Transfer Learning for Speech and Language Processing

Dong Wang and Thomas Fang Zheng
1. Center for Speech and Language Technologies (CSLT)
Research Institute of Information Technology, Tsinghua University
2. Tsinghua National Lab for Information Science and Technology
Beijing, 100084, P.R.China

***Abstract*—Transfer learning is a vital technique that generalizes models trained for one setting or task to other settings or tasks. For example in speech recognition, an acoustic model trained for one language can be used to recognize speech in another language, with little or no re-training data. Transfer learning is closely related to multi-task learning (cross-lingual vs. multilingual), and is traditionally studied in the name of 'model adaptation'. Recent advance in deep learning shows that transfer learning becomes much easier and more effective with high-level abstract features learned by deep models, and the 'transfer' can be conducted not only between data distributions and data types, but also between model structures (e.g., shallow nets and deep nets) or even model types (e.g., Bayesian models and neural models). This review paper summarizes some recent prominent research towards this direction, particularly for speech and language processing. We also report some results from our group and highlight the potential of this very interesting research field[1].**

I. Introduction

Machine learning (ML) techniques have been extensively exploited in modern speech and language processing research [1], [2], [3]. Among the rich family of ML models and algorithms, transfer learning is among the most interesting. Generally speaking, transfer learning involves all methods that utilize any auxiliary resources (data, model, labels, etc.) to enhance model learning for the target task [4], [5], [6], [7]. This is very important for speech and language research, since human speech and languages are so diverse and imbalanced. There are more than $5,000$ languages around the world, and the number is even bigger if dialects are counted. Among this big family, 389 languages (nearly 6%) account for 94% of the word's population, and the rest thousands languages are spoken by very few people.[2] Even for the 389 'big' languages, only very few possess adequate resources (speech signal, text corpus, lexicon, phonetic/syntactic regulations, etc.) for speech and language research. If we talk about 'rich-resource' languages, perhaps only English is in that category. Additionally, resources in different domains are also highly imbalanced, even for English. This means that almost all research in speech and language confront the challenge of data sparsity. More seriously, human language is such dynamic that new words and domains emerge every day, and so no models learned at a particular time will remain valid forever.

With such diversity, variation, imbalance and dynamics, it is almost impossible for speech and language researchers to learn a model from one single data resource and then put it on the shelf. We have to resort to some more smart algorithms that can learn from multiple languages, multiple data, multiple domains and keep the model adapted. On the other hand, it would not be very controversial to argue that human speech and languages hold some common statistical patterns at both the signal and symbolic levels, so that learning from multiple resources is possible.

In fact, transfer learning has been studied for a long time in a multitude of research fields in speech and language processing, e.g., speaker adaptation and multilingual modeling in speech recognition, cross-language document classification and sentiment analysis. Most of the studies, however, are task-driven in their own research fields and seldom hold deep understanding about the position of their research in the whole picture of transfer learning. This prevents researchers from answering some important questions: how and in which conditions their methods work, what are possible alternatives of their methods, and what advantages can be achieved with different alternatives? In this paper, we will give a brief summary of the most promising transfer learning methods, particularly within the modern deep learning paradigm. Special focus will be put on the application of transfer learning in speech and language processing, and some recent results from our research team will be presented.

We highlight that it is not our goal to present an entire list of the transfer learning methods in this paper. Instead, the focus is put on the most promising approaches for speech and language processing. Even with such a constraint, the work on transfer learning is still too much to be enumerated, and we can only touch a small part of the plenty techniques. We decide to focus on two specific domains: speech recognition and document classification, particularly the most recent advances based on deep learning which is most relevant to our research. For more detailed surveys on transfer learning in broad research fields, readers are referred to the nice review articles from Pan, Taylor, Bengio and Lu [4], [5], [6], [7] and the references therein.

The paper is organized as follows: Section II gives a quick review of the transfer learning approach, and Section III and Section IV discuss application of transfer learning in speech processing and language processing respectively. The paper is concluded in Section V, with some discussions for the future research directions in this very promising field.

[1] This survey will be continuously updated online (http://arxiv.org/abs/1511.06066) to reflect the recent progress on transfer learning.

[2] https://www.ethnologue.com/statistics

978-988-14768-0-7©2015 APSIPA

II. Transfer learning: A quick review

The motivation of transfer learning can be found in the idea of "Learning to Learn", which stats that learning from scratch (tabula rasa learning) is often limited, and so past experience should be used as much as possible [8]. For instance, once we learned that a hard apple is often sour, this experience can be used when we select pears: we conjecture that hard pears are also sour. This idea and associated research trace back to 20 years ago and were summarized in the NIPS 95 workshop on 'Learning to Learn: Knowledge Consolidation and Transfer in Inductive Systems' [9]. Many ideas and research goals raised in that workshop last two decades and influence our research till today, though the data, models, algorithms, computing power have dramatically changed. Some of the recent developments were discussed in several workshops, e.g., the ICML 2011 workshop on unsupervised and transfer learning[3]; the NIPS 2013 workshop on new directions in transfer and multitask[4]; the ICDM 2015 workshop on practical transfer learning[5]. In this section, we review some of the most prominent approaches to transfer learning, particularly those have been applied to or are potential for speech and language processing.

A. Categories of transfer learning

The initial idea of transfer learning is to reuse the experience/knowledge obtained already to enhance learning for new things. Depending on the relation of the 'old things' (source) that we have learned and the 'new things' (target) that we want to learn, a large amount of methods have been devised, in different names by different authors. A short list of these names include multitask learning, lifelong learning, knowledge transfer, knowledge consolidation, model adaptation, concept drift, covariance shift, etc. Different researchers hold different views for the categorization of these methods. For example, Pan and Yang [4] believed transfer learning should really 'transfer' something so multitask learning should be regarded as a different approach, while Bengio [6] treated transfer learning and multitask learning as synonyms.

In our opinion, the different learning methods mentioned above can be regarded as particular implementations of transfer learning applied in different conditions or by different ways. For example, model adaptation is applied to conditions where the data distributions of the source and target domains are clearly different, while covariance drift is applied to conditions where the distribution changes gradually. As another example, knowledge transfer is applied to the condition where the source model and target model are trained sequentially, while multi-task learning is applied to the condition where the source and target models are trained simultaneously. No matter what forms and properties the learning methods hold, what they all have in common is 'the attempt to transfer knowledge from other sources to benefit the current inductive task', and the benefit of the transfer involves faster convergence, more robust models and less data sensitivity.

We can thus categorize transfer learning into several classes according to the conditions that they apply to. Following the taxonomy in [4], we use *data* and *task* as two conditional factors of transfer learning. For the data condition, it involves the feature space $\mathcal{X}$ (e.g., audio or text) and the distribution $P(X)$ of the feature (e.g., financial news and scientific papers); for the task condition, it involves the label space $\mathcal{Y}$ (e.g., speech phones or speaker identity) and the model $M(x)$ (e.g., probabilistic models or neural models). Any of the two components of the two conditional factors can be the same or different for the learning in the source and target domains, and their relation is shown in Fig. 1. Note that if the feature space is different for the source and target domains, then their distributions are certainly different. Similarly, if the labels are different, then the models are regarded as different, although models from the same family might be used in the source and target domains.

