1 Challenges and Applications in Multimodal Machine Learning

Tadas Baltrušaitis, Chaitanya Ahuja, Louis-Philippe Morency

1.1 Introduction

The world surrounding us involves multiple modalities. We see objects, hear sounds, feel texture, smell odors, and so on. In general terms, a *modality* refers to the way in which something happens or is experienced. Most people associate the word modality with the *sensory modalities* which represent our primary channels of communication and sensation, such as vision or touch. In this chapter we focus primarily, but not exclusively, on three such modalities: *linguistic* modality which can be both written or spoken; *visual* modality which is often represented with images or videos; and *vocal* modality which encodes sounds and para-verbal information such as prosody and vocal expressions.

In order for Artificial Intelligence to make progress in understanding the world around us, it needs to be able to interpret and reason about multimodal messages. *Multimodal machine learning* aims to build models that can process and relate information from multiple modalities. From early research on audio-visual speech recognition to the recent explosion of interest in language and vision models, multimodal machine learning is a vibrant multi-disciplinary field of increasing importance and with extraordinary potential.

Glossary

Representation learns how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy of multiple modalities. The heterogeneity of multimodal data makes it challenging to construct such representations. For example, language is often symbolic while audio and visual modalities will be represented as signals.

Translation addresses how to translate (map) 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, there exist a number of *correct* ways to describe an image and one perfect translation may not exist.

Alignment identifies the direct relations between (sub)elements from two or more different modalities. For example, we may want to align the steps in a recipe to a video showing the dish being made. To tackle this challenge we need to measure similarity between different modalities and deal with possible long-range dependencies and ambiguities.

Fusion joins information from two or more modalities to perform a prediction. For example, for audio-visual speech recognition, the visual description 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.

Co-learning transfers knowledge between modalities, their representation, and their predictive models. This is exemplified by algorithms of co-training, conceptual grounding, and zero shot learning. Co-learning explores how knowledge learning from one modality 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., few annotated data).

The research field of multimodal machine learning brings some unique challenges for computational researchers given the heterogeneity of the data. Learning from multimodal sources offers the possibility of capturing correspondences between modalities and gaining an in-depth understanding of natural phenomena. In a recent survey paper, Baltrušaitis et al. [2017] identify five core technical challenges (and related sub-challenges) surrounding multimodal machine learning. They are (a) ***representation***, (b) ***translation***, (c) ***alignment***, (d) ***fusion***, and (e) ***co-learning*** (for definitions see the Glossary). They are central to the multimodal setting and need to be tackled in order to progress the field.

We start with a discussion of main applications of multimodal machine learning (Section 1.2). In this chapter we focus on two out of the five core technical challenges facing multimodal machine learning: representation (Section 1.3) and co-learning (Section 1.4). The fusion challenge is addressed in Chapters 2, 4 and Jameson and Kristensson [2017], and part of translation challenge is discussed in Chapter 4. More details about all five challenges are also available in the survey paper [Baltrušaitis et al. 2017].

1.2 Multimodal Applications

Multimodal machine learning enables a wide range of applications: from audio-visual speech recognition to image captioning. In this section we present a brief history of multimodal applications, from its beginnings in audio-visual speech recognition to a recently renewed interest in language and vision applications.

One of the earliest examples of multimodal research is audio-visual speech recognition (AVSR) [Yuhas et al. 1989]. It was motivated by the McGurk effect [McGurk and MacDonald 1976], an interaction between hearing and vision during speech perception. When human subjects heard the syllable /ba-ba/ while watching the lips of a person saying /ga-ga/, they perceived a third sound: /da-da/. These results motivated many researchers from the speech community to extend their approaches with visual information. Given the prominence of hidden Markov models (HMMs) in the speech community at the time [Juang and Rabiner 1991], it is without surprise that many of the early models for AVSR were based on various HMM extensions [Bourlard and Dupont 1996, Brand et al. 1997]. While research into AVSR is not as common these days, it has seen renewed interest from the deep learning community [Ngiam et al. 2011].

While the original vision of AVSR was to improve speech recognition performance (e.g., word error rate) in all contexts, the experimental results showed that the main advantage of visual information was when the speech signal was noisy (i.e., low signal-to-noise ratio) [Yuhas et al. 1989, Gurban et al. 2008, Ngiam et al. 2011]. In other words, the captured interactions between modalities were supplementary rather than complementary. The same information was captured in both, improving the robustness of the multimodal models but not improving the speech recognition performance in noiseless scenarios.

A second important category of multimodal applications comes from the field of multimedia content indexing and retrieval [Snoek and Worring 2005, Atrey et al. 2010]. With the advance of personal computers and the internet, the quantity of digitized multimedia content has increased dramatically.[1] While earlier approaches for indexing and searching these multimedia videos were keyword-based [Snoek and Worring 2005], new research problems emerged when trying to search the visual and multimodal content directly. This led to new research topics in multimedia content analysis such as automatic shot-boundary detection [Lienhart 1999] and video summarization [Evangelopoulos et al. 2013]. These research projects were supported by the TrecVid initiative from the National Institute of Standards and Technologies which introduced many high-quality datasets, including the multimedia event detection (MED) tasks started in 2011.[2]

