Sensory Media Association through Reciprocating Training*

Qiong Liu[1], Hao Hu[1], Ray Yuan[2], Yanxia Zhang[1] and Yan-Ying Chen[1]

Abstract

Machine learning achieved great progress in recent years. However, state-of-the-art machine learning systems are still far behind biological learning systems on learning directly from sensors without offline labeling. This paper proposes an approach for automating machine learning from multi-modal sensors. In this learning setup, the system has no access to any human labeling tool which is not available to a biological learning system such as a dog or a newborn baby. We tested the learning proposal with audiovisual data. The testing system contains two deep autoencoders, one for learning speech representations and another for learning image representations. Two deep networks are trained to bridge the latent spaces of two autoencoders, yielding representation mappings for both speech-to-image and image-to-speech. To improve feature clustering in both latent spaces, the system alternately uses one modality to guide the learning of another modality. Different from traditional technology that uses a fixed modality for supervision (e.g. using text labels for image classification), the proposed approach facilitates a machine to learn from sensory inputs of two or more modalities through alternating guidance among these modalities. We evaluate the proposed model with MNIST digit images and corresponding digit speeches in the Google Command Digit Dataset (GCDD) and got very promising results.

1. Introduction

How to automate the learning from multi-modal sensors is an interesting and irresistible research idea [1], [2]. With methods that can directly learn from sensor inputs without offline labeling, machines can learn from daily activities, and a robot can continuously observe an environment and learn from it. These methods will also alleviate the dependence of machine learning on the availability of human text labelers. Moreover, this type of algorithm also provides a path to learn from multimodal data that has no labels. In this paper, we will present ideas in this direction.

With the increase of computing resources and the improvement of learning technology, researchers revisited the cross-modal/cross-media learning problem with deep neural network approaches [3], [4]. These approaches use common latent spaces achieved through simple concatenation, shared weights, or distance minimization. Common latent space of multiple modalities significantly increases the dimensionality of the data representation. It also creates interference between modalities. Different from using shared latent space, we propose transfer mapping across modalities that allows our system to encode images and audios in different latent spaces. This arrangement is more reliable when the data from both modalities are highly imbalanced.

Vukotic et al. [5] proposed an architecture to learn cross-modal data transfer. Its representation is more optimal than using a shared latent space. However, this approach requires data from both modalities available for learning. This requirement reduces its learning ability when a system only has data from one-modality, which is common in a real learning environment. Our proposed approach overcomes above issue by assigning an autoencoder learning module for each modality and perform cross-modal data-transfer learning across representation layers instead of raw data layers.

Article [6] used trained dot-products of image and audio features to illustrate relations between objects in an image and words in a sentence. While it provides a great heat-map visualization in a compressed space-and-time domain, the relation indication is still weak and the network cannot use input in one modality to generate output in another modality like a human does. To overcome above issues, we propose Sensory Media Association through Reciprocating Training (SMART). Similar to [6], the proposed model can learn without an offline labeling process. Different from [6], it can get input in one modality and generate output in another modality instead of saving the original datasets. Figure 1 shows a simple illustration of the idea.

*This work was fully supported by Fuji-Xerox Palo Alto Lab

[1]Authors are with Fuji-Xerox Palo Lab (FXPAL), 3174 Porter Drive, Palo Alto, CA 94304, USA liu at fxpal.com

[2]Author is with the EECS Department, University of California - Berkeley, 387 Soda Hall, Berkeley, CA 94720, USA