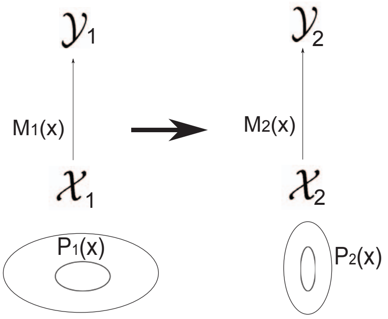

Fig. 1. Relation of the conditional factors in the transfer learning paradigm. $\mathcal{X}_1$ and $\mathcal{Y}_1$ are the feature and label spaces respectively for the learning task in the source domain, and $\mathcal{X}_2$ and $\mathcal{Y}_2$ are for the learning task in the target domain. $M_1(x)$ and $M_2(x)$ represent the models in the source and target domains, respectively.

According to whether the conditional factors (data and task) of the learning in the source and target domains are different or not, transfer learning methods can be categorized into several classes. Table I shows some of the most popular transfer learning approaches that are applicable in different conditions. In the table, '+' means the corresponding conditional factor is the same for the source and target domains, while '-' means different. Note that transfer learning is such a large research field and it is impossible to classify all the methods in such a simple way. For example, an important factor that discriminates different learning methods is whether or not the data in the source and target domains are labelled, which is not clearly reflected in the table (though we will discuss the related issue in the next section). Anyway, Table I gives a rough picture how big the family of transfer learning methods and how they can be categorized according to the conditional factors.

B. Transfer learning methods

We give a short description of the learning methods appearing in Table I. For each method, only the general idea is presented, and application of these methods to speech and language processing is left to the next sections.

1) Model adaptation and incremental training: The simplest transfer learning is to adapt an existing model to meet the change of data distribution. Both the feature and label spaces are the same for the source and target domains, and the models are the same. There are various approaches for

[3] http://clopinet.com/isabelle/Projects/ICML2011/

[4] https://sites.google.com/site/learningacross/

[5] https://sites.google.com/site/icdmwptl2015/home

TABLE I
Categories of transfer learning

		$\mathcal{Y}+$		$\mathcal{Y}-$
		M(x)+	M(x) -	
$\mathcal{X}+$	P(X)+	Conventional ML	Model transfer[10]	Multitask learning[11]
	P(X)-	Model Adaptation[12], [13], incremental learning[14]		
$\mathcal{X}-$			Co-training[15] Heterogeneous transfer learning[16], [17]	Analogy learning [18]

model adaptation. For example, the maximum *a posterior* (MAP) [12] estimation and the maximum likelihood linear regression (MLLR) algorithm [13]. If the distribution changes gradually, then incremental or online learning is often used, e.g. [14], [19], [20].

Note that the adaptation can be either supervised or unsupervised. In the supervised learning, the data in the target domain are labelled, while in the unsupervised learning, no labels are available and they have to be generated by the model in the source domain before the adaptation can be performed. The latter case is often referred to as semi-supervised learning [21]. Note that semi-supervised learning is a general framework to deal with unlabelled data, and can be applied to any conditions where the label spaces are the same in the source and target domains. We will come back to this method in heterogeneous transfer learning that will be discussed shortly. Another approach to dealing with unlabelled data is to use them to derive new features (e.g., by linear projection) where the distributions of the data in the source domain and the target domain are close to each other. An interesting work towards this direction is the approach based on transfer component analysis (TCA) [22].

In another configuration, some unlabelled data are available but the distribution is different from that of the target domain. These data cannot be used for adaptation (either by semi-supervised learning or TCA) otherwise the model will be adapted to a biased condition. However, it can be used to assist deriving more robust features. The idea is similar to TCA, but the unlabelled data are not used as supervision about the target domain, instead as an auxiliary information to derive more domain-independent features. This approach is often referred to as self-taught learning [23], and it essentially holds the same idea as the more recent deep representation learning that will be discussed in Section II-C.

2) Heterogeneous transfer learning: A more complex transfer learning scenario is to keep the labels and model unchanged, however the features are different in the source and target domains. The transfer learning in this scenario is often called heterogeneous transfer learning. The basic assumption for heterogeneous transfer learning is that some correspondence between the source and target domains exist, and this correspondence can be used to transfer knowledge in one domain to another. For example, speech and text are two domains, and there is clear correspondence between the two domains based on human concepts: no matter we speak or write ‘chicken’, it is clear that we refer to the same bird that has wings but can not fly much.

The early research tried to define and utilize the correspondence between the instances of the source and target domains. For example, [24] employed an oracle word translator to define some pivot words that were used to establish the cross-domain correspondence by learning multiple linear classifiers that predict the ‘joint existence’ of these words in the multi-domain data. In [25] some instance-level co-occurrence data were used to estimate the correspondence in the form of joint or conditional probabilities; this correspondence was then used to improve the model in the target domain by risk-minimization inference. Asymmetric regularized cross-domain transformation was proposed in [26], which tries to learn a non-linear transform between the source and target domains by class-labeled instances from both source and target domains. Although an instance does not necessarily possess features of both domains, the class labels offer the correspondence information.

More recent approaches prefer to finding common representations of the source and target domains, for example by matrix factorization [17], RBM-based latent factor learning [27], or joint transfer optimization [28], [16], [29]. More recently, deep learning and heterogeneous transfer learning are combined where high-level features are derived by deep learning and inter-domain transforms are learned by transfer learning [30].

We emphasize that most of the approaches discussed above assume that the label space does not change when transferring from the source domain to the target domain. A more ambitious task is to learn from very different tasks for which the label space is different from the target domain. For example, the task in the source domain is to classify document sentiment, while in the target domain the task is to classify image aesthetic value. This two tasks are fundamentally different, however some analogy does exist between them. Learning correspondence between two independent but analogous domains is easy for humans [31], [32], [33], however it is very difficult for machines. There has been long-term interest in analogy learning among artificial intelligence researchers, e.g., [34], [18], though not too much achievement yet. Interestingly, the recent improvement in deep learning methods seems provide more hope in this direction, by a unified framework for representation learning and multitask learning. This will be discussed in Section II-C.

3) Multiview co-training: A special case of heterogeneous transfer learning is the multi-view co-training, which assumes that each training instance involves features of both the source and target domains, but only the feature in the target domain is available at runtime. In this condition, heterogeneous transfer learning is not very effective since the training instances in the source domain are the same as the instances in the target domain and so does not provide much additional information, at least with supervised learning. However, the multi-view property of the training data indeed can be used to improve unsupervised learning with unlabelled data, by the approach called co-training [15]. Specifically, co-training trains two separate models with features of the source and target domains respectively, and then generates labels for the unlabelled data using one model, which are in turn used to update the other model. This process iterates until convergence is obtained. It is well-known that co-training leads to better models than training with the feature of the target domain only.

4) Model transfer: If the feature and label spaces are the same however the models are different for the source and target domains, the knowledge learned by the source model can be transferred to the target model by model transfer. For example, in the source domain the model is a Gaussian mixture model (GMM), while in the target domain the model is a deep neural network (DNN). The transfer learning then exploits the GMM to initialize and boost the DNN. This is the general recipe in the modern DNN-based speech recognition system. Recently, this model transfer has gained much attention in the deep learning community. For example, it is possible to learn simple neural nets from a complex DNN model, or vice versa [10], [35], [36]. Some interesting work in this direction will be presented in the next sections.

5) Multitask learning: In the case where the feature spaces of the source and target domains are the same but the task labels are significantly different, multitask learning is more applicable [11], [37], [38]. The basic assumption of this learning approach is that the source and target tasks are closely related, either positively or negatively, so that learning for one task helps learning the other in the form of mutual regularization. Multitask learning is a general approach that can be applied to boost various types of models including kernel regression, k-nearest neighbour, and it can be even employed to learn 'opposite' tasks simultaneously, e.g., text content classification and emotion detection [39].