A third category of applications was established in the early 2000s around the emerging field of multimodal interaction with the goal of understanding human multimodal behaviors during social interactions. One of the first landmark datasets collected in this field is the AMI Meeting Corpus which contains more than 100 hours of video recordings of meetings, all fully transcribed and annotated [Carletta et al. 2005]. Another important dataset is the SEMAINE corpus which allowed to study interpersonal dynamics between speakers and listeners [McKeown et al. 2010]. This dataset formed the basis of the first audio-visual emotion challenge (AVEC) organized in 2011 [Schuller et al. 2011]. The fields of emotion recognition and affective computing bloomed in the early 2010s thanks to strong technical advances in automatic face detection, facial landmark detection, and facial expression recognition [De la Torre and Cohn 2011]. The AVEC challenge continued annually afterward with the later instantiation including healthcare applications such as automatic assessment of depression and anxiety [Valstar et al. 2013]. A great summary of recent progress in multimodal affect recognition was published by D'Mello and Kory [2015]. Their meta-analysis revealed that a majority of recent work on multimodal affect recognition show improvement when using more than one modality, but this improvement is reduced when recognizing naturally occurring emotions.

Most recently, a new category of multimodal applications emerged with an emphasis on language and vision: media description. One of the most representative applications is image captioning where the task is to generate a text description of the input image [Hodosh et al. 2013]. This is motivated by the ability of such systems to help the visually impaired in their daily tasks [Bigham et al. 2010]. The main challenges media description is evaluation: how to evaluate the quality of the

[1] http://www.youtube.com/intl/en-US/yt/about/press/ (accessed May 2018)

[2] http://www.nist.gov/multimodal-information-group/trecvid-multimedia-event-detection-2011-evaluation (accessed May 2018)

predicted descriptions. The task of visual question-answering (VQA) was recently proposed to address some of the evaluation challenges [Antol et al. 2015], where the goal is to answer a specific question about the image.

In order to bring some of the mentioned applications to the real world we need to address a number of technical challenges facing multimodal machine learning. We summarize the relevant technical challenges for the above-mentioned application areas in Table 1.1. One of the most important challenges is multimodal representation, the focus of our next section.

1.3 Multimodal Representations

Representing raw data in a format that a computational model can work with has always been a big challenge in machine learning. Following the work of Bengio et al. [2013], we use the term feature and representation interchangeably, with each referring to a vector or tensor representation of an entity, be it an image, audio sample, individual word, or a sentence. A multimodal representation is a representation of data using information from multiple such entities. Representing multiple modalities poses many difficulties: how to combine the data from heterogeneous sources; how to deal with different levels of noise; and how to deal with missing data. The ability to represent data in a meaningful way is crucial to multimodal problems, and forms the backbone of any model.

Good representations are important for the performance of machine learning models, as evidenced behind the recent leaps in performance of speech recognition [Amodei et al. 2016, Hinton et al. 2012] and visual object classification [Krizhevsky et al. 2012] systems. Bengio et al. [2013] identify a number of properties for good representations: smoothness, temporal and spatial coherence, sparsity, and natural clustering, among others. Srivastava and Salakhutdinov [2012b] identify additional desirable properties for multimodal representations: similarity in the representation space should reflect the similarity of the corresponding concepts, the representation should be easy to obtain even in the absence of some modalities, and, finally, it should be possible to fill-in missing modalities given the observed ones.

The development of unimodal representations has been extensively studied [Anagnostopoulos et al. 2015, Bengio et al. 2013, Li et al. 2015]. In the past decade there has been a shift from hand-designed representations to data-driven ones. For example, one of the most famous image descriptors in the early 2000s, the scale invariant feature transform (SIFT) was hand designed [Lowe 2004], but currently most visual descriptions are learned from data using neural architectures

Table 1.1 **A summary of applications enabled by multimodal machine learning. For each application area we identify the core technical challenges that need to be addressed in order to tackle it.**

	Challenges				
Applications	Representation	Translation	Fusion	Alignment	Co-learning
Speech Recognition and Synthesis					
AVSR	✓		✓	✓	✓
Speech synthesis	✓	✓			
Event Detection					
Action Classification	✓		✓		✓
Multimedia Event Detection	✓		✓		✓
Emotion and Affect					
Recognition	✓		✓	✓	✓
Synthesis	✓	✓			
Media Description					
Image Description	✓	✓		✓	✓
Video Description	✓	✓	✓	✓	✓
Visual Question-Answering	✓		✓	✓	✓
Media Summarization	✓	✓	✓		
Multimedia Retrieval					
Cross Modal retrieval	✓	✓		✓	✓
Cross Modal hashing	✓				✓

such as convolutional neural networks (CNN) [Krizhevsky et al. 2012]. Similarly, in the audio domain, acoustic features such as Mel-frequency cepstral coefficients (MFCC) have been superseded by data-driven deep neural networks in speech recognition [Amodei et al. 2016, Hinton et al. 2012] and recurrent neural networks for para-linguistic analysis [Trigeorgis et al. 2016]. In natural language processing, the textual features initially relied on counting word occurrences in documents, but have been replaced by data-driven word embeddings that exploit the word context [Mikolov et al. 2013]. While there has been a huge amount of work on unimodal representation, up until recently most multimodal representations involved simple concatenation of unimodal ones [D'Mello and Kory 2015], but this has been rapidly changing.

