11

Selected Applications in Multimodal and Multi-task Learning

Multi-task learning is a machine learning approach that learns to solve several related problems at the same time, using a shared representation. It can be regarded as one of the two major classes of transfer learning or learning with knowledge transfer, which focuses on generalizations across distributions, domains, or tasks. The other major class of transfer learning is adaptive learning, where knowledge transfer is carried out in a sequential manner, typically from a source task to a target task [95]. Multi-modal learning is a closely related concept to multi-task learning, where the learning domains or "tasks" cut across several modalities for human–computer interactions or other applications embracing a mixture of textual, audio/speech, touch, and visual information sources.

The essence of deep learning is to automate the process of discovering effective features or representations for any machine learning task, including automatically transferring knowledge from one task to another concurrently. Multi-task learning is often applied to conditions where no or very little training data are available for the target task domain, and hence is sometimes called zero-shot or one-shot learning. It is evident that difficult multi-task leaning naturally fits the paradigm of deep learning or representation learning where the shared

representations and statistical strengths across tasks (e.g., those involving separate modalities of audio, image, touch, and text) is expected to greatly facilitate many machine learning scenarios under low- or zero-resource conditions. Before deep learning methods were adopted, there had been numerous efforts in multi-modal and multi-task learning. For example, a prototype called MiPad for multi-modal interactions involving capturing, leaning, coordinating, and rendering a mix of speech, touch, and visual information was developed and reported in [175, 103]. And in [354, 443], mixed sources of information from multiple-sensory microphones with separate bone-conductive and airborn paths were exploited to de-noise speech. These early studies all used shallow models and learning methods and achieved worse than desired performance. With the advent of deep learning, it is hopeful that the difficult multi-modal learning problems can be solved with eventual success to enable a wide range of practical applications. In this chapter, we will review selected applications in this area, organized according to different combinations of more than one modalities or learning tasks. Much of the work reviewed here is on-going research, and readers should expect follow-up publications in the future.

11.1 Multi-modalities: Text and image

The underlying mechanism for potential effectiveness of multi-modal learning involving text and image is the common semantics associated with the text and image. The relationship between the text and image may come, for example, from the text annotations of an image (as the training data for a multi-modal learning system). If the related text and image share the same representation in a common semantic space, the system can generalize to the unseen situation where either text or image is unavailable. It can thus be naturally used for zero-shot learning for image or text. In other words, multi-modality learning can use text information to help image/visual recognition, and vice versa. Exploiting text information for image/visual recognition constitutes most of the work done in this space, which we review in this section below.

The deep architecture, called DeViSE (deep visual-semantic embedding) and developed by Frome et al. [117], is a typical example of the multi-modal learning where text information is used to improve the image recognition system, especially for performing zero-shot learning. Image recognition systems are often limited in their ability to scale to large number of object categories, due in part to the increasing difficulty of acquiring sufficient training data with text labels as the number of image categories grows. The multi-modal DeViSE system is aimed to leverage text data to train the image models. The joint model is trained to identify image classes using both labeled image data and the semantic information learned from unannotated text. An illustration of the DeViSE architecture is shown in the center portion of Figure 10.1. It is initialized with the parameters pre-trained at the lower layers of two models: the deep-CNN for image classification in the left portion of the figure and the text embedding model in the right portion of the figure. The part of the deep CNN, labeled "core visual model" in Figure 10.1, is further learned to predict the target word-embedding vector using a projection layer labeled "transformation" and using a similarity metric. The loss function used in training adopts a combination of dot-product similarity and max-margin, hinge rank loss. The former is the un-normalized version of the cosine loss function used for training the DSSM model in [170] as described in Section 9.3. The latter is similar to the earlier joint image-text model called WSABIE (web scale annotation by image embedding developed by Weston et al. [388, 389]. The results show that the information provided by text improves zero-shot image predictions, achieving good hit rates (close to 15%) across thousands of the labels never seen by the image model.

The earlier WSABIE system as described in [388, 389] adopted a shallow architecture and trained a joint embedding model of both images and labels. Rather than using deep architectures to derive the highly nonlinear image (as well as text-embedding) feature vectors as in DeViSE, the WSABIE uses simple image features and a linear mapping to arrive at the joint embedding space. Further, it uses an embedding vector for each possible label. Thus, unlike DeViSE, WSABIE could not generalize to new classes.