A particular issue of multitask learning is how to evaluate the relevance of two tasks so that whether they can be learned together can be determined. Although there is not a simple solution yet, [38] indeed provided an interesting approach that estimates the relevance between tasks by evaluating the overlap of different tasks in the same semantic space.

C. Transfer learning in deep learning era

Deep learning almost changed everything, including transfer learning. Because deep learning gains so much success in speech and language processing [40], [41], [42], [43], we put more emphasis on transfer learning methods based on deep models in this paper. Roughly speaking, deep learning consists of various models that involve multi-level representations and the associated training/inference algorithms. Typical deep models include deep belief networks (DBNs) [44], deep Boltzmann machines (DBMs) [45], deep auto encoders (DAEs) [46], [47], deep neural networks (DNNs) [48], [41] and deep recurrent neural networks (RNNs) [49].

The success of deep models is largely attributed to their capability of learning multi-level representations (features), which simulates the processing pipeline of human brains where information is processed in a hierarchical way. The multi-level feature learning possesses several advantages. First, it can learn high-level features which are more robust against data variation than features at low-levels; second, it offers a hierarchal parameter sharing that holds great expressive power [50]; third, the feature learning can be easily conducted without any labelled data and so is cheap; fourth, with a little supervised training (fine-tuning), the learned models can be well adapted to specific tasks [11], [51], [52].

For these reasons, deep learning provides a graceful framework for transfer learning, which unifies almost all the approaches listed in Table I as *representation learning*. The basic idea is to learn some high-level robust features that are shared by multiple features and multiple tasks, so that all the knowledge/model transfers are implemented as feature transfer. This approach was advocated in the NIPS95 workshop as a major research direction, but it was not such successful until deep learning became a main stream in machine learning and related fields [53], [6], [54], [55].

The transfer learning architecture based on deep representation learning is illustrated in Fig.2. The left part of this figure is the joint training phase where heterogeneous input features are projected onto a common semantic space by different pre-processing networks, and the shared features involve rich explanatory factors that can be used to perform multiple tasks. The right part of the picture illustrates the adaptation phase, where some data $\mathcal{X}_2$ for the target task $\mathcal{Y}_2$ have been provided, either with or without labels, and the model is updated with the new data which follows a distribution $P_2'(x)$ that is different from the original distribution $P_2(x)$ in the joint training phase.

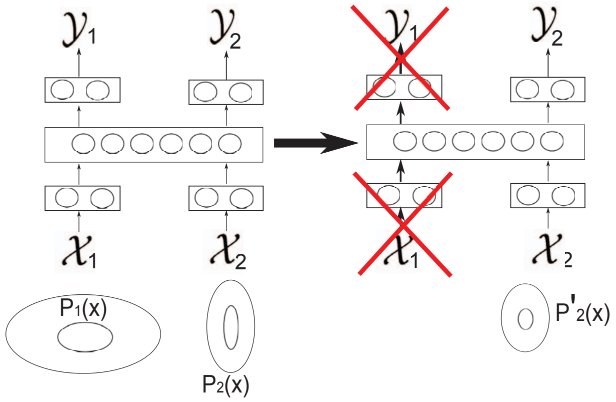

Fig. 2. Transfer learning architecture with deep representation learning. $\mathcal{X}_1$ and $\mathcal{Y}_1$ are the feature and label spaces respectively for the learning task in the source domain, and $\mathcal{X}_2$ and $\mathcal{Y}_2$ are for the learning task in the target domain. At the runtime, only the target domain is concerned.

The framework in Fig. 2 is very flexible and covers almost all the methods in Table I. For example, without the adaptation phase, it is basically a multitask learning, while using multi-domain data also implements structural correspondence learning and latent representation learning. If the joint training phase involves only a single task, then the adaptation phase implements the conventional model adaptation. It should be highlighted that a particular advantage of the representation learning framework is that the feature extractor can be trained in an unsupervised way, e.g., by restricted Boltzmann machines (RBMs) [56] or auto-associators [46], therefore little or no labelled data are required. According to [6], as long as the distribution $P(X)$ is relevant to the class-conditional distribution $P(Y|X)$, the unsupervised learning can improve the target supervised learning, in terms of convergence speed, amount of labelled data required and model quality.

An early work based on deep representation learning is [57], where the authors used unsupervised learning (denoising autoencoders) to extract high-level features, and trained a sentiment analysis system in one domain (e.g., book review). They found that the system could be directly migrated to a new domain (e.g., DVD review) and achieved much better performance than competitive approaches including structural correspondence learning (SCL) and spectral feature alignment (SFA). This work demonstrated that high-level abstract features are highly domain-independent and could be easily transferred across domains, even without any adaptation. As another example, [58] showed that CNN-based representations learned from a large image database imageNet were successfully applied to represent images in another database PASCAL VOC. A similar study was proposed recently in [59] where CNN features trained on multiple tasks were successfully applied to analyze biological images in multiple domains.

In another example called 'one-short learning' [60], high-level features trained on a large image database were found to be highly generalizable, and a very few labeled data could adapt models to recognize unseen objects by identifying the most relevant features. In a more striking configuration, the learning task can be specified as an input vector (task vector, e.g., a vector that represents a subregion of the data where the classification takes place) and fed into the deep nets together with the input data. The network then can learn the complex relationship between the data vector, the task vector, and the task labels. As long as the new task can be related to the tasks seen in the training phase (which can be obtained by a distributed task vector with which the relation between tasks can be estimated from the distance between task vectors), the new task can be well performed without any adaptation. This leads to the zero-data learning [61] and zero-shot learning [62].

III. Transfer learning in speech processing

Speech signals are pseduo-stationary and change vastly according to a large number of factors (language, gender, speaker, channel, environment, emotion, ...). Dealing with these varieties is the core challenge of the speech processing research, and transfer learning is an important tool to solve the problem. It is not possible to cover all the researches in a short paper, so we select three most prominent fields where transfer learning has gained much success: transfer across languages, transfer across speakers, and transfer across models.

A. Cross-lingual and multilingual transfer

It is natural to believe that some common patterns are shared across languages. For example, many consonants and vowels are shared across languages, defined by some universal phone sets, e.g., IPA. This sharing among human languages have been utilized explicitly or implicitly to improve statistical strength in multilingual conditions, and has delivered better models than training with monolingual data, especially for low-resource languages. This advantage has been demonstrated in a multitude of research fields, though our review simply focuses on speech recognition and speech enhancement.

Early approaches to employing cross-lingual or multilingual resources is via some linguistic correspondence, e.g., by a universal phone set or a pair-wised phone mapping [63], [64]. With the popularity of deep learning, the DNN-based multilingual approach in the form of representation learning gained much interest. The basic idea is that the features learned by DNN models tend to be language-independent at low layers and more language-dependent at high layers. Therefore multilingual data can be used to train a multilingual DNN where the low-level layers are shared by all languages, while the high-level layers are language specific. This is fully consistent with the representation learning framework shown in Fig. 2, where $\mathcal{Y}_1$ and $\mathcal{Y}_2$ represent two languages. By this sharing diagram, the features can be better learned with multilingual data, and for each language, training only the language-specific part is much easier than training the entire network.