To help understand the breadth of work, we propose two categories of multimodal representation: *joint* and *coordinated*. Joint representations combine the unimodal signals into the same representation space, while coordinated representations process unimodal signals separately, but enforce certain similarity constraints on them to bring them to what we term a coordinated space. An illustration of different multimodal representation types can be seen in Figure 1.1.

Mathematically, the joint representation is expressed as:

$$\mathbf{x}_m = f(\mathbf{x}_1, \ldots, \mathbf{x}_n), \tag{1.1}$$

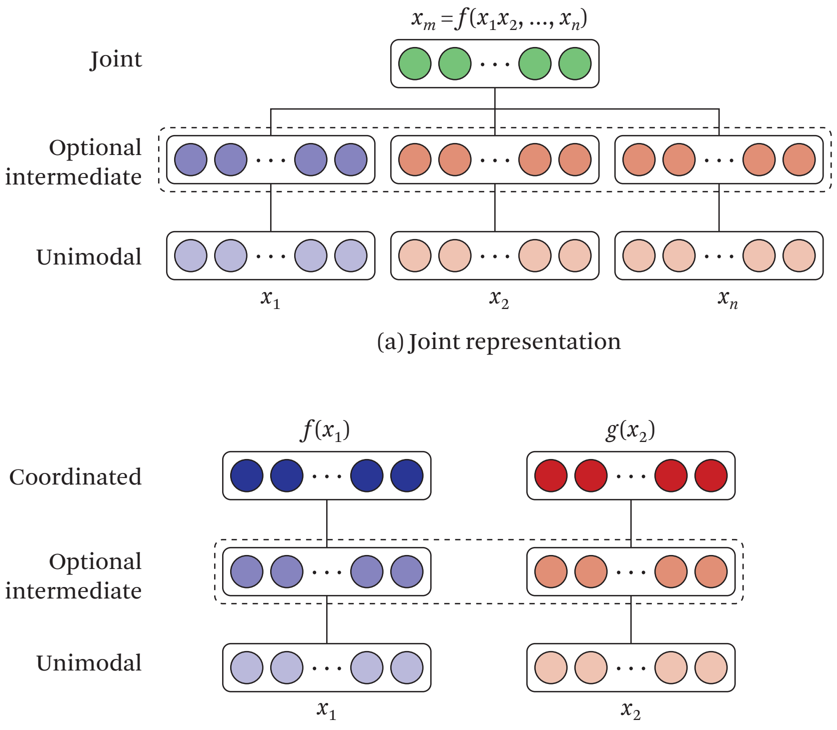
where the multimodal representation $\mathbf{x}_m$ is computed using function f (e.g., a deep neural network, restricted Boltzmann machine, or a recurrent neural network) that relies on unimodal representations $\mathbf{x}_1, \ldots \mathbf{x}_n$. On the other hand, the coordinated representation is as follows:

$$f(\mathbf{x}_1) \sim g(\mathbf{x}_2), \tag{1.2}$$

where each modality has a corresponding projection function (f and g above) that maps it into a coordinated multimodal space. While the projection into the multimodal space is independent for each modality, but the resulting space is coordinated between them (indicated as $\sim$). Examples of such coordination include minimizing cosine distance [Frome et al. 2013], maximizing correlation [Andrew et al. 2013], and enforcing a partial order [Vendrov et al. 2016] between the resulting spaces.

1.3.1 Joint Representations

We start our discussion with joint representations that project unimodal representations together into a multimodal space (Equation 1.1). Joint representations are



(a) Joint representation

(b) Coordinated representations

Figure 1.1 Structure of *joint* and *coordinated* representations. Joint representations are projected to the same space using all of the modalities as input. Coordinated representations, on the other hand, exist in their own space, but are coordinated through a similarity (e.g., euclidean distance) or structure constraint (e.g., partial order).

mostly (but not exclusively) used in tasks where multimodal data is present both during training and inference steps. The simplest example of a joint representation is a concatenation of individual modality features (also referred to as early fusion [D'Mello and Kory 2015]). In this section we discuss more advanced methods for creating joint representations starting with neural networks, followed by graphical models and recurrent neural networks (representative works can be seen in Table 1.2).

Neural networks have become a very popular method for unimodal data representation [Bengio et al. 2013]. They are used to represent visual, acoustic, and textual data, and are increasingly used in the multimodal domain [Ngiam et al. 2011, Wang et al. 2015a, Ouyang et al. 2014]. In this section we describe how neu-

Table 1.2 **A summary of multimodal representation techniques. We identify three subtypes of joint representations (Section 1.3.1) and two subtypes of coordinated ones (Section 1.3.2). For modalities + indicates the modalities combined.**

Representation	Modalities	Reference
Joint		
Neural networks	Images + Audio	[Ngiam et al. 2011, Mroueh et al. 2015]
	Images + Text	[Silberer and Lapata 2014]
Graphical models	Images + Text	[Srivastava and Salakhutdinov 2012b]
	Images + Audio	[Kim et al. 2013]
Sequential	Audio + Video	[Kahou et al. 2016, Nicolaou et al. 2011]
	Images + Text	[Rajagopalan et al. 2016]
Coordinated		
Similarity	Images + Text	[Frome et al. 2013, Kiros et al. 2015]
	Video + Text	[Xu et al. 2015, Pan et al. 2016]
Structured	Images + Text	[Cao et al. 2016, Vendrov et al. 2016]
	Audio + Articulatory	[Wang et al. 2015b]

ral networks can be used to construct a joint multimodal representation, and what advantages they offer.

