Chapter 1
Representation Learning

Liang Zhao, Lingfei Wu, Peng Cui and Jian Pei

Abstract In this chapter, we first describe what representation learning is and why we need representation learning. Among the various ways of learning representations, this chapter focuses on deep learning methods: those that are formed by the composition of multiple non-linear transformations, with the goal of resulting in more abstract and ultimately more useful representations. We summarize the representation learning techniques in different domains, focusing on the unique challenges and models for different data types including images, natural languages, speech signals and networks. Last, we summarize this chapter.

1.1 Representation Learning: An Introduction

The effectiveness of machine learning techniques heavily relies on not only the design of the algorithms themselves, but also a good representation (feature set) of data. Ineffective data representations that lack some important information or contains incorrect or huge redundant information could lead to poor performance of the algorithm in dealing with different tasks. The goal of representation learning is to extract sufficient but minimal information from data. Traditionally, this can be achieved via human efforts based on the prior knowledge and domain expertise on the data and tasks, which is also named as feature engineering. In deploying ma-

Liang Zhao
Department of Computer Science, Emory University, e-mail: liang.zhao@emory.edu

Lingfei Wu
JD.COM Silicon Valley Research Center, e-mail: lwu@email.wm.edu

Peng Cui
Department of Computer Science, Tsinghua University, e-mail: cuip@tsinghua.edu.cn

Jian Pei
Department of Computer Science, Simon Fraser University, e-mail: jpei@cs.sfu.ca

chine learning and many other artificial intelligence algorithms, historically a large portion of the human efforts goes into the design of prepossessing pipelines and data transformations. More specifically, feature engineering is a way to take advantage of human ingenuity and prior knowledge in the hope to extract and organize the discriminative information from the data for machine learning tasks. For example, political scientists may be asked to define a keyword list as the features of social-media text classifiers for detecting those texts on societal events. For speech transcription recognition, one may choose to extract features from raw sound waves by the operations including Fourier transformations. Although feature engineering is widely adopted over the years, its drawbacks are also salient, including: 1) Intensive labors from domain experts are usually needed. This is because feature engineering may require tight and extensive collaboration between model developers and domain experts. 2) Incomplete and biased feature extraction. Specifically, the capacity and discriminative power of the extracted features are limited by the knowledge of different domain experts. Moreover, in many domains that human beings have limited knowledge, what features to extract itself is an open questions to domain experts, such as cancer early prediction. In order to avoid these drawbacks, making learning algorithms less dependent on feature engineering has been a highly desired goal in machine learning and artificial intelligence domains, so that novel applications could be constructed faster and hopefully addressed more effectively.

The techniques of representation learning witness the development from the traditional representation learning techniques to more advanced ones. The traditional methods belong to "shallow" models and aim to learn transformations of data that make it easier to extract useful information when building classifiers or other predictors, such as Principal Component Analysis (PCA) (Wold et al, 1987), Gaussian Markov random field (GMRF) (Rue and Held, 2005), and Locality Preserving Projections (LPP) (He and Niyogi, 2004). Deep learning-based representation learning is formed by the composition of multiple non-linear transformations, with the goal of yielding more abstract and ultimately more useful representations. In the light of introducing more recent advancements and sticking to the major topic of this book, here we majorly focus on deep learning-based representation learning, which can be categorized into several types: (1) Supervised learning, where a large number of labeled data are needed for the training of the deep learning models. Given the well-trained networks, the output before the last fully-connected layers is always utilized as the final representation of the input data; (2) Unsupervised learning (including self-supervised learning), which facilitates the analysis of input data without corresponding labels and aims to learn the underlying inherent structure or distribution of data. The pre-tasks are utilized to explore the supervision information from large amounts of unlabelled data. Based on this constructed supervision information, the deep neural networks are trained to extract the meaningful representations for the future downstream tasks; (3) Transfer learning, which involves methods that utilize any knowledge resource (i.e., data, model, labels, etc.) to increase model learning and generalization for the target task. Transfer learning encompasses different scenarios including multi-task learning (MTL), model adaptation, knowledge transfer, co-variance shift, etc. There are also other important representation learning meth-

ods such as reinforcement learning, few-shot learning, and disentangled representation learning.