The initial investigation was proposed in [65], where multilingual data were used to initialize the DNN model for the target language. Interesting improvement was reported and this approach was quickly followed by researchers, with both the DNN-HMM hybrid setting and the tandem setting.

With the hybrid setting, DNNs are used to replace the conventional GMMs to estimate the likelihood of HMM states. In the multilingual scenario, the hidden layers of the DNN structure are shared across languages and each language holds its own output layer [66], [67], [68]. The training process then learns a shared feature extractor as well as language-dependent classifiers. This approach was proposed independently by three research groups in 2013, and tested on three different databases: English and Mandarin data [66], eleven Romance languages [67] and the global phone dataset with 19 languages [68]. A simple extension of the above setting is to involve multiple layers in the language-specific part, or simply use different classifiers (the default is software regression), although the latter is much similar to the tandem approach discussed below.

With the tandem setting, DNNs are used as feature exactors, based on which posterior features or bottleneck features are obtained and are used to train conventional GMM-HMM systems. In [69], [70], the same DNN structure as in the hybrid setting was used to train a multilingual DNN, however the model was used to produce features (from the last hidden layer) instead of state likelihood. It was showed that the features generated by multilingual DNNs are rather language-independent and can be used directly for new languages. With limited adaptation data in the target language, additional performance could be obtained. The same approach was proposed in [71], though the features were read from a hidden layer in the middle layer (the bottle net layer with less neurons than other layers) instead of the last hidden layer. The features produced in this way are often referred to as bottleneck (BN) features. Combing with a universal phone set, the language-independent BN features can be used to train models for languages even without any labelled data [72].

The hybrid setting and tandem setting can be combined. For example, in [73], the BN feature was first derived from a multilingual DNN, and then it was combined with the original feature to train a hybrid system. A similar approach was proposed in [74], where the BN feature extractor for each language was regarded as a module, and another DNN combined the BN features from the modules of multiple languages to construct the hybrid system.

The multilingual DNN approach described above belongs to multitask learning which can be extended to more general settings. For example, in [75] phone recognition and grapheme recognition were treated as two different tasks to supervise the DNN acoustic model training. They tested on three low-resource south African languages and showed that the mutitask training indeed improved performance. They also compared the multitask training with the conditional training where the grapheme recognition provided additional input for the phone recognition, instead of co-supervision.

In a slightly different configuration, we reported a multitask learning which learns speech content and speaker accent together [76]. In this approach, a pronunciation vector that represents the accent of a speaker is generated by either an i-vector system [77] or a DNN system [78]. This pronunciation vector can be integrated in the input or hidden layers as additional features (the conditional learning), or used as an auxiliary output of a hidden layer (the multitask learning). In the latter setting, the pronunciation vector plays the role of a regularization to help learn better representations that can disentangle the underlying factors of the speech signal. We tested the method in an English ASR task where the speech data are in multiple accents (British and Chinese). We found that both the two approach could improve performance for utterances in both British and Chinese accents. An advantage with the second setting, however, is that the pronunciation vector is required only at the training phase. This is actually a heterogeneous multitask learning that has been proposed for a long time [11] but has not been studied much in speech processing.

Besides speech recognition, cross-lingual and multilingual transfer were also proposed for speech enhancement. The assumption is that the noise and reverberation that need to be removed are largely language-independent, and therefore an enhancement model trained with the data in one language can be applied directly to other languages. For example, in [79], an DNN architecture trained in English data was demonstrated to be highly effective for enhancing speech signals in Chinese, by re-using the first several layers which were assumed to be language-independent. Another study published recently from our group demonstrated that a DNN structure can be used to remove music from speech in multilingual conditions [80].

B. Speaker adaptation

Speaker adaptation is another domain in which transfer learning has gained brilliant success. In the paradigm of parametric statistic models (e.g., Gaussian models or Gaussian mixture models), maximum a posterior (MAP) estimation [12] and maximum likelihood linear regression (MLLR) [13] are two most successful methods to adapt a general model to a specific speaker. These methods are still the research focus of some authors, e.g. [81], [82], [83]. A short survey for these early-stage techniques can be found in [84].

In the deep learning era, DNN models are widely used nearly everywhere. However, adapting neural network, particular a deep one, is not simple, because DNN is a highly compact distributed model. It is not easy to learn a simple form (with limited amount of parameters) such as MLLR to update all parameters of the network. However, recent research shows that with some particular constrains on the adaptation structure, speaker adaptation is possible for DNN models.

An early study reported in [85] introduced a user vector (user code) to represent a speaker, and the vector was augmented to the input and hidden layers. The learning then trained the network and the speaker code simultaneously. To adapt to a new speaker, the network was fixed while the speaker vector was inferred by the conventional back-propagation algorithm [86]. This approach was extended in [87] by involving a transform matrix before the speaker vector was augmented to the input and hidden layers, possibly in the form of low-rank matrices.

In a similar setting, the speaker code can be replaced by a more general speaker vector produced by exotic models, e.g., the famous i-vector [77]. Different from the speaker code approach, these speaker vectors do not need to be adapted (although it is possible) [88], [89], [90], [91]. An advantage of using exotic speaker vectors is that the speaker vector model can be trained with a large unlabelled database in an unsupervised fashion. A disadvantage is that no phone information is considered when deriving the vectors, at least it is case with the i-vector model. A careful analysis for the i-vector augmentation was conducted in [92], which showed that i-vectors not only compensate for speaker variance but also acoustic variance.

In contrast to involving an speaker vector, the second approach to speaker adaptation for DNN models is to update the DNN model directly, with some constraints on which components of the DNN should be adapted. For example, the adaptation can be conducted on the input layer [93], [94], the activations of hidden layers [95], [96], [97], or the output layer [94]. Some comparison for adaptation on different components can be found in [98], [99]. In order to constrain the adaptation more aggressively, [100], [101] studied a singular value decomposition (SVD) approach which decomposes a weight matrix as production of low rank matrices, and only the singular values are updated for each speaker. Another constraint for speaker adaptation is based on a prior distribution over the output of the adapted network, which is imposed by the output of the speaker-independent DNN, in the form of KL-divergence [102].

Another interesting approach to speaker adaptation for DNN models is to apply transfer learning to project features to a canonical speaker-independent space where a model can be well trained. For example, the famous constrained MLLR (CMLLR) in the HMM-GMM architecture [13]. Recently, an auto-encoder trained with speaker vectors (obtained from a regression-based speaker transform) was used to produce speaker-independent BN features [103]. A similar approach was studied in [104], though an i-vector was used as the speaker representation.

Most of the above researches are based on the DNN structure. Recent research shows that RNNs can be adapted in a similar way. For example, [105] reported an extensive study on speaker adaptation for LSTMs. It was found that LSTMs can be effectively adapted by using speaker-adaptive (SA) front-end (e.g., a speaker-aware DNN projection [104]), or by inserting speaker-dependent layers.

It should be noted that DNN itself possesses great advantage of learning multiple conditions. Therefore, DNN models trained with a large amount of data of multiple speakers can deal with speaker variation pretty well. This conjecture was demonstrated by [99], which showed that the adaptation methods provide some improvement if the network is small and the amount of training data is medium, however for a large network trained with a large mount of data, the improvement is insignificant.