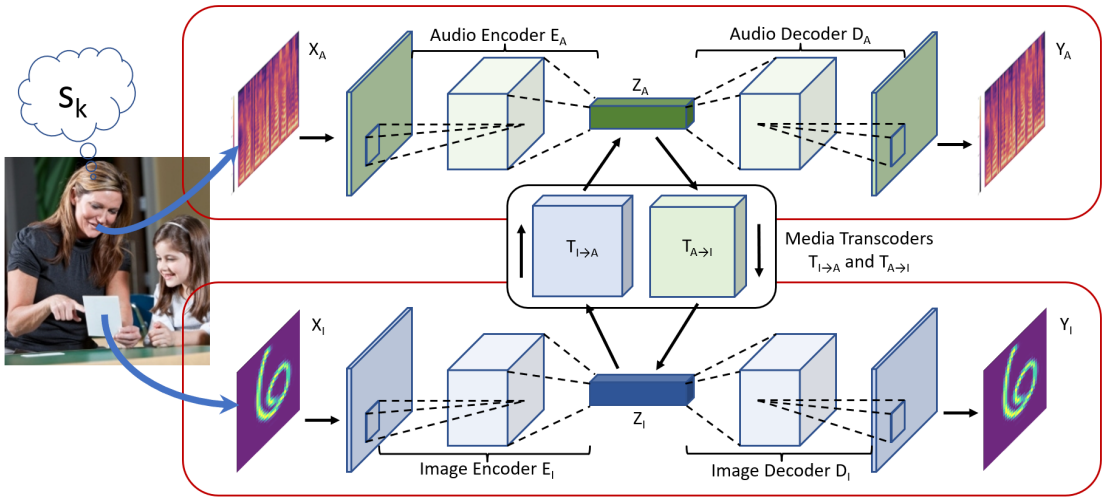

Figure 1. Proposed architecture for audiovisual cross-media learning. We use encoder and decoder architecture to let audio module and image module learn audio representations and image representations. Two deep networks are used to learn mappings from audio representation to image representation and image representation to audio representation respectively.

To evaluate the viability of the model, we set up a signal generation machine that can generate co-occurred multi-modal signals based on the same underlying concept. The machine mimics a human instructor's teaching process (e.g. an instructor shows handwriting of a digit when s/he pronounces the digit). In this process, the co-occurred signals in different media channels (i.e. speech and vision) are related to the same underlying concept, a digit. With this signal-generation machine, we can generate training signals for the learning module. After training, we feed the model with test signals in one modality to generate corresponding output signals in another modality. That is similar to showing a handwriting to a student to get a corresponding audio response or pronouncing a word to the student to get a handwriting response. By checking the underlying concept alignments of input and output signals in different modalities, we can evaluate the learning performance in the same way that a human teacher does.

2. Methodology

Figure 1 shows the proposed network architecture. Audio and image are two modalities for our experiments. In this setup, an encoder and decoder architecture is used for both the audio module and the image module to learn their corresponding representations respectively. To reduce memory usage, audio signals are converted to 2D logarithmic melspectrograms. With this conversion, the audio module and image module can use very similar convolutional autoencoders. Having latent representations for each modality, we use one multi-layer fully connected neural network to map an audio representation to an image representation and use a similar network to map in the opposite direction. By using these networks, an audio input can activate an audio output as well as an image output. Similarly, an image input can activate an image output as well as an audio output. Compared to a shared latent space, mapping across modalities allows the SMART system to encode images and audios in different latent spaces. Using cross-modality mapping in our system can avoid some disadvantages described early.

Having noisy data clustered well in a meaningful way is a challenging task in real environment. If the latent spaces of both modalities have clean data that already can be clustered well within one modality, guidance from another modality will not have significant impact on the clustering results. However, data from real environment are noisy at most of the time, especially when a machine needs to handle a large amount of data. In this situation, supervision from another dataset becomes important. To achieve the supervision, the system may use representations in one latent space to supervise noisy data learning in another latent space. This idea also aligns with our observation of a baby teaching process while a mom hold a digit card and pronounce it again and again or pronounce a word and show different images corresponding to the word.

2.1. Models

To make the model and teaching process explanation clear, we define notations as follows:

- $\mathbf{A}/\mathbf{I}$ represents audio/image modality respectively.
- $\mathbf{X}/\mathbf{Y}/\mathbf{Z}$ represents input/output/latent space respectively and $\mathbf{x}/\mathbf{y}/\mathbf{z}$ represents a sample in these spaces respectively.
- B is the training batch size. C indicates a selected representation closest to the center of a data cluster.
- $\mathbf{E}/\mathbf{D}/\mathbf{T}$ represents encoder/decoder/transcoder respectively so that $\mathbf{z} = \mathbf{E}(\mathbf{x})$, $\mathbf{y} = \mathbf{D}(\mathbf{z})$, $\mathbf{z_I} = \mathbf{T_{A\to I}}(\mathbf{z_A})$, $\mathbf{z_A} = \mathbf{T_{I\to A}}(\mathbf{z_I})$.
- S is the conceptual state space and s is a sample state in the space. Conceptual states are the underlying states for defining signal co-occurrences in different modalities.
- k, m, n are indexes for underlying states, speech samples, and image samples respectively and K, M, N are their maximum values respectively.
- p, and q are the indices of training iterations.