It is important to define what is a good representation. As the definition by Bengio (2008), representation learning is about learning the (underlying) features of the data that make it easier to extract useful information when building classifiers or other predictors. Thus, the evaluation of a learned representation is closely related to its performance on the downstream tasks. For example, in the data generation task based on a generative model, a good representation is often the one that captures the posterior distribution of the underlying explanatory factors for the observed input. While for a prediction task, a good representation is the one that captures the minimal but sufficient information of input data to correctly predict the target label. Besides the evaluation from the perspective of the downstream tasks, there are also some general properties that the good representations may hold, such as the smoothness, the linearity, capturing multiple explanatory and casual factors, holding shared factors across different tasks and simple factor dependencies.

1.2 Representation Learning in Different Areas

In this section, we summarize the development of representation learning on four different representative areas: (1) image processing; (2) speech recognition; (3) Natural language processing; and (4) network analysis. For the representation learning in each research area, we consider some of the fundamental questions that have been driving research in this area. Specifically, what makes one representation better than another, and how should we compute its representation? Why is the representation learning important in that area? Also, what are appropriate objectives for learning good representations? We also introduce the relevant typical methods and their development from the perspective of three main categories: supervised representation learning, unsupervised learning and transfer learning, respectively.

1.2.1 Representation Learning for Image Processing

Image representation learning is a fundamental problem in understanding the semantics of various visual data, such as photographs, medical images, document scans, and video streams. Normally, the goal of image representation learning for image processing is to bridge the semantic gap between the pixel data and semantics of the images. The successful achievements of image representation learning have enpowered many real-world problems, including but not limited to image search, facial recognition, medical image analysis, photo manipulation and target detection.

In recent years, we have witnessed a fast advancement of image representation learning from handcrafted feature engineering to that from scratch through deep neural network models. Traditionally, the patterns of images are extracted with the

help of hand-crafted features by human beings based on prior knowledge. For example, Huang et al (2000) extracted the character's structure features from the strokes, then use them to recognize the handwritten characters. Rui (2005) adopted the morphology method to improve local feature of the characters, then use PCA to extract features of characters. However, all of these methods need to extract features from images manually and thus the prediction performances strongly rely on the prior knowledge. In the field of computer vision, manual feature extraction is very cumbersome and impractical because of the high dimensionality of feature vectors. Thus, representation learning of images which can automatically extract meaningful, hidden and complex patterns from high-dimension visual data is necessary. Deep learning-based representation learning for images is learned in an end-to-end fashion, which can perform much better than hand-crafted features in the target applications, as long as the training data is of sufficient quality and quantity.

Supervised Representation Learning for image processing. In the domain of image processing, supervised learning algorithm, such as Convolution Neural Network (CNN) and Deep Belief Network (DBN), are commonly applied in solving various tasks. One of the earliest deep-supervised-learning-based works was proposed in 2006 (Hinton et al, 2006), which is focused on the MNIST digit image classification problem, outperforming the state-of-the-art SVMs. Following this, deep convolutional neural networks (ConvNets) showed amazing performance which is greatly depends on their properties of shift in-variance, weights sharing and local pattern capturing. Different types of network architectures were developed to increase the capacity of network models, and larger and larger datasets were collected these days. Various networks including AlexNet (Krizhevsky et al, 2012), VGG (Simonyan and Zisserman, 2014b), GoogLeNet (Szegedy et al, 2015), ResNet (He et al, 2016a), and DenseNet (Huang et al, 2017a) and large scale datasets, such as ImageNet and OpenImage, have been proposed to train very deep convolutional neural networks. With the sophisticated architectures and large-scale datasets, the performance of convolutional neural networks keeps outperforming the state-of-the-arts in various computer vision tasks.