The techniques discussed above are mostly applied to speech recognition, however they can be easily migrated to other applications. For example in HMM-based speech synthesis, model adaptation based on MAP and MLLR has been widely used to produce specific voice, e.g., [106], [107], [108], [109]. Particularly, speaker adaptation is often coupled with language adaptation to obtain multilingual synthesis, e.g., by state mapping [107], [110], [111]. For DNN-based speech synthesis [112], [113], [114], it is relatively new and the adaptation methods have not been extensively studied, except a few exceptions [115], [116].

C. Model transfer

A recent progress in transfer learning is to learn a new model (child model) from an existing model (teacher model), which is known as model transfer. This was mentioned in the seminal paper of multitask learning [11] and was recently rediscovered by several researchers in the context of deep learning [117], [10], [118]. The initial idea was that the teacher model learns rich knowledge from the training data and this knowledge can be used to guide the training of child models which are simple and hence unable to learn many details without the teacher's guide. To distill the knowledge from the teacher model, the logit matching approach proposed by Ba [117] teaches the child model by encouraging its logits (activations before softmax) close to those generated by the teacher model in terms of square error, and the dark knowledge distiller model proposed by Hinton [10] encourages the output of the child model close to those of the teacher model in terms of cross entropy.

This approach has been applied to learn simple models from complex models so that the simple model can approach the performance of the complex model. For example, [118] utilized the output of a complex DNN as regularization to learn a small DNN that is suitable for speech recognition on mobile devices. [119] used a complex RNN to train a DNN. Recently, a new architecture called FitNet was proposed [120]. Instead of regularizing the output, FitNet regularizes hidden units so that knowledge learned by the intermediate representations can be transferred to the target model, which is suitable for training a model whose label space is different from that of the teacher model. This work was further extended in [121], where multiple hidden layers were regularized by the teacher model. Another example is to transfer heterogeneous models. For instance, in [122], unsupervised learning models (PCA and ICA) were used to model the outputs of a DNN model. This in fact treats the DNN output as an intermediate feature, and uses the feature for general tasks, e.g., classifying instances from novel classes.

Our recent work [35] showed that this model transfer can not only learn simple models from complex models, but also the reverse: a weak model can be used to teach a stronger model. In our work [35], a DNN model was used to train a powerful complex RNN. We found that by the model transfer learning, RNNs can be learned pretty well with the regularization of a DNN model, though the teacher model is weaker than the target one. In a related work [36], we found that the model transfer learning can be used as a new pre-training approach, and it even works in some scenarios where the RBM pre-training and layer-wised discriminative pre-training do not work. Additionally, combining the RMB-based pre-training and the model transfer pre-training can offer additional gains, at least in our experimental setting where the training data is not very abundant.

IV. Transfer learning in language processing

As in speech processing, the basic assumption of transfer learning for language processing is also intuitive: all human languages share some common semantic structures (e.g., concepts and syntactic rules). Following this idea, the simple way of transfer learning in multilingual or multi-domain scenarios is to construct some cross-lingual/cross-domain correspondence so that knowledge learned in one language/domain can be transferred and reused in another language/domain. For example, a bi-lingual lexicon can be used to provide instance-level correspondence so that syntactic knowledge learned in one language can be used to improve the syntactic learning in the second language [123]. Another approach that gained more attention recently is to learn a common latent space that are shared by different languages or domains, so that knowledge can be aggregated, leading to improved statistic strength for probabilistic modeling in each single language or domain.

Once again, transfer learning is such a broad research field and the research of language processing is even more broad itself, which makes a detailed review for all the research fields impossible in such a short paper. We will focus on two particular fields: cross-lingual learning and cross-domain learning, particularly for the document classification task.

A. Cross-lingual and multilingual transfer learning

A straightforward way to transfer knowledge between languages is to translate words from one language to another by a bi-lingual lexicon. For example, this approach was used in [124] to translate a maximum entropy (ME) classifier trained in English data to a classifier used for classifying Chinese documents. In another work from our group, we have applied this approach successfully to train multilingual language models, where some foreign words need to be addressed [125]. Word-by-word translation, however, is obviously not ideal since no syntactic constraints in different languages are considered. A more complicated approach is to translate the whole sentence by machine translation [126], so that any labelling or classification tasks in one language can be conducted with models trained in another language.

A more recent approach to multilingual learning is to learn some common latent structures/representations based on multilingual data. For example, the multilingual LDA model proposed in [127] assumes a common latent topic space, so that words from multiple languages can share the same topics. This is similar to the RMB-based heterogeneous factor learning [27]: both are based on unsupervised learning with weak supervision, i.e., no word alignment is required.

A similar approach proposed in [128] learns multilingual word clusters, where a cluster may involve words from different languages. This was achieved by means of a probabilistic model over large amounts of monolingual data in two languages, coupled with parallel data through which cross-lingual correspondence was obtained. Applying to the NER task, it was found that up to 26% performance improvement was observed with the multi-lingual model. This work was extend in [129] where cross-lingual clusters were used to 'directly' transfer an NER model trained in the source language to the target language.

Another approach to constructing common latent space is by linear projection instead of statistical models. For example, in the heterogeneous feature augmentation (HFA) approach proposed in [29], two linear projections are learned to project features in different languages to a common space. In their study, these projections were used to produce additional features that were augmented to the original features to train the model in the target language. An interesting part of their approach is to train the supervision model (e.g., SVM) in the source and target languages *simultaneously*. This leads to a joint optimization for the common space projections as well as the classifiers. The approach was tested on a text classification task with the Reuters multilingual database and obtained good performance. In another work [24], a linear projection was learned by optimizing a set of multi-lingual classifier, each of which predicted the existence of the words of a bi-lingual word-pair. The approach was tested on cross-lingual topic discovery and sentiment classification.

Recently, word embedding becomes a hot topic [130], [131], [132], [133], [134]. Intuitively, word embedding represents each word as a low-dimensional dense vector (word vector) with the constraint that relevant words are located more closely than irrelevant words. This embedding enables semantic computing over words, and provides new ways for mulitilingual learning: if word vectors can be trained in a multilingual fashion, regressors/classifiers trained on these vectors naturally apply to multiple languages.

A simple approach is to map word vectors trained in individual languages to a single space. For example, in [135], it was found that a linear transform can project word vectors trained in one languages to word vectors in another language so that relevant words are put closely, in spite of their languages. This projection can be learned simply by some pivot word pairs from the two languages. We extended this work in [136] by modeling the transfer as an orthogonal transform. A more systematic approach was proposed by [137], where different languages were projected to the same space by different projections, and the projections were determined by maximizing the canonical correlation of the corresponding words in the projected space. This approach requires one-to-one word correspondence, which was obtained by aligned parallel data.

A potential problem of the above approaches is that the word vectors and projections are learned separately. The approach proposed in [138] does not learn any projection, instead the bi-lingual correspondence was taken into account in the embedding process. This work was based on the neural LM model [130] and changed the objective function by involving an extra term that encourages relevant words in different languages located together. The relevance of words in different languages was derived from aligned parallel data.

In another work [139], the relevance constraint was employed at the sentence level. Word vectors were aggregated into a sentence embedding, and relevant sentences were embedded closely. This approach does not require word alignment and so can be easily implemented. Additionally, this approach can be simply extended to document level models, for which only document pairs are required, without any sentence-level alignment. This approach was tested on a multilingual classification task.

A similar work was proposed by [140]. As in [139], only sentence pairs are required in the learning; the difference is that the embedding leveraged both monolingual data and bi-lingual data, and employd noise-contrastive training to improve efficiency. Good performance was obtained in both cross-lingual document classification and word-level translation.