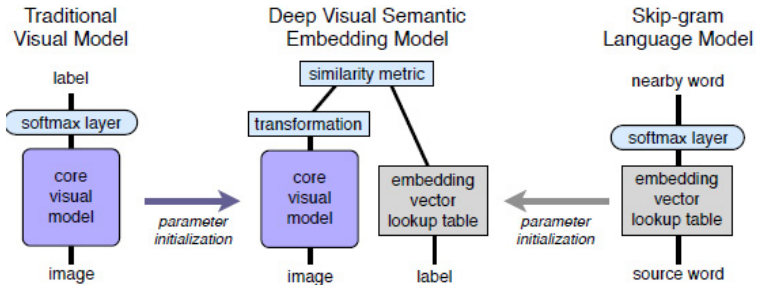

Figure 11.1: Illustration of the multi-modal DeViSE architecture. The left portion is an image recognition neural network with a softmax output layer. The right portion is a skip-gram text model providing word embedding vectors; see Section 8.2 and Figure 8.3 for details. The center is the joint deep image-text model of DeViSE, with the two Siamese branches initialized by the image and word embedding models below the softmax layers. The layer labeled "transformation" is responsible for mapping the outputs of the image (left) and text (right) branches into the same semantic space. [after [117], @NIPS].

It is also interesting to compare the DeViSE architecture of Figure 11.1 with the DSSM architecture of Figure 9.2 in Section 9. The branches of "Query" and "Documents" in DSSM are analogous to the branches of "image" and "text-label" in DeViSE. Both DeViSE and DSSM use the objective function related to cosine distance between two vectors for training the network weights in an end-to-end fashion. One key difference, however, is that the two sets of inputs to the DSSM are both text (i.e., "Query" and "Documents" designed for IR), and thus mapping "Query" and "Documents" to the same semantic space is conceptually more straightforward compared with the need in DeViSE for mapping from one modality (image) to another (text). Another key difference is that the generalization ability of DeViSE to unseen image classes comes from computing text embedding vectors for many unsupervised text sources (i.e., with no image counterparts) that would cover the text labels corresponding to the unseen classes. The generalization ability of the DSSM over unseen words, however, is derived from a special coding scheme for words in terms of their constituent letters.

The DeViSE architecture has inspired a more recent method, which maps images into the semantic embedding space via convex

combination of embedding vectors for the text label and the image classes [270]. Here is the main difference. DeViSE replaces the last, softmax layer of a CNN image classifier with a linear transformation layer. The new transformation layer is then trained together with the lower layers of the CNN. The method in [270] is much simpler — keeping the softmax layer of the CNN while not training the CNN. For a test image, the CNN first produces top N-best candidates. Then, the convex combination of the corresponding N embedding vectors in the semantic space is computed. This gives a deterministic transformation from the outputs of the softmax classifier into the embedding space. This simple multi-modal learning method is shown to work very well on the ImageNet zero-shot learning task.

Another thread of studies separate from but related to the above work on multi-modal learning involving text and image have centered on the use of multi-modal embeddings, where data from multiple sources with separate modalities of text and image are projected into the same vector space. For example, Socher and Fei-Fei [341] project words and images into the same space using kernelized canonical correlation analysis. Socher et al. [342] map images to single-word vectors so that the constructed multi-modal system can classify images without seeing any examples of the class, i.e., zero-shot learning similar to the capability of DeViSE. The most recent work by Socher et al. [343] extends their earlier work from single-word embeddings to those of phrases and full-length sentences. The mechanism for mapping sentences instead of the earlier single words into the multi-modal embedding space is derived from the power of the recursive neural network described in Socher et al. [347] as summarized in Section 8.2, and its extension with dependency tree.

In addition to mapping text to image (or vice versa) into the same vector space or to creating the joint image/text embedding space, multi-modal learning for text and image can also be cast in the framework of language models. In [196], a model of natural language is made conditioned on other modalities such as image as the focus of the study. This type of multi-modal language model is used to (1) retrieve images given complex description queries, (2) retrieve phrase descriptions given image queries, and (3) generate text conditioned on images.