Unsupervised Representation Learning for image processing. Collection and annotation of large-scale datasets are time-consuming and expensive in both image datasets and video datasets. For example, ImageNet contains about 1.3 million labeled images covering 1,000 classes while each image is labeled by human workers with one class label. To alleviate the extensive human annotation labors, many unsupervised methods were proposed to learn visual features from large-scale unlabeled images or videos without using any human annotations. A popular solution is to propose various pretext tasks for models to solve, while the models can be trained by learning objective functions of the pretext tasks and the features are learned through this process. Various pretext tasks have been proposed for unsupervised learning, including colorizing gray-scale images (Zhang et al, 2016d) and image inpainting (Pathak et al, 2016). During the unsupervised training phase, a predefined pretext task is designed for the models to solve, and the pseudo labels for the pretext task are automatically generated based on some attributes of data. Then the models are trained according to the objective functions of the pretext tasks. When trained

with pretext tasks, the shallower blocks of the deep neural network models focus on the low-level general features such as corners, edges, and textures, while the deeper blocks focus on the high-level task-specific features such as objects, scenes, and object parts. Therefore, the models trained with pretext tasks can learn kernels to capture low-level features and high-level features that are helpful for other downstream tasks. After the unsupervised training is finished, the learned visual features in this pre-trained models can be further transferred to downstream tasks (especially when only relatively small data is available) to improve performance and overcome over-fitting.

Transfer Learning for image processing. In real-world applications, due to the high cost of manual labeling, sufficient training data that belongs to the same feature space or distribution as the testing data may not always be accessible. Transfer learning mimics the human vision system by making use of sufficient amounts of prior knowledge in other related domains (i.e., source domains) when executing new tasks in the given domain (i.e., target domain). In transfer learning, both the training set and the test set can contribute to the target and source domains. In most cases, there is only one target domain for a transfer learning task, while either single or multiple source domains can exist. The techniques of transfer learning in images processing can be categorized into feature representation knowledge transfer and classifier-based knowledge transfer. Specifically, feature representation transfer methods map the target domain to the source domains by exploiting a set of extracted features, where the data divergence between the target domain and the source domains can be significantly reduced so that the performance of the task in the target domain is improved. For example, classifier-based knowledge-transfer methods usually share the common trait that the learned source domain models are utilized as prior knowledge, which are used to learn the target model together with the training samples. Instead of minimizing the cross-domain dissimilarity by updating instances' representations, classifier-based knowledge-transfer methods aim to learn a new model that minimizes the generalization error in the target domain via the provided training set from both domains and the learned model.

Other Representation Learning for Image Processing. Other types of representation learning are also commonly observed for dealing with image processing, such as reinforcement learning, and semi-supervised learning. For example, reinforcement learning are commonly explored in the task of image captioning Liu et al (2018a); Ren et al (2017) and image editing Kosugi and Yamasaki (2020), where the learning process is formalized as a sequence of actions based on a policy network.

1.2.2 Representation Learning for Speech Recognition

Nowadays, speech interfaces or systems have become widely developed and integrated into various real-life applications and devices. Services like Siri [1], Cortana [2], and Google Voice Search [3] have become a part of our daily life and are used by millions of users. The exploration in speech recognition and analysis has always been motivated by a desire to enable machines to participate in verbal human-machine interactions. The research goals of enabling machines to understand human speech, identify speakers, and detect human emotion have attracted researchers' attention for more than sixty years across several distinct research areas, including but not limited to Automatic Speech Recognition (ASR), Speaker Recognition (SR), and Speaker Emotion Recognition (SER).

Analyzing and processing speech has been a key application of machine learning (ML) algorithms. Research on speech recognition has traditionally considered the task of designing hand-crafted acoustic features as a separate distinct problem from the task of designing efficient models to accomplish prediction and classification decisions. There are two main drawbacks of this approach: First, the feature engineering is cumbersome and requires human knowledge as introduced above; and second, the designed features might not be the best for the specific speech recognition tasks at hand. This has motivated the adoption of recent trends in the speech community towards the utilization of representation learning techniques, which can learn an intermediate representation of the input signal automatically that better fits into the task at hand and hence lead to improved performance. Among all these successes, deep learning-based speech representations play an important role. One of the major reasons for the utilization of representation learning techniques in speech technology is that speech data is fundamentally different from two-dimensional image data. Images can be analyzed as a whole or in patches, but speech has to be formatted sequentially to capture temporal dependency and patterns.