In general, neural networks are made up of successive layers of inner products followed by nonlinear activation functions. In order to use a neural network as a way to represent data, it is first trained to perform a specific task (e.g., recognizing objects in images). Due to the multilayer nature of deep neural networks, each successive layer is hypothesized to represent the data in a more abstract way [Bengio et al. 2013], hence it is common to use the final or penultimate neural layers as a form of data representation. To construct a multimodal representation using neural networks each modality starts with several individual neural layers followed by a hidden layer that projects the modalities into a joint space [Wu et al. 2014, Mroueh et al. 2015, Antol et al. 2015, Ouyang et al. 2014]. The joint multimodal representation is then passed through multiple hidden layers or used directly for prediction. Such models can be trained end-to-end, learning both to represent the data and to perform a particular task. This results in a close relationship between multimodal representation learning and multimodal fusion when using neural networks.

As neural networks require a lot of labeled training data, one approach is to pre-train such representations using an autoencoder on unsupervised data [Hinton and Zemel 1994]. The model proposed by Ngiam et al. [2011] extended the idea of using autoencoders to the multimodal domain. They used stacked denoising autoencoders to represent each modality individually and then fused them into a multimodal representation using another autoencoder layer. Similarly, Silberer and Lapata [2014] proposed using a multimodal autoencoder for the task of semantic concept grounding (see Section 1.4.2). In addition to using a reconstruction loss to train the representation they introduce a term into the loss function that uses the representation to predict object labels. It is also common to fine-tune the resulting representation on a particular task at hand as the representation constructed using an autoencoder is generic and not necessarily optimal for a specific task [Wang et al. 2015a].

The major advantage of neural network-based joint representations comes from their often superior performance and the ability to pre-train the representations in an unsupervised manner. The performance gain is, however, dependent on the amount of data available for training. One of the disadvantages comes from the model not being able to handle missing data naturally, although there are ways to alleviate this issue [Ngiam et al. 2011, Wang et al. 2015a]. Finally, deep networks are often difficult to train [Glorot and Bengio 2010], but the field is making progress in better training techniques [Srivastava et al. 2014].

Probabilistic graphical models are another popular way to construct representations through the use of latent random variables [Bengio et al. 2013]. In this section we describe how probabilistic graphical models are used to represent unimodal and multimodal data.

One approach for graphical model-based representation is deep Boltzmann machines (DBM) [Salakhutdinov and Hinton 2009], which stack restricted Boltzmann machines (RBM) [Hinton et al. 2006] as building blocks. Similar to neural networks, each successive layer of a DBM is expected to represent the data at a higher level of abstraction. The appeal of DBMs comes from the fact that they do not need supervised data for training [Salakhutdinov and Hinton 2009]. As they are graphical models, the representation of data is probabilistic. However it is possible to convert them to a deterministic neural network, but this loses the generative aspect of the model [Salakhutdinov and Hinton 2009].

Work by Srivastava and Salakhutdinov [2012a] introduced multimodal deep belief networks as a multimodal representation. Kim et al. [2013] used a deep belief network for each modality and then combined them into joint representation for audiovisual emotion recognition. Huang and Kingsbury [2013] used a similar model for AVSR, and Wu and Shao [2014] for audio and skeleton joint-based gesture recognition.

Multimodal deep belief networks have been extended to multimodal DBMs by Srivastava and Salakhutdinov [2012b]. Multimodal DBMs are capable of learning joint representations from multiple modalities by merging two or more undirected graphs using a binary layer of hidden units on top of them. They allow for the low-level representations of each modality to influence each other after the joint training due to the undirected nature of the model.

Ouyang et al. [2014] explore the use of multimodal DBMs for the task of human pose estimation from multi-view data. They demonstrate that integrating the data at a later stage, after unimodal data underwent nonlinear transformations, was beneficial for the model. Similarly, Suk et al. [2014] use multimodal DBM representation to perform Alzheimer's disease classification from positron emission tomography and magnetic resonance imaging data.

One of the big advantages of using multimodal DBMs for learning multimodal representations is their generative nature, which allows for an easy way to deal with missing data even if a whole modality is missing, the model has a natural way to cope. It can also be used to generate samples of one modality in the presence of the other one, or both modalities from the representation. Similar to autoencoders, the representation can be trained in an unsupervised manner enabling the use of unlabeled data. The major disadvantage of DBMs is the difficulty of training them, the high computational cost, and the need to use approximate variational training methods [Srivastava and Salakhutdinov 2012b].

Sequential Representation. Sequential Representations are designed to be able to represent sequences of varying lengths. This is in contrast with the approaches previously described which are for static data or datasets with fixed length. In this section we describe models that can be used to represent such sequences.

Recurrent neural networks (RNNs), and their variants such as long-short term memory (LSTMs) networks [Hochreiter and Schmidhuber 1997], have recently gained popularity due to their success in sequence modeling across various tasks [Bahdanau et al. 2014, Venugopalan et al. 2015]. So far, RNNs have mostly been used to represent unimodal sequences of words, audio, or images, with most success in the language domain. Similar to traditional neural networks, the hidden state of an RNN can be seen as a representation of the data, i.e., the hidden state of RNN at timestep t can be seen as the summarization of the sequence up to that timestep. This is especially apparent in RNN encoder-decoder frameworks where the task of an encoder is to represent a sequence in the hidden state of an RNN in such a way that a decoder could reconstruct it [Bahdanau et al. 2014].