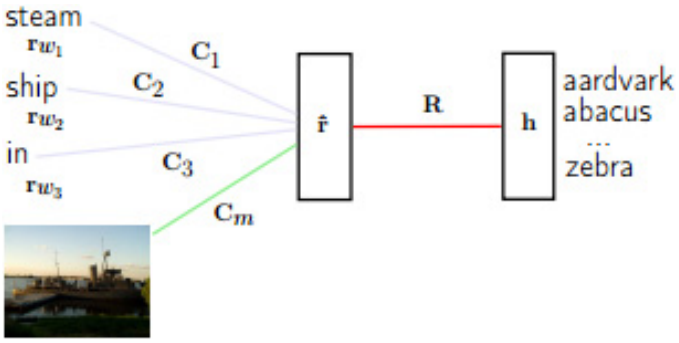

Figure 11.2: Illustration of the multi-modal DeViSE architecture. The left portion is an image recognition neural network with a softmax output layer. The right portion is a skip-gram text model providing word embedding vectors; see Section 8.2 and Figure 8.3 for details. The center is the joint deep image-text model of DeViSE, with the two Siamese branches initialized by the image and word embedding models below the softmax layers. The layer labeled "transformation" is responsible for mapping the outputs of the image (left) and text (right) branches into the same semantic space. [after [196], @NIPS].

Word representations and image features are jointly learned by training the multi-modal language model together with a convolutional network. An illustration of the multi-modal language model is shown in Figure 11.2.

11.2 Multi-modalities: Speech and image

Ngiam et al. [268, 269] propose and evaluate an application of deep networks to learn features over audio/speech and image/video modalities. They demonstrate cross-modality feature learning, where better features for one modality (e.g., image) is learned when multiple modalities (e.g., speech and image) are present at feature learning time. A bi-modal deep autoencoder architecture for separate audio/speech and video/image input channels are shown in Figure 11.3. The essence of this architecture is to use a shared, middle layer to represent both types of modalities. This is a straightforward generalization from the single-modal deep autoencoder for speech shown in Figure 4.1 of Section 4 to bi-modal counterpart. The authors further show how to

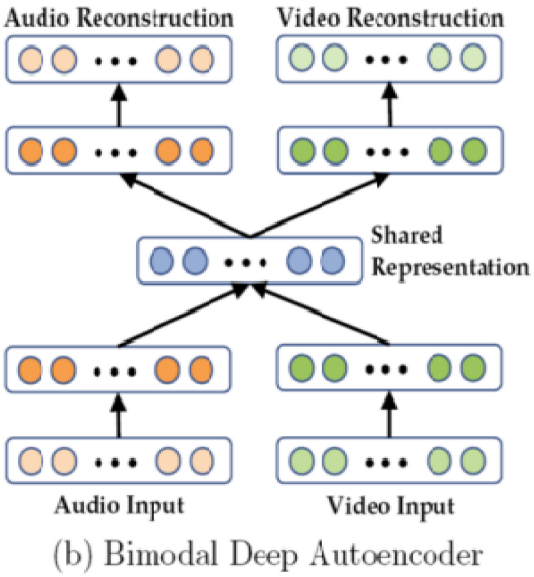

Figure 11.3: The architecture of a deep denoising autoencoder for multi-modal audio/speech and visual features. [after [269], @ICML].

learn a shared audio and video representation, and evaluate it on a fixed task, where the classifier is trained with audio-only data but tested with video-only data and vice versa. The work concludes that deep learning architectures are generally effective in learning multi-modal features from unlabeled data and in improving single modality features through cross modality information transfer. One exception is the cross-modality setting using the CUAVE dataset. The results presented in [269, 268] show that learning video features with both video and audio outperforms that with only video data. However, the same paper also shows that a model of [278] in which a sophisticated signal processing technique for extracting visual features, together with the uncertainty-compensation method developed originally from robust speech recognition [104], gives the best classification accuracy in the cross-modal learning task, beating the features derived from the generative deep architecture designed for this task.