Supervised representation learning for speech recognition. In the domain of speech recognition and analyzing, supervised representation learning methods are widely employed, where feature representations are learned on datasets by leveraging label information. For example, restricted Boltzmann machines (RBMs) (Jaitly and Hinton, 2011; Dahl et al, 2010) and deep belief networks (DBNs) (Cairong et al, 2016; Ali et al, 2018) are commonly utilized in learning features from speech for different tasks, including ASR, speaker recognition, and SER. For example, in 2012, Microsoft has released a new version of their MAVIS (Microsoft Audio Video Indexing Service) speech system based on context-dependent deep neural networks (Seide et al, 2011). These authors managed to reduce the word error rate on four major benchmarks by about 30% (e.g., from 27.4% to 18.5% on RT03S) com-

[1] Siri is an artificial intelligence assistant software that is built into Apple's iOS system.

[2] Microsoft Cortana is an intelligent personal assistant developed by Microsoft, known as "the world's first cross-platform intelligent personal assistant".

[3] Google Voice Search is a product of Google that allows you to use Google to search by speaking to a mobile phone or computer, that is, to use the legendary content on the device to be identified by the server, and then search for information based on the results of the recognition

pared to the traditional models based on Gaussian mixtures. Convolutional neural networks are another popular supervised models that are widely utilized for feature learning from speech signals in tasks such as speech and speaker recognition (Palaz et al, 2015a,b) and SER Latif et al (2019); Tzirakis et al (2018). Moreover, it has been found that LSTMs (or GRUs) can help CNNs in learning more useful features from speech by learning both the local and long-term dependency (Dahl et al, 2010).

Unsupervised Representation Learning for speech recognition. Unsupervised representation learning from large unlabelled datasets is an active area of speech recognition. In the context of speech analysis, it is able to exploit the practically available unlimited amount of unlabelled corpora to learn good intermediate feature representations, which can then be used to improve the performance of a variety of downstream supervised learning speech recognition tasks or the speech signal synthetic tasks. In the tasks of ASR and SR, most of the works are based on Variational Auto-encoder (VAEs), where a generative model and an inference model are jointly learned, which allows them to capture latent representations from observed speech data (Chorowski et al, 2019; Hsu et al, 2019, 2017). For example, Hsu et al (2017) proposed a hierarchical VAE to capture interpretable and disentangled representations from speech without any supervision. Other auto-encoding architectures like Denoised Autoencoder(DAEs) are also found very promising in finding speech representations in an unsupervised way, especially for noisy speech recognition (Feng et al, 2014; Zhao et al, 2015). Beyond the aforementioned, recently, adversarial learning (AL) is emerging as a powerful tool in learning unsupervised representation for speech, such as generative adversarial nets (GANs). It involves at least a generator and a discriminator, where the former tries to generates as realistic as possible data to obfuscate the latter which also tries its best to deobfuscate. Hence both of the generator and discriminator can be trained and improved iteratively in an adversarial way, which result in more discriminative and robust features. Among these, GANs (Chang and Scherer, 2017; Donahue et al, 2018), adversarial autoencoders (AAEs) Sahu et al (2017) are becoming mostly popular in modeling speech not only in ASR but also SR and SER.

Transfer Learning for speech recognition. Transfer learning (TL) encompasses different approaches, including MTL, model adaptation, knowledge transfer, covariance shift, etc. In the domain of speech recognition, representation learning gained much interest in these approaches of TL including but not limited to domain adaptation, multi-task learning, and self-taught learning. In terms of Domain Adaption, speech is a typical example of heterogeneous data and thus, a mismatch always exists between the probability distributions of source and target domain data. To build more robust systems for speech-related applications in real-life, domain adaptation techniques are usually applied in the training pipeline of deep neural networks to learn representations which are able to explicitly minimize the difference between the distribution of data in the source and target domains (Sun et al, 2017; Swietojanski et al, 2016). In terms of MTL, representations learned can successfully increases the performance of speech recognition without requiring contextual speech data, since speech contains multi-dimensional information (message, speaker, gender, or emotion) that can be used as auxiliary tasks. For example, In the task of ASR, by us-

ing MTL with different auxiliary tasks including gender, speaker adaptation, speech enhancement, it has been shown that the learned shared representations for different tasks can act as complementary information about the acoustic environment and give a lower word error rate (WER) (Parthasarathy and Busso, 2017; Xia and Liu, 2015).