With training procedure described in previous paragraphs, we define loss function for training $\mathbf{E_I}, \mathbf{D_I}$, and $\mathbf{T_{I\to A}}$ as:

$$\mathbf{L_{I\to A}} = \sum_{n_k=0}^{B} \Big(\big|\big| \mathbf{x}_{\mathbf{I},n_k} - \mathbf{D}_{\mathbf{I},p}\big(\mathbf{E}_{\mathbf{I},p}(\mathbf{x}_{\mathbf{I},n_k})\big) \big|\big| + \big|\big| \mathbf{z}_{\mathbf{A},m_{c,k}} - \mathbf{T}_{\mathbf{I}\to\mathbf{A},p}\big(\mathbf{E}_{\mathbf{I},p}(\mathbf{x}_{\mathbf{I},n_k})\big) \big|\big| \Big), \tag{1}$$

and define loss function for training $\mathbf{E_A}, \mathbf{D_A}$, and $\mathbf{T_{A\to I}}$ as:

$$\mathbf{L_{A\to I}} = \sum_{m_k=0}^{B} \Big(\big|\big| \mathbf{x}_{\mathbf{A},m_k} - \mathbf{D}_{\mathbf{A},q}\big(\mathbf{E}_{\mathbf{A},q}(\mathbf{x}_{\mathbf{A},m_k})\big) \big|\big| + \big|\big| \mathbf{z}_{\mathbf{I},n_{c,k}} - \mathbf{T}_{\mathbf{A}\to\mathbf{I},q}\big(\mathbf{E}_{\mathbf{A},q}(\mathbf{x}_{\mathbf{A},m_k})\big) \big|\big| \Big), \tag{2}$$

where $\mathbf{z}_{\mathbf{A},m_{c,k}}$ and $\mathbf{z}_{\mathbf{I},n_{c,k}}$ are tentative frozen representations (details in the evaluation section) for guiding the training of another modality. In both equations, the first terms enforce the learning of the data representation manifold while the second terms enforce the margin increase among data clusters with different underlying states. The increased margins learned from cross-modality transformation may help the system to reduce communication error related to concepts.

The reciprocating training process of SMART has similarities and dissimilarities with traditional supervised learning and unsupervised learning. Similar to supervised learning, SMART does have two or more signal spaces and guides learning process in one space with signals from another space. On the other hand, the feature of sensory data may take guiding or guided role in turn. Unlike traditional supervised learning, which explicitly specify supervisory signal such as text labels in data, SMART does not use explicit labels. Different from unsupervised learning which infers a signal-structure function without any guidance from another signal space, SMART forms signal-structure function with guidance from co-occurred data points in another signal space. These similarities and dissimilarities enables the SMART approach to form meaningful associations without connecting data in two modalities to explicit labels.

3. Evaluations

In the evaluation network, both the audio module and image module have 21 convolution layers, 2 fully connected layers, and matched up/down sampling layers, normalization layers, Leaky ReLU layers. Each of these two modules has a 512-dimension latent space. Both cross-modal networks have 5 fully connected layers. The image module has 32×32 input and output. The audio module has 96×96 input and output. Since the long-term goal of the SMART learning system is to autonomously learn from different sensory data like a baby, we need to test the learning system with simulated setups before we connect sensors to the system for learning in real environment.

To simulate human teaching/learning scenarios, we choose Google Command Digit Dataset (GCDD) and MNIST dataset as audio and image datasets for creating a signal generation machine that mimics a human teacher. In the machine, we set 10 underlying states 0-9 for generating co-occurred image and audio signals. With this setup, when the machine wants to teach a concept (e.g. '0'), it picks a '0' sample from GCDD and a '0' sample from MNIST and forward them together to the learning system. This is similar to the teaching process when a teacher writes down '0' and pronounce '0' at the same time. Beyond generating co-occurred media in two channels, the machine is also capable to generate one channel output (i.e. audio or image only). That is similar to seeing digits without knowing their pronunciations or hearing digit pronunciations without knowing corresponding hand-writings.

Since the GCDD has 23,666 speech wave-forms with around 2,300 wave-forms for each of 10 digits, we pick first 2,000 wave-forms from each digit category for training and use the rest as testing data. With this selection, we have 20,000 audio waveforms for training and 3,666 audio waveforms for testing. To

reduce memory usage and model sizes, we use audio tools by Kastner [7] to convert audio waveforms to 2D logarithmic mel-spectrograms. The logarithmic mel-spectrogram transformation has 96 Mel filter bands, a 93ms transformation window, and a 58ms shift step. This setup allows the system to handle speech segments up-to 5 seconds with a 96×96 2D logarithmic mel-spectrogram. In the MNIST dataset, there are 60,000 digit images for training and 10,000 images for testing. To make our max pooling and upsampling adjustment easier, we resize these training and testing images to 32×32 dimension.