While the deep generative architecture for multimodal learning described in [268, 269] is based on non-probabilistic autoencoder neural

nets, a probabilistic version based on deep Boltzmann machine (DBM) has appeared more recently for the same multimodal application. In [348], a DBM is used to extract a unified representation integrating separate modalities, useful for both classification and information retrieval tasks. Rather than using the "bottleneck" layers in the deep autoencoder to represent multimodal inputs, here a probability density is defined on the joint space of multimodal inputs, and states of suitably defined latent variables are used for the representation. The advantage of this probabilistic formulation, possibly lacking in the traditional deep autoencoder, is that the missing modality's information can be filled in naturally by sampling from its conditional distribution. More recent work on autoencoders [22, 30] shows the capability of generalized denoising autoencoders in carrying out sampling, thus they may overcome the earlier problem of filling-in the missing modality's information. For the bi-modal data consisting of image and text, the multimodal DBM was shown to slightly outperform the traditional version of the deep multimodal autoencoder as well as multimodal DBN in classification and information retrieval tasks. No results on the comparisons with the generalized version of deep autoencoders has been reported but may appear soon.

The several architectures discussed so far in this chapter for multimodal processing and learning can be regarded as special cases of more general multi-task learning and transfer learning [22, 47]. Transfer learning, encompassing both adaptive and multi-task learning, refers to the ability of a learning architecture and technique to exploit common hidden explanatory factors among different learning tasks. Such exploitation permits sharing of aspects of diverse types of input data sets, thus allowing the possibility of transferring knowledge across seemingly different learning tasks. As argued in [22], the learning architecture shown in Figure 11.4 and the associated learning algorithms have an advantage for such tasks because they learn representations that capture underlying factors, a subset of which may be relevant for each particular task. We will discuss a number of such multi-task learning applications in the remainder of this chapter that are confined with a single modality of speech, natural language processing, *or* image domain.

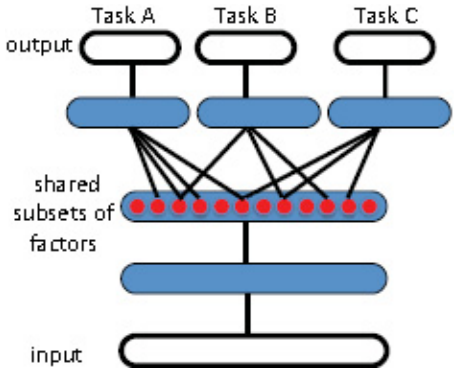

Figure 11.4: A DNN architecture for multitask learning that is aimed to discover hidden explanatory factors shared among three tasks A, B, and C. [after [22], @IEEE].

11.3 Multi-task learning within the speech, NLP or image domain

Within the speech domain, one most interesting application of multi-task learning is multi-lingual or cross-lingual speech recognition, where speech recognition for different languages is considered as different tasks. Various approaches have been taken to attack this rather challenging acoustic modeling problem for speech recognition, where the difficulty lies in the lack of transcribed speech data due to economic considerations in developing speech recognition systems for all languages in the world. Cross-language data sharing and data weighing are common and useful approaches for the GMM–HMM system [225]. Another successful approach for the GMM–HMM is to map pronunciation units across languages either via knowledge-based or data-driven methods [420]. But they are much inferior to the DNN–HMM approach which we now summarize.

In recent papers of [94, 170] and [150], two research groups independently developed closely related DNN architectures with multi-task learning capabilities for multilingual speech recognition. See Figure 11.5 for an illustration of this type of architecture. The idea behind these architectures is that the hidden layers in the DNN, when learned

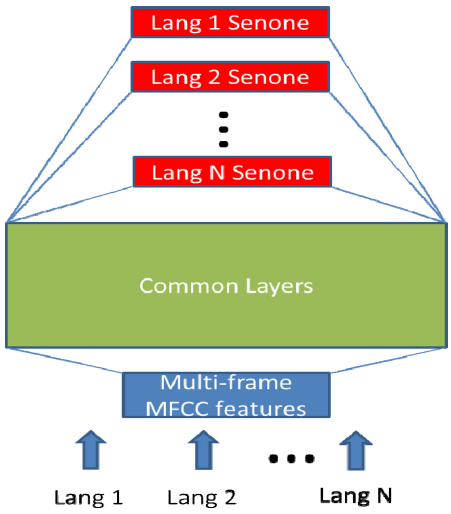

Figure 11.5: A DNN architecture for multilingual speech recognition. [after [170], @IEEE].