The use of RNN representations has not been limited to the unimodal domain. An early use of constructing a multimodal representation using RNNs comes from work by Cosi et al. [1994] on AVSR. They have also been used for representing audio-visual data for affect recognition [Nicolaou et al. 2011, Chen et al. 2015] and to represent multi-view data such as different visual cues for human behavior analysis [Rajagopalan et al. 2016].

1.3.2 Coordinated Representations

An alternative to a joint multimodal representation is a coordinated representation. Instead of projecting the modalities together into a joint space, we learn separate representations for each modality but coordinate them through a constraint. We start our discussion with coordinated representations that enforce similarity between representations, moving on to coordinated representations that enforce more structure on the resulting space (representative works of different coordinated representations can be seen in Table 1.2).

Similarity models minimize the distance between modalities in the coordinated space. For example, such models encourage the representation of the word *dog* and an image of a dog to have a smaller distance between them than distance between the word *dog* and an image of a car [Frome et al. 2013]. One of the earliest examples of such a representation comes from the work by Weston et al. [2011] on the WSABIE (web scale annotation by image embedding) model, where a coordinated space was constructed for images and their annotations. WSABIE constructs a simple linear map from image and textual features such that corresponding annotation and image representation would have a higher inner product (smaller cosine distance) between them than non-corresponding ones.

More recently, neural networks have become a popular way to construct coordinated representations, due to their ability to learn representations. Their advantage lies in the fact that they can jointly learn coordinated representations in an end-to-end manner. An example of such coordinated representation is DeViSE, a deep visual-semantic embedding [Frome et al. 2013]. DeViSE uses a similar inner product and ranking loss function to WSABIE but uses more complex image and word embeddings. Kiros et al. [2015] extended this to sentence and image coordinated representation by using an LSTM model and a pairwise ranking loss to coordinate the feature space. Socher et al. [2014] tackle the same task, but extend the language model to a dependency tree RNN to incorporate compositional semantics. A similar model was also proposed by Pan et al. [2016], but using videos instead of images. Xu et al. [2015] also constructed a coordinated space between videos and sentences using a ⟨subject, verb, object⟩ compositional language model and a deep video model. This representation was then used for the task of cross-modal retrieval and video description.

While the above models enforced similarity between representations, **structured coordinated space** models go beyond that and enforce additional constraints between the modality representations. The type of structure enforced is often based on the application, with different constraints for hashing, cross-modal retrieval, and image captioning.

Structured coordinated spaces are commonly used in cross-modal hashing which is compression of high-dimensional data into compact binary codes with similar binary codes for similar objects [Wang et al. 2014]. The idea of cross-modal hashing is to create such codes for cross-modal retrieval [Kumar and Udupa 2011, Bronstein et al. 2010, Jiang et al. 2015]. Hashing enforces certain constraints on the resulting multimodal space: (1) it has to be an N-dimensional Hamming space, a binary representation with controllable number of bits; (2) the same object from different modalities has to have a similar hash code; and (3) the space has to be similarity-preserving. Learning how to represent the data as a hash function attempts to enforce all of these three requirements [Kumar and Udupa 2011, Bronstein et al. 2010]. For example, Jiang and Li [2017] introduced a method to learn such common binary space between sentence descriptions and corresponding images using end-to-end trainable deep learning techniques. While Cao et al. [2016] extended the approach with a more complex LSTM sentence representation and introduced an outlier insensitive bit-wise margin loss and a relevance feedback based semantic similarity constraint. Similarly, Wang et al. [2016] constructed a coordinated space in which images (and sentences) with similar meanings are closer to each other.

Another example of a structured coordinated representation comes from order-embeddings of images and language [Vendrov et al. 2016, Zhang et al. 2016]. The model proposed by Vendrov et al. [2016] enforces a dissimilarity metric that is asymmetric and implements the notion of partial order in the multimodal space. The idea is to capture a partial order of the language and image representations, enforcing a hierarchy on the space; for example, image of "a woman walking her dog" → text "woman walking her dog" → text "woman walking." A similar model using denotation graphs was also proposed by Young et al. [2014] where denotation graphs are used to induce a partial ordering. Lastly, Zhang et al. [2016] present how exploiting structured representations of text and images can create concept taxonomies in an unsupervised manner.

A special case of a structured coordinated space is one based on canonical correlation analysis (CCA) [Hotelling 1936]. CCA computes a linear projection which maximizes the correlation between two random variables (in our case modalities) and enforces orthogonality of the new space. CCA models have been used extensively for cross-modal retrieval [Hardoon et al. 2004, Rasiwasia et al. 2010, Klein et al. 2015] and audiovisual signal analysis [Sargin et al. 2007, Slaney and Covell 2001]. Extensions to CCA attempt to construct a correlation maximizing nonlinear projection [Lai and Fyfe 2000, Andrew et al. 2013]. Kernel canonical correlation analysis (KCCA) [Lai and Fyfe 2000] uses reproducing kernel Hilbert spaces for projection. However, as the approach is nonparametric it scales poorly with the size of the training set and has issues with very large real-world datasets. Deep canonical correlation analysis (DCCA) [Andrew et al. 2013] was introduced as an alternative to KCCA and addresses the scalability issue, it was also shown to lead to better correlated representation space. Similar correspondence autoencoder [Feng et al. 2014] and deep correspondence RBMs [Feng et al. 2015] have also been proposed for cross-modal retrieval.