Table 1. Comparison between SMART conversions and text label supervised classifiers. Bottom two lines are results from testing proposed approaches.

APPROACHES	ACC(%)
MNIST LABEL CLASSIFIER (MLC)	99.47
GCDD LABEL CLASSIFIER (GLC)	96.15
MNIST $\rightarrow \mathbf{E_I} \rightarrow \mathbf{T_{I \rightarrow A}} \rightarrow \mathbf{D_A} \rightarrow$ GLC (HYBRID STYLE TRAINING)	**99.59**
GCDD $\rightarrow \mathbf{E_A} \rightarrow \mathbf{T_{A \rightarrow I}} \rightarrow \mathbf{D_I} \rightarrow$ MLC (HYBRID STYLE TRAINING)	**97.33**
MNIST $\rightarrow \mathbf{E_I} \rightarrow \mathbf{T_{I \rightarrow A}} \rightarrow \mathbf{D_A} \rightarrow$ GLC (CLASSROOM STYLE TRAINING)	99.05
GCDD $\rightarrow \mathbf{E_A} \rightarrow \mathbf{T_{A \rightarrow I}} \rightarrow \mathbf{D_I} \rightarrow$ MLC (CLASSROOM STYLE TRAINING)	97.19

To avoid hiring human graders to check underlying-state matches for 3,666 testing audio inputs and 10,000 testing image inputs and make the evaluation process more objective, we trained an audio pattern classifier and an image pattern classifier with labeled GCDD and MNIST training data. The image classifier has a 99.47% labeling accuracy on 10,000 MNIST testing data and the audio classifier has a 96.15% labeling accuracy on 3,666 GCDD testing data. These data are shown in Table 1. Due to high testing accuracies, we believe the trained classifiers are reliable to judge the state matching correctly.

After classifiers' training, we trained the proposed model with a hybrid approach that separately trains autoencoders in one modality first and then retrain the whole network with pre-trained autoencoders. Then, we feed 10,000 MNIST testing images to the SMART system to get 10,000 audio outputs, and these 10,000 audio outputs are fed to the audio classifier to predict the underlying states. Similarly, we feed 3,666 GCDD testing audio segments to the system to get 3,666 image outputs. These 3,666 output images are then fed to the image classifier for predicting their underlying states. The matching percentages of the predicted and ground-truth states are reported in the 3-4 lines of Table 1.

To our surprise, the image-to-audio activation result can reach 99.59% which is higher than the image classifier accuracy. Similarly, the audio-to-image activation result, 97.33% is also higher than the audio label classifier labeling accuracy. After thinking the process carefully and compare it with a similar human operation process, we think this is possible and reasonable. In reality, we occasionally see people read handwriting numbers and have another person type these numbers in a computer. In this process, the first person performs image-to-audio conversion and second person give the digit pronunciations labels in the computer. If the first person is well trained on handwriting recognition and have a clear pronunciation and the second person is less experienced on handwriting recognition, the combined labeling accuracy may be higher than the labeling accuracy when the second person is assigned to perform the handwriting-to-label task alone. Similar results will happen when one person hear digit pronunciations with dialect and write them down in a easy to recognize format and the second person who is not used to the dialect inputs the digit handwriting in a computer.

References

[1] N. Wiener. 1948.Cybernetics or control and communication in the animal andmachine(2nd ed.). MIT Press.

[2] A. M. Turing. 1950. Computing machinery and intelligence.Mindcol, 59 (1950),433460.

[3] S. Chaudhury, S. Dasgupta, A. Munawar, Md. A. S. Khan, and R. Tachibana. 2017.Conditional generation of multi-modal data using constrained embedding spacemapping. InProceedings of ICML Workshop on Implicit Models. Sydney, Australia.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. 2011. Multimodal deeplearning. InProceedings of International Conference on Machine Learning 2011.Bellevue, Washington, USA.

[5] V. Vukotic, C. Raymond, and G. Gravier. 2016. Bidirectional Joint Represen-tation Learning with Symmetrical Deep Neural Networks for Multimodal andCross-modal Applications. InProceedings of ACM International Conference onMultimedia Retrieval. New York, United States.

[6] D. Harwath, A. Recasens, D. Surs, G. Chuang, A. Torralba, and J.R. Glass.2018. Jointly Discovering Visual Objects and Spoken Words from Raw SensoryInput. InProceedings of European Conference on Computer Vision (ECCV) 2018. Munich, Germany.

[7] Kyle Kastner. 2018. Audio tools for numpy/python.