Multimodal Machine Learning: Integrating Language, Vision and Speech

Louis-Philippe Morency
Language Technologies Institute
Carnegie Mellon University
morency@cs.cmu.edu

Tadas Baltrušaitis
Language Technologies Institute
Carnegie Mellon University
tbaltrus@cs.cmu.edu

Abstract

Multimodal machine learning is a vibrant multi-disciplinary research field which addresses some of the original goals of artificial intelligence by integrating and modeling multiple communicative modalities, including linguistic, acoustic and visual messages. With the initial research on audio-visual speech recognition and more recently with language & vision projects such as image and video captioning and visual question answering, this research field brings some unique challenges for multimodal researchers given the heterogeneity of the data and the contingency often found between modalities.

Tutorial overview

The present tutorial will review fundamental concepts of machine learning and deep neural networks before describing the five main challenges in multimodal machine learning:

1. **Representation**: A first fundamental challenge is to learn how to represent and summarize the multimodal data to highlight the complementarity and synchrony between modalities. The heterogeneity of multimodal data makes it particularly challenging for coordinated and joint representations. For example, language is often seen as symbolic while audio and visual modalities will be represented as signals.

2. **Translation**: A second challenge is how to translate data from one modality to another. Not only is the data heterogeneous, but the relationship between modalities is often open-ended or subjective. For example, when describing a specific image verbally, more than one description can be correct. The evaluation and characterization of the multimodal translation may be subjective.

3. **Alignment**: A third challenge is to identify the direct relations between elements from two or more different modalities. For example, when analyzing the speech and gestures of a human subject, how can we align specific gestures with the spoken words or utterances? This alignment between modalities may be based on long-range dependencies and the segmentation is often ambiguous (e.g., words or utterances).

4. **Fusion**: A fourth challenge is to join information from two or more modalities to perform a prediction, discrete or continuous. For example, for audio-visual speech recognition, the visual descriptoin of the lip motion is fused with the speech signal to predict spoken words. The information coming from different modalities may have varying predictive power and noise topology. With possibly missing data in at least one of the modalities. Multimodal fusion needs to handle such variations.

5. **Co-learning**: A fifth challenge is to transfer knowledge between modalities and their representations. Exemplified by algorithms of co-training, conceptual grounding and zero shot learning, how does knowledge learning from one modality (e.g., predicted labels or representation) can help a computational model trained on a different modality? This challenge is particularly relevant when one of the modalities has limited resources (e.g., annotated data).

The tutorial will also present state-of-the-art algorithms that were recently proposed to solve mul-

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-Tutorial Abstracts, pages 3–5
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-5002

timodal applications such as image captioning, video descriptions and visual question-answer. We will also discuss the current and upcoming challenges.

Structure

We plan to follow a similar structure to our ICMI 2016 tutorial which was 3 hours long:

1. Intruduction
    - What is Multimodal?
        - Historical view, multimodal vs multimedia
    - Why multimodal?
        - Multimodal applications: image captioning, video description, AVSR,
    - Core technical challenges
        - Representation learning, translation, alignment, fusion and co-learning
2. Basic concepts — Part 1
    - Linear models
        - Score and loss functions, regularization
    - Neural networks
        - Activation functions, multi-layer perceptron
    - Optimization
        - Stochastic gradient descent, backpropagation
3. Unimodal representations
    - Language representations
        - Distributional hypothesis and word embedding
    - Visual representations
        - Convolutional neural networks
    - Acoustic representations
        - Spectrograms, auto-encoders
4. Multimodal representations
    - Joint representations
        - Visual semantic spaces, multimodal auto-encoder
    - Coordinated representations
        - Component analysis
        - Similarity metrics, canonical correlation analysis

===Break===

1. Basic concepts — Part 2
    - Language models
        - Unigrams, bigrams, skip-grams, skip-thought
    - Unimodal sequence modeling
        - Recurrent neural networks, LSTMs
    - Optimization
        - Backpropagation through time
2. Multimodal translation and mapping
    - Encoder-decoder models
        - Machine translation, image captioning
    - Generative vs example based approaches
        - Viseme generation, visual puppetry
        - Model evaluation
3. Modality alignment
    - Latent alignment approaches
        - Attention models, multi instance learning
    - Explicit alignment
        - Dynamic time warping
4. Multimodal fusion and co-learning
    - Model free approaches
        - Early and late fusion, hybrid models
    - Kernel-based fusion
        - Multiple kernel learning
    - Multimodal graphical models
        - Factorial HMM, Multi-view Hidden CRF
    - Co-learning
        - Parallel, non-parallel and hybrid data
5. Future directions and concluding remarks

About the speakers

Louis-Philippe Morency (https://www.cs.cmu.edu/~morency/) is Assistant Professor in the Language Technology Institute at the Carnegie Mellon University where he leads the Multimodal Communication and Machine Learning Laboratory (MultiComp Lab). He received

his Ph.D. and Master degrees from MIT Computer Science and Artificial Intelligence Laboratory. In 2008, Dr. Morency was selected as one of "AI's 10 to Watch" by IEEE Intelligent Systems. He has received 7 best paper awards in multiple ACM- and IEEE-sponsored conferences for his work on context-based gesture recognition, multimodal probabilistic fusion and computational models of human communication dynamics. Dr. Morency was General Chair for the International Conference on Multimodal Interaction (ICMI 2012) and the NIPS 2010 workshop on Modeling Human Communication Dynamics. He was Program Chair for ICMI 2011, 2014 and 2016, as well as the Tenth International Conference on Creating, Connecting and Collaborating through Computing in January 2012.

Tadas Baltrušaitis (http://www.cl.cam.ac.uk/~tb346/) is a post-doctoral associate at the Language Technologies Institute, Carnegie Mellon University. Before this, he was a post-doctoral research at the University of Cambridge, where he also received his PhD degree in 2014. His primary research interests lie in the automatic understanding of non-verbal human behaviour, computer vision, and multimodal machine learning. His papers have won a number of awards for his work on non-verbal human behavior analysis, including ICMI 2014 best student paper award, and ETRA 2016 emerging investigator award. He is also a winner of several challenges in computer vision and multi-modal machine learning, including FERA 2015, and AVEC 2011.