speech recognition research database: Specification and status. In *DARPA Speech Recognition Workshop*, pp. 249–249, 1986.

Gurban, M. and Thiran, J.P. Information theoretic feature extraction for audio-visual speech recognition. *IEEE Trans. on Sig. Proc.*, 57(12):4765–4776, 2009.

Hardoon, David R., Szedmak, Sandor R., and Shawe-taylor, John R. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16:2639–2664, 2004.

Hinton, G. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 2002.

Hinton, G. and Salakhutdinov, R. Reducing the dimensionality of data with neural networks. *Science*, 313(5786): 504–507, 2006.

Lee, H., Ekanadham, C., and Ng, A. Sparse deep belief net model for visual area V2. In *NIPS*, 2007.

Lucey, P. and Sridharan, S. Patch-based representation of visual speech. In *HCSNet Workshop on the Use of Vision in Human-Computer Interaction*, 2006.

Matthews, I., Cootes, T.F., Bangham, J.A., and Cox, S. Extraction of visual features for lipreading. *PAMI*, 24: 198 –213, 2002.

McGurk, H. and MacDonald, J. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

Meier, U., Hürst, W., and Duchnowski, P. Adaptive Bimodal Sensor Fusion For Automatic Speechreading. In *ICASSP*, pp. 833–836, 1996.

Papandreou, G., Katsamanis, A., Pitsikalis, V., and Maragos, P. Multimodal fusion and learning with uncertain features applied to audiovisual speech recognition. In *MMSP*, pp. 264–267, 2007.

Papandreou, G., Katsamanis, A., Pitsikalis, V., and Maragos, P. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE TASLP*, 17(3):423–435, 2009.

Patterson, E., Gurbuz, S., Tufekci, Z., and Gowdy, J. CUAVE: A new audio-visual database for multimodal human-computer interface research. 2:2017–2020, 2002.

Pitsikalis, V., Katsamanis, A., Papandreou, G., and Maragos, P. Adaptive multimodal fusion by uncertainty compensation. In *ICSLP*, pp. 2458–2461, 2006.

Potamianos, G., Neti, C., Luettin, J., and Matthews, I. Audio-visual automatic speech recognition: An overview. In *Issues in Visual and Audio-Visual Speech Processing*. MIT Press, 2004.

Salakhutdinov, R. and Hinton, G. Semantic hashing. *IJAR*, 50(7):969–978, 2009.

Summerfield, Q. Lipreading and audio-visual speech perception. *Trans. R. Soc. Lond.*, pp. 71–78, 1992.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.A. Extracting and composing robust features with denoising autoencoders. In *ICML*, pp. 1096–1103. ACM, 2008.

Yuhas, B. P., Goldstein, M. H., and Sejnowski, T. J. Integration of acoustic and visual speech signals using neural networks. *IEEE Comm. Magazine*, pp. 65 –71, 1989.

Zhao, G. and Barnard, M. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11 (7):1254–1265, 2009.