An interesting research that involves much ingredient of deep learning was proposed by [30]. The basic idea is to learn high-level document features individually in each language by unsupervised learning (i.e., mSDA in that paper), and then learn the correspondence (transform) using parallel data. The raw and high-level features can be combined to train the classifier in one language, and documents in another language can be transferred to the rich language and are classified there. The idea of applying unsupervised learning to learn high-level features is prominent, which may help remove noises in the raw data thus leading to more reliable transform estimation. The approach was tested on several multilingual sentiment classification tasks where the raw document feature was TF-IDF and the high-level features were learned by mSDA. Good performance was reported.

B. *Cross-domain transfer learning*

Cross-domain transfer learning has two different meaning: when the domain refers to applications, then the difference is in the data distribution; when it refers to features, then the difference is in feature types or modalities, e.g., audio feature or image feature. We focus on the feature domain transfer, which is relatively new and invokes much interest recently. With the simplest approach, multi-modal features can be combined either at the feature level or the score level. For example on the semantic relatedness task, [141] concatenated visual and textual features to train multi-stream systems; in [142], the scores predicted by multiple models based on different features are combined. A more complex setting involves transferring knowledge between models built with heterogeneous features. Note that some authors regard different languages as different domains, e.g., [30]. However, we focus on transfer learning between different feature modalities.

An example is the work proposed in [25], where the authors used co-occurrence data to estimate the correspondence between different features, i.e., image and text. The estimated correspondence was then used to assist the classification task in the target domain, by transferring the target features to the source domain where a good classification model had been constructed. The authors formulated this transfer process using a Markov chain and risk minimization inference. The method was tested on a text-aided image classification task and achieved significant performance improvement.

The common latent space approach was studied in [143], with the task of image segmentation and labelling. The model was based on kernelized canonical correlation analysis which finds a mapping between visual and textual representations by projecting them into a latent semantic space.

Deep learning provides an elegant way for cross-domain transfer learning, with its great power in learning high-level representations shared by multiple modalities [54]. For example, in [62], [144], images and words are embedded in the same low-dimensional space via neural networks, by which image classification can be improved by the word embedding, even for classes without any image training data. [145] proposed a multi-modal neural language modeling approach with which the history and prediction can be both text and images, so that the prediction between multiple modalities becomes possible. In [146], an RNN structure based on dependency-tree was proposed to embed textual sentences into compositional vectors, which were then projected together with image representations to a common space. Within this common space, multi-modal retrieval and annotation can be easily conducted. The same idea was proposed by [147], though deep Boltzmann machines were used instead of DNNs to infer the common latent space.

C. Model transfer

Model transfer, which aims to learn one model from another, has not yet been extensively studied in language processing. A recent work [148] studied a knowledge distilling approach on the sentiment classification task. The original classifier was a large neural net with large word vectors as input, and a small network was learned in two ways: either using the output of the large network as supervision or directly transferring large word vectors to smaller ones.

In a recent study [149], we show that it is possible to learn a neural model using supervision from a Bayesian model. Specifically, we tried to learn a document vector from the raw TF input using a neural net, supervised by the vector representation produced by latent Dirichlet allocation (LDA). Our experimental results showed that with a two-layer neural network, it is possible to learn document vectors that are quite similar to the ones produced by LDA, while the inference is hundreds of times faster.

V. Perspective and conclusions

We gave a very brief review of transfer learning, and introduced some applications of this approach in speech and language processing. Due to the broad landscape of this research and the limited knowledge of the authors, only very limited areas were touched. Also, many important contributions in the 'history' had to be omitted, for the sake of emphasis on more recent directions in the past few years, especially deep learning. Even with such a limited review, we can still clearly see the important role that transfer learning plays and how fast it has evolved recently. For speech and language processing, transfer learning is essentially important as both speech and language are diverse, imbalance, dynamic and inter-linked, which makes transfer learning inevitable.

Transfer learning can be conducted in very different manners. It can be conducted as a shared learning that learns various domains and tasks together, or as a tandem learning which learns a model in one domain/task and migrates the model to another domain/task. It can be conducted with a supervised way where labeled data are used to refine the classifier, or an unsupervised way where numerous unlabelled data are used to learn better representations. It can be used to transfer instances, representations, structures and models. It can transfer between different distributions, different features and different tasks.

Go back to the NIPS 95 workshop, where some questions were raised by the famous researchers at that time. Two decades later, we can answer some of the questions, while other remains mystery:

- *What do we mean by related tasks and how can we identify them?* It is still difficult to measure relatedness, particularly with the complex configurations of transfer learning. However, we do know some possible metrics, e.g., the relatedness between marginal and conditional distributions [6] in unsupervised feature learning, or representation overlap in model adaptation [38]. Particularly, we now know that even two tasks are intuitively unrelated (e.g., speech recognition and speaker recognition), transfer learning still works by utilizing the fact that the tasks are unrelated [39].
- *How do we predict when transfer will help (or hurt)?* Again, it is not easy to find a complete solution. However some approaches indeed can alleviate negative transfer, e.g., [150], [38]. With deep learning, the risk of negative transfer seems substantially reduced. For example, any data in related domains can be used to assist learning abstract features, even they are sampled from a distribution different from the target domain [23]. This is not the case twenty years ago.
- *What are the benefits: speed, generalization, intelligibility,...?* Seems all of these can be improved by transfer learning.
- *What should be transferred: internal representations, parameter settings, features,...?* We now know all these components can be transferred.
- *How should it be transferred: weight initialization, biasing the error metric,...?* All these methods can be used, although it seems that the regularization view is more attractive and it is related to modifying the objective function.
- *How do we look inside to see what has been transferred?* This question is more related to model adaptation and the answer is model-dependent. For example with a DNN model which is highly compact, it is not simple to investigate which part of the model has been changed after adaptation.

Transfer learning has been widely studied in speech and language processing, particularly for model adaptation. Recent advance in multilingual learning and heterogeneous feature transform demonstrates the power of transfer learning in a more clear way. Nevertheless, compared to the very diverse methods studied in the machine learning community, application of transfer learning in speech and language research is still very limited. There are many questions remain unanswered, for example: can we learn common representations for both speech, language and speaker recognition? Can we learn acoustic models for voiced speech and whistle speech together? How about sign language? How to use large volume of unlabeled video data to regularize speech models? How pronunciation models can be used to regularize NLP tasks? How to involve heterogeneous resources including audio, visual, language to solve the most challenging tasks in the respective research fields? How to utilize the large amount of unlabeled data more efficiently in the big-data era? To solve these problems, we believe collaboration among researchers who have been used to work independently in their own areas is mostly required.

ACKNOWLEDGEMENT

This research was supported by the National Science Foundation of China (NSFC) under the project No. 61271389 and No. 61371136, and the National Basic Research Program (973 Program) of China under Grant No. 2013CB329302. It was also supported by the MESTDC PhD Foundation Project No.20130002120011, as well as Sinovoice and Huilan Ltd. Thanks to Zhiyuan Tang for the careful reference collection.

REFERENCES

[1] J. H. Martin and D. Jurafsky, "Speech and language processing," *International Edition*, 2000.

[2] J. Benesty, *Springer handbook of speech processing*. Springer Science & Business Media, 2008.

[3] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 5, pp. 1060–1089, 2013.

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.

[5] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *The Journal of Machine Learning Research*, vol. 10, pp. 1633–1685, 2009.

[6] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *ICML Unsupervised and Transfer Learning*, 2012.