Other Representation Learning for speech recognition. Other than the above-mentioned three categories of representation learning for speech signals, there are also some other representation learning techniques commonly explored, such as semi-supervised learning and reinforcement learning. For example, in the speech recognition for ASR, semi-supervised learning is mainly used to circumvent the lack of sufficient training data. This can be achieved either by creating features fronts ends (Thomas et al, 2013), or by using multilingual acoustic representations (Cui et al, 2015), or by extracting an intermediate representation from large unpaired datasets (Karita et al, 2018). RL is also gaining interest in the area of speech recognition, and there have been multiple approaches to model different speech problems, including dialog modeling and optimization (Levin et al, 2000), speech recognition (Shen et al, 2019), and emotion recognition (Sangeetha and Jayasankar, 2019).

1.2.3 Representation Learning for Natural Language Processing

Besides speech recognition, there are many other Natural Language Processing (NLP) applications of representation learning, such as the text representation learning. For example, Google's image search exploits huge quantities of data to map images and queries in the same space (Weston et al, 2010) based on NLP techniques. In general, there are two types of applications of representation learning in NLP. In one type, the semantic representation, such as the word embedding, is trained in a pre-training task (or directly designed by human experts) and is transferred to the model for the target task. It is trained by using language modeling objective and is taken as inputs for other down-stream NLP models. In the other type, the semantic representation lies within the hidden states of the deep learning model and directly aims for better performance of the target tasks in an end-to-end fashion. For example, many NLP tasks want to semantically compose sentence or document representation, such as tasks like sentiment classification, natural language inference, and relation extraction, which require sentence representation.

Conventional NLP tasks heavily rely on feature engineering, which requires careful design and considerable expertise. Recently, representation learning, especially deep learning-based representation learning is emerging as the most important technique for NLP. First, NLP is typically concerned with multiple levels of language entries, including but not limited to characters, words, phrases, sentences, paragraphs, and documents. Representation learning is able to represent the semantics of these multi-level language entries in a unified semantic space, and model complex semantic dependence among these language entries. Second, there are various NLP tasks that can be conducted on the same input. For example, given a sentence, we

can perform multiple tasks such as word segmentation, named entity recognition, relation extraction, co-reference linking, and machine translation. In this case, it will be more efficient and robust to build a unified representation space of inputs for multiple tasks. Last, natural language texts may be collected from multiple domains, including but not limited to news articles, scientific articles, literary works, advertisement and online user-generated content such as product reviews and social media. Moreover, texts can also be collected from different languages, such as English, Chinese, Spanish, Japanese, etc. Compared to conventional NLP systems which have to design specific feature extraction algorithms for each domain according to its characteristics, representation learning enables us to build representations automatically from large-scale domain data and even add bridges among these languages from different domains. Given these advantages of representation learning for NLP in the feature engineering reduction and performance improvement, many researchers have developed efficient algorithms on representation learning, especially deep learning-based approaches, for NLP.

Supervised Representation Learning for NLP. Deep neural networks in the supervised learning setting for NLP emerge from distributed representation learning, then to CNN models, and finally to RNN models in recent years. At early stage, distributed representations are first developed in the context of statistical language modeling by Bengio (2008) in so-called neural net language models. The model is about learning a distributed representation for each word (i.e., word embedding). Following this, the need arose for an effective feature function that extracts higher-level features from constituting words or n-grams. CNNs turned out to be the natural choice given their properties of excellent performance in computer vision and speech processing tasks. CNNs have the ability to extract salient n-gram features from the input sentence to create an informative latent semantic representation of the sentence for downstream tasks. This domain was pioneered by Collobert et al (2011) and Kalchbrenner et al (2014), which led to a huge proliferation of CNN-based networks in the succeeding literature. The neural net language model was also improved by adding recurrence to the hidden layers (Mikolov et al, 2011a) (i.e., RNN), allowing it to beat the state-of-the-art (smoothed n-gram models) not only in terms of perplexity (exponential of the average negative log-likelihood of predicting the right next word) but also in terms of WER in speech recognition. RNNs use the idea of processing sequential information. The term "recurrent" applies as they perform the same computation over each token of the sequence and each step is dependent on the previous computations and results. Generally, a fixed-size vector is produced to represent a sequence by feeding tokens one by one to a recurrent unit. In a way, RNNs have "memory" over previous computations and use this information in current processing. This template is naturally suited for many NLP tasks such as language modeling (Mikolov et al, 2010, 2011b), machine translation (Liu et al, 2014; Sutskever et al, 2014), and image captioning (Karpathy and Fei-Fei, 2015).