CCA, KCCA, and DCCA are unsupervised techniques and only optimize the correlation over the representations, thus mostly capturing what is shared across the modalities. Deep canonically correlated autoencoders [Wang et al. 2015b] also include an autoencoder based data reconstruction term. This encourages the representation to also capture modality specific information. Semantic correlation maximization method [Zhang and Li 2014] also encourages semantic relevance, while retaining correlation maximization and orthogonality of the resulting space, leading to a combination of CCA and cross-modal hashing techniques.

1.3.3 Discussion

In this section we identified two major types of multimodal representations: joint and coordinated. Joint representations project multimodal data into a common space and are best suited for situations when all of the modalities are present during inference. They have been extensively used for AVSR, affect, and multimodal gesture recognition. Coordinated representations, on the other hand, project each modality into a separate but coordinated space, making them suitable for applications where only one modality is present at test time, such as: multimodal retrieval and translation, conceptual grounding (Section 1.4.2), and zero shot learning (Section 1.4.2). Finally, while joint representations have been used in situations to construct representations of more than two modalities, coordinated spaces have, so far, been mostly limited to two modalities.

1.4 Co-learning

The final multimodal challenge in our taxonomy is co-learning that aids the modeling of a (resource poor) modality by exploiting knowledge from another (resource rich) modality. It is particularly relevant when one of the modalities has limited resources, lack of annotated data, noisy input, and unreliable labels. We call this challenge co-learning as most often the helper modality is used only during model training and is not used during test time. We identify three types of co-learning approaches based on their training resources: parallel, non-parallel, and hybrid. *Parallel-data* approaches require training datasets where the observations from one modality are directly linked to the observations from other modalities. In other words, when the multimodal observations are from the same instances, such as in an audio-visual speech dataset where the video and speech samples are from the same speaker. In contrast, *non-parallel data* approaches do not require direct links between observations from different modalities. These approaches usually achieve co-learning by using overlap in terms of categories, e.g., in zero shot learning when the conventional visual object recognition dataset is expanded with a second text-only dataset from Wikipedia to improve the generalization of visual object recognition. In the *hybrid* data setting the modalities are *bridged* through a shared modality or a dataset. An overview of the taxonomy in co-learning can be seen in Table 1.3 and summary of data parallelism in Figure 1.2.

1.4.1 Parallel Data

In parallel data co-learning both modalities share a set of instances, audio recordings with the corresponding videos, images, and their sentence descriptions. This allows for two types of algorithms to exploit that data to better model the modalities: co-training and representation learning.

Co-training is the process of creating more labeled training samples when we have few labeled samples in a multimodal problem [Blum and Mitchell 1998]. The basic algorithm builds weak classifiers in each modality to bootstrap each other with labels for the unlabeled data. It has been shown in the seminal work of Blum and Mitchell [1998] that more training samples for web page classification can be discovered on the web page itself and hyper-links leading to it. By definition this task requires parallel data as it relies on the overlap of multimodal samples.

Co-training has been used for statistical parsing [Sarkar 2001] to build better visual detectors [Levin et al. 2003] and for audio-visual speech recognition [Christoudias et al. 2006]. It has also been extended to deal with disagreement between modalities by filtering out unreliable samples [Christoudias et al. 2008].

Table 1.3 **A summary of co-learning taxonomy, based on data parallelism. Parallel data—multiple modalities can see the same instance. Non-parallel data—unimodal instances are independent of each other. Hybrid data—the modalities are *pivoted* through a shared modality or dataset.**

Data parallelism	Task	Reference
Parallel		
Co-training	Mixture	[Blum and Mitchell 1998]
Transfer learning	AVSR	[Ngiam et al. 2011]
	Lip reading	[Moon et al. 2015]
Non-parallel		
Transfer learning	Visual classification	[Frome et al. 2013]
	Action recognition	[Mahasseni and Todorovic 2016]
Concept grounding	Metaphor class.	[Shutova et al. 2016]
	Word similarity	[Kiela and Clark 2015]
Zero shot learning	Image class.	[Socher et al. 2013]
	Thought class.	[Palatucci et al. 2009]
Hybrid Data		
Bridging	MT and image ret.	[Rajendran et al. 2015]
	Transliteration	[Nakov and Ng 2012]

While co-training is a powerful method for generating more labeled data, it can also lead to biased training samples resulting in overfitting.

Transfer learning is another way to exploit co-learning with parallel data. Multimodal representation learning (Section 1.3.1) approaches such as multimodal deep Boltzmann machines [Srivastava and Salakhutdinov 2012b] and multimodal autoencoders [Ngiam et al. 2011] transfer information from representation of one modality to that of another. This not only leads to multimodal representations, but also to better unimodal ones, with only one modality being used during test time.