[7] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: A survey," *Knowledge-Based Systems*, vol. 80, pp. 14–23, 2015.

[8] S. Thrun and L. Pratt, *Learning to learn*. Springer Science & Business Media, 2012.

[9] "NIPS 95 workshop on learning to learn: Knowledge consolidation and transfer in inductive systems," 1995. [Online]. Available: http://socrates.acadiau.ca/courses/comp/dsilver/NIPS95_LTL/transfer.workshop.1995.html

[10] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS 2014 Deep Learning Workshop*, 2014.

[11] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[12] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and audio processing*, vol. 2, no. 2, pp. 291–298, 1994.

[13] C. J. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[14] P. E. Utgoff, "Incremental induction of decision trees," *Machine learning*, vol. 4, no. 2, pp. 161–186, 1989.

[15] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.

[16] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 1, 2011, p. 1541.

[17] Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang, "Heterogeneous transfer learning for image classification." in *AAAI*, 2011.

[18] H.-Y. Wang and Q. Yang, "Transfer learning by structural analogy," in *AAAI*. Citeseer, 2011.

[19] O. Arandjelovic and R. Cipolla, "Incremental learning of temporally-coherent Gaussian mixture models," *Society of Manufacturing Engineers (SME) Technical Papers*, pp. 1–1, 2006.

[20] A. Declercq and J. H. Piater, "Online learning of Gaussian mixture models-a two-level approach." in *VISAPP (1)*, 2008, pp. 605–611.

[21] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences TRP 1530, University of Wisconsin C Madison, 2005.

[22] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *Neural Networks, IEEE Transactions on*, vol. 22, no. 2, pp. 199–210, 2011.

[23] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 759–766.

[24] P. Prettenhofer and B. Stein, "Cross-lingual adaptation using structural correspondence learning," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 1, p. 13, 2011.

[25] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu, "Translated learning: Transfer learning across different feature spaces," in *Advances in neural information processing systems*, 2008, pp. 353–360.

[26] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1785–1792.

[27] B. Wei and C. J. Pal, "Heterogeneous transfer learning with RBMs." in *AAAI*, 2011.

[28] X. Shi, Q. Liu, W. Fan, P. S. Yu, and R. Zhu, "Transfer learning on heterogenous feature spaces via spectral transformation," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 1049–1054.

[29] L. Duan, D. Xu, and I. Tsang, "Learning with augmented features for heterogeneous domain adaptation," *arXiv preprint arXiv:1206.4660*, 2012.

[30] J. T. Zhou, S. J. Pan, I. W. Tsang, and Y. Yan, "Hybrid heterogeneous transfer learning through deep learning," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[31] D. Gentner, "Structure-mapping: A theoretical framework for analogy," *Cognitive science*, vol. 7, no. 2, pp. 155–170, 1983.

[32] D. Gentner and K. J. Holyoak, "Reasoning and learning by analogy: Introduction." *American Psychologist*, vol. 52, no. 1, p. 32, 1997.

[33] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the 2006 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2006, pp. 120–128.

[34] J. G. Carbonell, *Learning by analogy: Formulating and generalizing plans from past experience*. Springer, 1983.

[35] D. Wang, C. Liu, Z. Tang, Z. Zhang, and M. Zhao, "Recurrent neural network training with dark knowledge transfer," *arXiv preprint arXiv:1505.04630*, 2015.

[36] Z. Tang, D. Wang, Y. Pan, and Z. Zhang, "Knowledge transfer pre-training," *arXiv preprint arXiv:1506.02256*, 2015.

[37] J. Baxter, "A model of inductive bias learning," *J. Artif. Intell. Res.(JAIR)*, vol. 12, pp. 149–198, 2000.

[38] J. Guinney, Q. Wu, and S. Mukherjee, "Estimating variable structure and dependence in multitask learning via gradients," *Machine Learning*, vol. 83, no. 3, pp. 265–287, 2011.

[39] B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil, "Exploiting unrelated tasks in multi-task learning," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 951–959.

[40] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[41] L. Deng and D. Yu, "Deep learning: Methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3-4, pp. 197–387, 2013. [Online]. Available: http://dx.doi.org/10.1561/2000000039

[42] X. He, J. Gao, and L. Deng, "Deep learning for natural language processing and related applications (Tutorial at ICASSP)," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[43] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.

[44] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[45] R. Salakhutdinov and G. E. Hinton, "Deep boltzmann machines," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 448–455.

[46] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle *et al.*, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.

[47] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a

deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[48] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[49] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 6645–6649.

[50] Y. Bengio and O. Delalleau, "On the expressive power of deep architectures," in *Algorithmic Learning Theory*. Springer, 2011, pp. 18–36.

[51] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[52] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams *et al.*, "Recent advances in deep learning for speech research at Microsoft," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8604–8608.

[53] S. M. Gutstein, *Transfer learning techniques for deep neural nets*. The University of Texas at El Paso, 2010.

[54] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[55] Y. Bengio, I. J. Goodfellow, and A. Courville, *Deep Learning*, 2015, book in preparation for MIT Press. [Online]. Available: http://www.iro.umontreal.ca/~bengioy/dlbook

[56] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[57] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 513–520.

[58] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1717–1724.

[59] W. Zhang, R. Li, T. Zeng, Q. Sun, S. Kumar, J. Ye, and S. Ji, "Deep model based transfer and multi-task learning for biological image analysis," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1475–1484.

[60] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 4, pp. 594–611, 2006.

[61] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks." in *AAAI*, vol. 1, no. 2, 2008, p. 3.

[62] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Advances in neural information processing systems*, 2013, pp. 935–943.

[63] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1, pp. 31–51, 2001.

[64] N. T. Vu, F. Kraus, and T. Schultz, "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5000–5003.

[65] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 246–251.

[66] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7304–7308.

[67] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8619–8623.

[68] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7319–7323.

[69] K. Vesely, M. Karafiát, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 336–341.

[70] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6704–6708.

[71] Z. Tuske, J. Pinto, D. Willett, and R. Schluter, "Investigation on cross-and multilingual mlp features under matched and mismatched acoustical conditions," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7349–7353.

[72] K. M. Knill, M. J. Gales, A. Ragni, and S. P. Rath, "Language independent and unsupervised acoustic models for speech recognition and keyword spotting," in *Proc. Interspeech14*, 2014.

[73] P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6975–6979.

[74] J. Gehring, Q. B. Nguyen, F. Metze, and A. Waibel, "DNN acoustic modeling with modular multi-lingual feature extraction networks," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 344–349.

[75] D. Chen, B. Mak, C.-C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5592–5596.

[76] X. Z. Zhiyuan Tang, "Speech recognition with pronunciation vecotrs," CSLT, Tsinghua University, 2015. [Online]. Available: http://cslt.riit.tsinghua.edu.cn/publications.php?Publication-trp

[77] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[78] V. Ehsan, L. Xin, M. Erik, L. M. Ignacio, and G.-D. Javier, "Deep neural networks for small footprint text-dependent speaker verification," *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, vol. 28, no. 4, pp. 357–366, 2014.

[79] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Cross-language transfer learning for deep neural network based speech enhancement," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, 2014, pp. 336–340.

[80] M. Zhao, D. Wang, Z. Zhang, and X. Zhang, "Music removal by convolutional denoising autoencoder in speech recognition," in *Interspeech 2015*, 2015.