Unsupervised Representation Learning for NLP. Unsupervised learning (including self-supervised learning) has made a great success in NLP, for the plain text itself contains abundant knowledge and patterns about languages. For example, in most deep learning based NLP models, words in sentences are first mapped to their corre-

sponding embeddings via the techniques, such as word2vec Mikolov et al (2013b), GloVe Pennington et al (2014), and BERT Devlin et al (2019), before sending to the networks. However, there are no human-annotated "labels" for learning those word embeddings. To acquire the training objective necessary for neural networks, it is necessary to generate "labels" intrinsically from the existing data. Language modeling is a typical unsupervised learning task, which can construct the probability distribution over sequences of words and does not require human annotations. Based on the distributional hypothesis, using the language modeling objective can lead to hidden representations that encode the semantics of words. Another typical unsupervised learning model in NLP is auto-encoder (AE), which consists of a reduction (encoding) phase and a reconstruction (decoding) phase. For example, recursive auto-encoders (which generalize recurrent networks with VAE) have been used to beat the state-of-the-art at the moment of its publication in full sentence paraphrase detection (Socher et al, 2011) by almost doubling the F1 score for paraphrase detection.

Transfer Learning for NLP. Over the recent years, the field of NLP has witnessed fast growth of transfer learning methods via sequential transfer learning models and architectures, which significantly improved upon the state-of-the-arts on a wide range of NLP tasks. In terms of domain adaption, the sequential transfer learning consists of two stages: a pretraining phase in which general representations are learned on a source task or domain followed by an adaptation phase during which the learned knowledge is applied to a target task or domain. The domain adaption in NLP is categorized into model-centric, data-centric, and hybrid approaches. Model-centric methods target the approaches to augmenting the feature space, as well as altering the loss function, the architecture, or the model parameters (Blitzer et al, 2006). Data-centric methods focus on the data aspect and involve pseudo-labeling (or bootstrapping) where only small number of classes are shared between the source and target datasets (Abney, 2007). Lastly, hybrid-based methods are built by both data- and model-centric models. Similarly, great advances have also been made into the multi-task learning in NLP, where different NLP tasks can result in better representation of texts. For example, based on a convolutional architecture, Collobert et al (2011) developed the SENNA system that shares representations across the tasks of language modeling, part-of-speech tagging, chunking, named entity recognition, semantic role labeling, and syntactic parsing. SENNA approaches or sometimes even surpasses the state-of-the-art on these tasks while is simpler and much faster than traditional predictors. Moreover, learning word embeddings can be combined with learning image representations in a way that allow associating texts and images.

Other Representation Learning for NLP. In NLP tasks, when a problem gets more complicated, it requires more knowledge from domain experts to annotate training instances for fine-grained tasks and thus increases the cost of data labeling. Therefore, sometimes it requires the models or systems can be developed efficiently with (very) few labeled data. When each class has only one or a few labeled instances, the problem becomes a one/few-shot learning problem. The few-shot learning problem is derived from computer vision and has also been studied in NLP

recently. For example, researchers have explored few-shot relation extractio (Han et al, 2018) where each relation has a few labeled instances, and low-resource machine translation (Zoph et al, 2016) where the size of the parallel corpus is limited.

1.2.4 Representation Learning for Networks

Beyond popular data like images, texts, and sounds, network data is another important data type that is becoming ubiquitous across a large scale of real-world applications ranging from cyber-networks (e.g., social networks, citation networks, telecommunication networks, etc.) to physical networks (e.g., transportation networks, biological networks, etc). Networks data can be formulated as graphs mathematically, where vertices and their relationships jointly characterize the network information. Networks and graphs are very powerful and flexible data formulation such that sometimes we could even consider other data types like images, and texts as special cases of it. For example, images can be considered as grids of nodes with RGB attributes which are special types of graphs, while texts can also be organized into sequential-, tree-, or graph-structured information. So in general, representation learning for networks is widely considered as a promising yet more challenging tasks that require the advancement and generalization of many techniques we developed for images, texts, and so forth. In addition to the intrinsic high complexity of network data, the efficiency of representation learning on networks is also an important issues considering the large-scale of many real-world networks, ranging from hundreds to millions or even billions of vertices. Analyzing information networks plays a crucial role in a variety of emerging applications across many disciplines. For example, in social networks, classifying users into meaningful social groups is useful for many important tasks, such as user search, targeted advertising and recommendations; in communication networks, detecting community structures can help better understand the rumor spreading process; in biological networks, inferring interactions between proteins can facilitate new treatments for diseases. Nevertheless, efficient and effective analysis of these networks heavily relies on good representations of the networks.