Moon et al. [2015] show how to transfer information from a speech recognition neural network (based on audio) to a lip-reading one (based on images), leading to a better visual representation, and a model that can be used for lip-reading without need for audio information during test time. Similarly, Arora and Livescu [2013] build better acoustic features using CCA on acoustic and articulatory (location of lips, tongue, and jaw) data. They use articulatory data only during CCA construction and use only the resulting acoustic (unimodal) representation during test time.

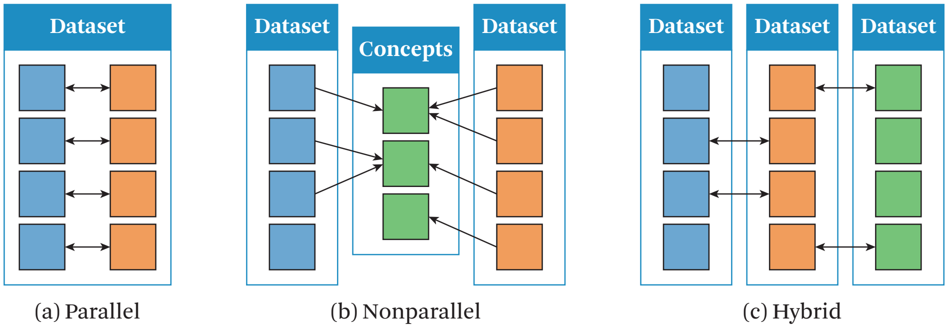


Figure 1.2 Types of data parallelism used in co-learning: *parallel*, modalities are from the same dataset and there is a direct correspondence between instances; *non-parallel*, modalities are from different datasets and do not have overlapping instances, but overlap in general categories or concepts; and *hybrid*, the instances or concepts are bridged by a third modality or a dataset.

1.4.2 Non-parallel Data

Methods that rely on non-parallel data do not require the modalities to have shared instances, but only shared categories or concepts. Non-parallel co-learning approaches can help when learning representations as they allow for better semantic concept understanding and even perform unseen object recognition.

Transfer learning is also possible on non-parallel data and allows for the learning of better representations through transferring information from a representation built using a data rich or clean modality to a data scarce or noisy modality. This type of transfer learning is often achieved by using coordinated multimodal representations (see Section 1.3.2). For example, Frome et al. [2013] used text to improve visual representations for image classification by coordinating CNN visual features with word2vec textual ones [Mikolov et al. 2013] trained on separate large datasets. Visual representations trained in such a way result in more meaningful errors, mistaking objects for ones of similar category [Frome et al. 2013]. Mahasseni and Todorovic [2016] demonstrated how to regularize a color video-based LSTM using an autoencoder LSTM trained on 3D skeleton data by enforcing similarities between their hidden states. Such an approach is able to improve the original LSTM and lead to state-of-the-art performance in action recognition.

Conceptual grounding refers to learning semantic meanings or concepts not purely based on language but also on additional modalities such as vision, sound, or even smell. While the majority of concept learning approaches are purely language-based, representations of meaning in humans are not merely a product of our linguistic exposure but are also *grounded* through our sensorimotor experience and perceptual system [Barsalou 2008, Louwerse 2011]. Human semantic knowledge relies heavily on perceptual information [Louwerse 2011] and many concepts are grounded in the perceptual system and are not purely symbolic [Barsalou 2008]. This implies that learning semantic meaning purely from textual information might not be optimal, and motivates the use of visual or acoustic cues to ground our linguistic representations.

Starting from work by Feng and Lapata [2010], grounding is usually performed by finding a common latent space between the representations [Feng and Lapata 2010, Silberer and Lapata 2012] (in case of parallel datasets) or by learning unimodal representations separately and then concatenating them to lead to a multimodal one [Regneri et al. 2013, Shutova et al. 2016, Kiela and Bottou 2014, Bruni et al. 2012] (in case of non-parallel data). Once a multimodal representation is constructed it can be used on purely linguistic tasks. Shutova et al. [2016] and Bruni et al. [2012] used grounded representations for better classification of metaphors and literal language. Such representations have also been useful for measuring conceptual similarity and relatedness, identifying how semantically or conceptually related two words are [Kiela and Bottou 2014, Bruni et al. 2014, Silberer and Lapata 2012] or actions [Regneri et al. 2013]. Furthermore, concepts can be grounded not only using visual signals, but also acoustic ones, leading to better performance especially on words with auditory associations [Kiela and Clark 2015], or even olfactory signals [Kiela et al. 2015] for words with smell associations. Finally, there is a lot of overlap between multimodal alignment and conceptual grounding, as aligning visual scenes to their descriptions leads to better textual or visual representations [Regneri et al. 2013, Plummer et al. 2015, Kong et al. 2014, Yu and Siskind 2013].

Conceptual grounding has been found to be an effective way to improve performance on a number of tasks. It also shows that language and vision (or audio) are complementary sources of information and combining them in multimodal models often improves performance. However, one has to be careful as grounding does not always lead to better performance [Kiela and Clark 2015, Kiela et al. 2015], and only makes sense when grounding has relevance for the task such as grounding using images for visually related concepts.

Zero shot learning (ZSL) refers to recognizing a concept without having explicitly seen any examples of it. For example, classifying a cat in an image without ever having seen (labeled) images of cats. This is an important problem to address as in a number of tasks such as visual object classification: it is prohibitively expensive to provide training examples for every imaginable object of interest.