appropriately, serve as increasingly complex feature transformations sharing common hidden factors across the acoustic data in different languages. The final softmax layer representing a log-linear classifier makes use of the most abstract feature vectors represented in the top-most hidden layer. While the log-linear classifier is necessarily separate for different languages, the feature transformations can be shared across languages. Excellent multilingual speech recognition results are reported, far exceeding the earlier results using the GMM–HMM based approaches [225, 420]. The implication of this set of work is significant and far reaching. It points to the possibility of quickly building a high-performance DNN-based system for a new language from an existing multilingual DNN. This huge benefit would require only a small amount of training data from the target language, although having more data would further improve the performance. This multitask learning approach can reduce the need for the unsupervised pre-training stage, and can train the DNN with much fewer epochs. Extension of this set of work would be to efficiently build a language-universal speech recognition system. Such a system cannot only recognize many languages and improve the accuracy for each individual language, but

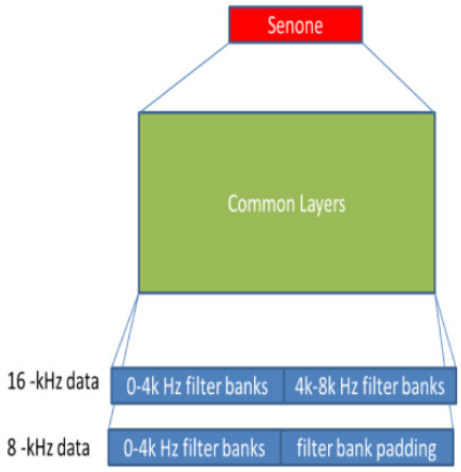

Figure 11.6: A DNN architecture for speech recognition trained with mixed-bandwidth acoustic data with 16-kHz and 8-kHz sampling rates; [after [221], @IEEE].

also expand the languages supported by simply stacking softmax layers on the DNN for new languages.

A closely related DNN architecture, as shown in Figure 11.6, with multitask learning capabilities was also recently applied to another acoustic modeling problem — learning joint representations for two separate sets of acoustic data [94, 221]. The set that consists of the speech data with 16 kHz sampling rate is of wideband and high quality, which is often collected from increasingly popular smart phones under the voice search scenario. Another, narrowband data set has a lower sampling rate of 8kHz, often collected using the telephony speech recognition systems.

As a final example of multi-task learning within the speech domain, let us consider phone recognition and word recognition as separate "tasks." That is, phone recognition results are used not for producing text outputs but for language-type identification or for spoken document retrieval. Then, the use of pronunciation dictionary in almost all speech systems can be considered as multi-task learning that share the tasks of phone recognition and word recognition. More advanced frameworks in speech recognition have pushed this direction further

by advocating the use of even finer units of speech than phones to bridge the raw acoustic information of speech to semantic content of speech via a hierarchy of linguistic structure. These atomic speech units include "speech attributes" in the detection-based and knowledge-rich modeling framework for speech recognition, whose accuracy has been significantly boosted recently by the use of deep learning methods [332, 330, 427].

Within the natural language processing domain, the best known example of multi-task learning is the comprehensive studies reported in [62, 63], where a range of separate "tasks" of part-of-speech tagging, chunking, named entity tagging, semantic role identification, and similar-word identification in natural language processing are attacked using a common representation of words and a unified deep learning approach. A summary of these studies can be found in Section 8.2.

Finally, within the domain of image/vision as a single modality, deep learning has also been found effective in multi-task learning. Srivastava and Salakhutdinov [349] present a multi-task learning approach based on hierarchical Bayesian priors in a DNN system applied to various image classification data sets. The priors are combined with a DNN, which improves discriminative learning by encouraging information sharing among tasks and by discovering similar classes among which knowledge is transferred. More specifically, methods are developed to jointly learn to classify images and a hierarchy of classes, such that "poor classes," for which there are relatively few training examples, can benefit from similar "rich classes," for which more training examples are available. This work can be considered as an excellent instance of learning output representations, in addition to learning input representation of the DNN as the focus of nearly all deep learning work reported in the literature.

As another example of multi-task learning within the single-modality domain of image, Ciresan et al. [58] applied the architecture of deep CNNs to character recognition tasks for Latin and for Chinese. The deep CNNs trained on Chinese characters are shown to be easily capable of recognizing uppercase Latin letters. Further, learning Chinese characters is accelerated by first pre-training a CNN on a small subset of all classes and then continuing to train on all classes.