Traditional feature engineering on network data usually focuses on obtaining a number of predefined straightforward features in graph levels (e.g., the diameter, average path length, and clustering co-efficient), node levels (e.g., node degree and centrality), or subgraph levels (e.g., frequent subgraphs and graph motifs). Those limited number of hand-crafted, well-defined features, though describe several fundamental aspects of the graphs, discard the patterns that cannot be covered by them. Moreover, real-world network phenomena are usually highly complicated require sophisticated, unknown combinations among those predefined features or cannot be characterized by any of the existing features. In addition, traditional graph feature engineering usually involve expensive computations with super-linear or exponential complexity, which often makes many network analytic tasks computationally expensive and intractable over large-scale networks. For example, in dealing with

the task of community detection, classical methods involve calculating the spectral decomposition of a matrix with at least quadratic time complexity with respect to the number of vertices. This computational overhead makes algorithms hard to scale to large-scale networks with millions of vertices.

More recently, network representation learning (NRL) has aroused a lot of research interest. NRL aims to learn latent, low-dimensional representations of network vertices, while preserving network topology structure, vertex content, and other side information. After new vertex representations are learned, network analytic tasks can be easily and efficiently carried out by applying conventional vector-based machine learning algorithms to the new representation space. Earlier work related to network representation learning dates back to the early 2000s, when researchers proposed graph embedding algorithms as part of dimensionality reduction techniques. Given a set of independent and identically distributed (i.i.d.) data points as input, graph embedding algorithms first calculate the similarity between pairwise data points to construct an affinity graph, e.g., the k-nearest neighbor graph, and then embed the affinity graph into a new space having much lower dimensionality. However, graph embedding algorithms are designed on i.i.d. data mainly for dimensionality reduction purpose, which usually have at least quadratic time complexity with respect to the number of vertices.

Since 2008, significant research efforts have shifted to the development of effective and scalable representation learning techniques that are directly designed for complex information networks. Many network representation learning algorithms (Perozzi et al, 2014; Yang et al, 2015b; Zhang et al, 2016b; Manessi et al, 2020) have been proposed to embed existing networks, showing promising performance for various applications. These methods embed a network into a latent, low-dimensional space that preserves structure proximity and attribute affinity. The resulting compact, low-dimensional vector representations can be then taken as features to any vector-based machine learning algorithms. This paves the way for a wide range of network analytic tasks to be easily and efficiently tackled in the new vector space, such as node classification (Zhu et al, 2007), link prediction (Lü and Zhou, 2011), clustering (Malliaros and Vazirgiannis, 2013), network synthesis (You et al, 2018b). The following chapters of this book will then provide a systematic and comprehensive introduction into network representation learning.

1.3 Summary

Representation learning is a very active and important field currently, which heavily influences the effectiveness of machine learning techniques. Representation learning is about learning the representations of the data that makes it easier to extract useful and discriminative information when building classifiers or other predictors. Among the various ways of learning representations, deep learning algorithms have increasingly been employed in many areas nowadays where the good representation can be learned in an efficient and automatic way based on large amount of complex

and high dimensional data. The evaluation of a representation is closely related to its performance on the downstream tasks. Generally, there are also some general properties that the good representations may hold, such as the smoothness, the linearity, disentanglement, as well as capturing multiple explanatory and casual factors.

We have summarized the representation learning techniques in different domains, focusing on the unique challenges and models for different areas including the processing of images, natural language, and speech signals. For each area, there emerges many deep learning-based representation techniques from different categories, including supervised learning, unsupervised learning, transfer learning, disentangled representation learning, reinforcement learning, etc. We have also briefly mentioned about the representation learning on networks and its relations to that on images, texts, and speech, in order for the elaboration of it in the following chapters.