There are two main types of ZSL: unimodal and multimodal. The unimodal ZSL looks at component parts or attributes of the object, such as phonemes to recognize an unheard word or visual attributes such as color, size, and shape to predict an unseen visual class [Farhadi et al. 2009]. The multimodal ZSL recognizes the objects in the primary modality through the help of the secondary one—in which the object has been seen. The multimodal version of ZSL is a problem facing non-parallel data by definition as the overlap of seen classes is different between the modalities.

Socher et al. [2013] map image features to a conceptual word space and are able to classify between seen and unseen concepts. The unseen concepts can be then assigned to a word that is close to the visual representation; this is enabled by the semantic space being trained on a separate dataset that has seen more concepts. Instead of learning a mapping from visual to concept space Frome et al. [2013] learn a coordinated multimodal representation between concepts and images that allows for ZSL. Palatucci et al. [2009] perform prediction of words people are thinking of based on functional magnetic resonance images; they show how it is possible to predict unseen words through the use of an intermediate semantic space. Lazaridou et al. [2014] present a fast mapping method for ZSL by mapping extracted visual feature vectors to text-based vectors through a neural network.

1.4.3 Hybrid Data

In the hybrid data setting two non-parallel modalities are bridged by a shared modality or a dataset (see Figure 1.2c). The most notable example is the Bridge Correlational Neural Network [Rajendran et al. 2015], which uses a pivot modality to learn coordinated multimodal representations in presence of non-parallel data. For example, in the case of multilingual image captioning, the image modality would always be paired with at least one caption in any language. Such methods have also been used to bridge languages that might not have parallel corpora but have access to a shared pivot language, such as for machine translation [Rajendran et al. 2015, Nakov and Ng 2012] and document transliteration [Khapra et al. 2010].

Instead of using a separate modality for bridging, some methods rely on existence of large datasets from a similar or related task to lead to better performance in a task that only contains limited annotated data. Socher and Fei-Fei [2010] use the existence of large text corpora in order to guide image segmentation. While Anne Hendricks et al. [2016] use separately trained visual model and a language model to lead to a better image and video description system, for which only limited data is available.

1.4.4 Discussion

Multimodal co-learning allows for one modality to influence the training of another, exploiting the complementary information across modalities. It is important to note that co-learning is task independent and could be used to create better fusion, translation, and alignment models. This challenge is exemplified by algorithms such as co-training, multimodal representation learning, conceptual grounding, and zero shot learning (ZSL) and has found many applications in visual classification, action recognition, audio-visual speech recognition, and semantic similarity estimation.

1.5 Conclusion

Multimodal machine learning is a vibrant multi-disciplinary field which aims to build models that can process and relate information from multiple modalities. As part of this chapter, presented the taxonomy of two challenges in multimodal machine learning: representation and co-learning [Baltrušaitis et al. 2017]. Some of them such as fusion have been studied for a long time, but more recent interest in alignment and translation have led to a large number of new multimodal algorithms and exciting multimodal applications.

Although the focus of this chapter was primarily on the last decade of multimodal research, it is important to address future challenges with a knowledge of past achievements. Moving forward, the proposed taxonomy gives researchers a framework to understand current research and identify understudied challenges for future research. We believe that all these aspects of multimodal research are needed if we want to build computers able to perceive, model and generate multimodal signals. One specific area of multimodal machine learning which seems to be under-studied is co-learning, where knowledge from one modality helps with modeling in another modality. This challenge is related to the concept of coordinated representations where each modality keeps its own representation but find a way to exchange and coordinate knowledge. We see these lines of research as promising directions for future research.

Focus Questions

1.1. Describe the different applications that have emerged in the multimodal domain. Also think of and list a few applications not listed in the chapter. Do these applications suggest that models using multiple modalities are essential when contrasted with unimodal models?

1.2. What are the 2 categories of Representation Learning that have been discussed in this chapter? Which of the 2 categories seems to be more explicit in estimating similarities between two modalities and how did you come to that conclusion?

1.3. Coordinated Representation methods have only been used for two modalities at a time. Can you come up with ideas to extend the current methods to multiple modalities?

1.4. Compare and contrast probabilistic graphical models and neural networks for representation learning. Can you think of a way to combine the best of both worlds for multimodal representations?

1.5. Limited and noisy data from a modality is a common problem in a hoards of real world tasks. Co-learning exploits knowledge from a resource rich modality to aid the resource poor modality. Can you think of conditions where the described methods could do more harm than good?

1.6. Transfer Learning could be done on both Parallel and Non-parallel data. What are the key differences between the approaches followed on both these kinds of data?

1.7. Although the chapter focuses more on visual, acoustic, and textual modalities there are other modalities (like olfactory signals) that can act as a bridge to ground associations made by humans. What are some ways that can help one decide which modalities are complimentary (help each other and boost performance on the task)?

1.8. Zero shot learning recognizes a concept without ever having explicitly seen it before. Which kind of representation is a good choice for this task and why?

1.9. Hybrid data is similar to Non-Parallel data apart form one key difference. What is that, and how are the models modified to take advantage of this kind of data?

References

D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pp. 173–182. 23, 25

C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos. 2015. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177. DOI: 10.1007/s10462-012-9368-5. 23