[81] M. L. Seltzer and A. Acero, "Separating speaker and environmental variability using factored transforms." in *INTERSPEECH*, 2011, pp. 1097–1100.

[82] D. Povey and K. Yao, "A basis representation of constrained MLLR transforms for robust adaptation," *Computer Speech & Language*, vol. 26, no. 1, pp. 35–51, 2012.

[83] Y. Miao, F. Metze, and A. Waibel, "Learning discriminative basis coefficients for eigenspace MLLR unsupervised adaptation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7927–7931.

[84] K. Shinoda, "Speaker adaptation techniques for automatic speech recognition," *Proc. APSIPA ASC 2011*, 2011.

[85] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7942–7946.

[86] ——, "Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition." in *INTERSPEECH*, 2013, pp. 1248–1252.

[87] S. Xue, O. Abdel-Hamid, H. Jiang, and L. Dai, "Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6339–6343.

[88] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 55–59.

[89] P. Karanasou, Y. Wang, M. J. Gales, and P. C. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *Proc Interspeech*, 2014.

[90] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *Proc. ICASSP*, 2014.

[91] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6334–6338.

[92] M. Rouvier and B. Favre, "Speaker adaptation of DNN-based ASR with i-vectors: Does it actually adapt models to speakers?" in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[93] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," in *Proc. EUROSPEECH'95*. International Speech Communication Association, 1995.

[94] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 366–369.

[95] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. De Mori, "Linear hidden transformations for adaptation of hybrid ANN/HMM models," *Speech Communication*, vol. 49, no. 10, pp. 827–835, 2007.

[96] S. M. Siniscalchi, J. Li, and C.-H. Lee, "Hermitian polynomial for speaker adaptation of connectionist speech recognition systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2152–2161, 2013.

[97] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 171–176.

[98] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *Interspeech'10*, 2010.

[99] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7947–7951.

[100] J. Xue, J. Li, D. Yu, M. Seltzer, and Y. Gong, "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6359–6363.

[101] S. Xue, H. Jiang, and L. Dai, "Speaker adaptation of hybrid NN/HMM model for speech recognition based on singular value decomposition," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, 2014, pp. 1–5.

[102] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7893–7897.

[103] Y. Tang, A. Mohan, R. C. Rose, and C. Ma, "Deep neural network trained with speaker representation for speaker normalization," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6329–6333.

[104] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," in *Interspeech'14*, 2014.

[105] Y. Miao and F. Metze, "On speaker adaptation of long short-term memory recurrent neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)(To Appear). ISCA*, 2015.

[106] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 805–808.

[107] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis." in *Interspeech*, 2009, pp. 528–531.

[108] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[109] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 66–83, 2009.

[110] H. Liang, J. Dines, and L. Saheer, "A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4598–4601.

[111] M. Gibson and W. Byrne, "Unsupervised intralingual and cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 895–904, May 2011.

[112] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2129–2139, 2013.

[113] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3844–3848.

[114] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4455–4459.

[115] B. Potard, P. Motlicek, and D. Imseng, "Preliminary work on speaker adaptation for DNN-based speech synthesis," Idiap, Tech. Rep., 2015.

[116] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Interspeech 2015*, 2015.

[117] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Advances in Neural Information Processing Systems*, 2014, pp. 2654–2662.

[118] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, September 2014. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=230080

[119] W. Chan, N. R. Ke, and I. Lane, "Transferring knowledge from a RNN to a DNN," *arXiv preprint arXiv:1504.01483*, 2015.

[120] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[121] M. Long and J. Wang, "Learning transferable features with deep adaptation networks," *arXiv preprint arXiv:1502.02791*, 2015.

[122] Y. Lu, "Unsupervised learning of neural network outputs," *arXiv preprint arXiv:1506.00990*, 2015.

[123] G. Durrett, A. Pauls, and D. Klein, "Syntactic transfer using a bilingual lexicon," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 1–11.

[124] L. Shi, R. Mihalcea, and M. Tian, "Cross language text classification by model translation and semi-supervised learning," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 1057–1067.

[125] X. Ma, X. Wang, and D. Wang, "Recognize foreign low-frequency words with similar pairs," in *Interspeech 2015*, 2015.

[126] P. Koehn, *Statistical machine translation*. Cambridge University Press, 2009.

[127] W. De Smet, J. Tang, and M.-F. Moens, "Knowledge transfer across multilingual corpora via latent topics," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2011, pp. 549–560.

[128] O. Täckström, R. McDonald, and J. Uszkoreit, "Cross-lingual word clusters for direct transfer of linguistic structure," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, pp. 477–487.

[129] O. Täckström, "Nudging the envelope of direct transfer methods for multilingual named entity recognition," in *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*. Association for Computational Linguistics, 2012, pp. 55–63.

[130] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models," in *Innovations in Machine Learning*. Springer, 2006, pp. 137–186.

[131] A. Mnih and G. E. Hinton, "A scalable hierarchical distributed language model," in *NIPS*, 2008, pp. 1081–1088.

[132] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2010, pp. 384–394.

[133] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *The Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[134] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[135] T. Mikolov, Q. V. Le, and I. Sutskever, “Exploiting similarities among languages for machine translation,” *arXiv preprint arXiv:1309.4168*, 2013.

[136] C. Xing, D. Wang, C. Liu, and Y. Lin, “Normalized word embedding and orthogonal transform for bilingual word translation,” in *NAACL'15*, 2015.

[137] M. Faruqui and C. Dyer, “Improving vector space word representations using multilingual correlation,” in *EACL'14*. Association for Computational Linguistics, 2014.

[138] A. Klementiev, I. Titov, and B. Bhattarai, “Inducing crosslingual distributed representations of words,” in *COLING'12*. Citeseer, 2012.

[139] K. M. Hermann and P. Blunsom, “Multilingual models for compositional distributed semantics,” *arXiv preprint arXiv:1404.4641*, 2014.

[140] S. Gouws, Y. Bengio, and G. Corrado, “Bilbowa: Fast bilingual distributed representations without word alignments,” *arXiv preprint arXiv:1410.2455*, 2014.

[141] E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran, “Distributional semantics in technicolor,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 136–145.

[142] C. W. Leong and R. Mihalcea, “Going beyond text: A hybrid image-text approach for measuring word relatedness.” in *IJCNLP*, 2011, pp. 1403–1407.

[143] R. Socher and L. Fei-Fei, “Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 966–973.

[144] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, “Devise: A deep visual-semantic embedding model,” in *Advances in Neural Information Processing Systems*, 2013, pp. 2121–2129.

[145] R. Kiros, R. Salakhutdinov, and R. Zemel, “Multimodal neural language models,” in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 595–603.

[146] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, “Grounded compositional semantics for finding and describing images with sentences,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014.

[147] N. Srivastava and R. R. Salakhutdinov, “Multimodal learning with deep boltzmann machines,” in *Advances in neural information processing systems*, 2012, pp. 2222–2230.

[148] L. Mou, G. Li, Y. Xu, L. Zhang, and Z. Jin, “Distilling word embeddings: An encoding approach,” *arXiv preprint arXiv:1506.04488*, 2015.

[149] D. Zhang, T. Luo, D. Wang, and R. Liu, “Learning from LDA using deep neural networks,” *arXiv preprint arXiv:1508.01011*, 2015.

[150] M. Long, J. Wang, G. Ding, D. Shen, and Q. Yang, “Transfer learning with graph co-regularization,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 7, pp. 1805–1818, 2014.