Information Sciences 585 (2022) 498–528

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A data-centric review of deep transfer learning with applications to text data

Samar Bashath [a,1], Nadeesha Perera [a,1], Shailesh Tripathi [a], Kalifa Manjang [a], Matthias Dehmer [b,c,d,e], Frank Emmert Streib [a,*]

[a] Predictive Society and Data Analytics Lab, Tampere University, Tampere, Korkeakoulunkatu 10, 33720 Tampere, Finland
[b] Department of Computer Science, Swiss Distance University of Applied Sciences, Brig, Switzerland
[c] School of Science, Xian Technological University, Xian, China
[d] College of Artificial Intelligence, Nankai University, Tianjin, China
[e] Department of Biomedical Computer Science and Mechatronics, The Health and Life Science University, UMIT, Hall in Tyrol, Austria

ARTICLE INFO

Article history:
Received 4 May 2021
Received in revised form 15 September 2021
Accepted 19 November 2021
Available online 27 November 2021

2010 MSC:
00-01
99-00

Keywords:
Transfer learning
Deep learning
Natural language processing
Machine learning
Domain adaptation

ABSTRACT

In recent years, many applications are using various forms of deep learning models. Such methods are usually based on traditional learning paradigms requiring the consistency of properties among the feature spaces of the training and test data and also the availability of large amounts of training data, e.g., for performing supervised learning tasks. However, many real-world data do not adhere to such assumptions. In such situations transfer learning can provide feasible solutions, e.g., by simultaneously learning from data-rich source data and data-sparse target data to transfer information for learning a target task. In this paper, we survey deep transfer learning models with a focus on applications to text data. First, we review the terminology used in the literature and introduce a new nomenclature allowing the unequivocal description of a transfer learning model. Second, we introduce a visual taxonomy of deep learning approaches that provides a systematic structure to the many diverse models introduced until now. Furthermore, we provide comprehensive information about text data that have been used for studying such models because only by the application of methods to data, performance measures can be estimated and models assessed.

1. Introduction

We added comments regrarding the tabels. Please send us an updated proof to check after you made the corrections (we do not know how to add a comment) Deep learning models consist of multiple layers which help the model to learn a representation or embedding of the data with multiple levels of abstraction [60,48,123]. Machine learning in general, including deep learning, is based on two main assumptions [12]. First, the training and testing data should be drawn from the same underlying distribution [32]. Second, training data should be large enough for learning patterns in the data, because it is known that deep learning models require large quantities of training data to learn latent patterns in the data [118,40].

* Corresponding author at: Predictive Society and Data Analytics Lab, Tampere University, Tampere, Korkeakoulunkatu 10, 33720 Tampere, Finland.
E-mail address: v@bio-complexity.com (F.E. Streib).
[1] Both authors contributed equally.

https://doi.org/10.1016/j.ins.2021.11.061
0020-0255/

Due to the fact that transfer learning provides means to soften both assumptions, this approach is promising for many real-world applications suffering, e.g., from limited training data. Unfortunately, so far, transfer learning, is still undervalued compared to traditional learning paradigms, e.g., supervised learning. This is especially the case for applications analyzing text data. For this reason, we survey recent deep transfer learning approaches with a particular focus on applications to text data. A key idea of transfer learning is to extend the concept of a domain and a task. Specifically, instead of having only one domain and one task, transfer learning considers a source domain and a target domain and a source task and a target task. From this, a model is learned by leveraging information provided in the source domain by optimizing the results of the target task. This is called transfer of knowledge between the source and target and can be realized in a number of different ways. Importantly, the model is only assessed for the target task while the source task serves merely as an auxiliary evaluation. This extension allows to systematically accommodate, e.g., differences in feature spaces, label spaces or prediction functions between the source and the target.

For applications, these formal extensions have beneficial consequences. For instance, while a shift in the distribution between the training and testing data usually requires the models to be newly rebuilt by using new training data from the new underlying distribution [56,31] because otherwise the performance of the model suffers [76,87], the above issue is addressed by approaches from heterogeneous transfer learning. Even more importantly, for the insufficient training data problem, which is notorious in certain application areas, e.g., medicine, transfer learning is capable of circumventing this, e.g., by parameter transfer between the source and target model [145]. We would like to note that in general transfer learning does not refer to one particular approach but rather to a family of (very different) strategies. Hence, there are vast differences between transfer learning models and the way they address such problems. Also such strategies depend on the underlying data and the application domain. For this reason, we focus in this paper on deep transfer learning methods for analyzing text data.

Despite the fact that there are many deep transfer learning approaches for text applications, so far there is no dedicated review paper about this domain in the context of transfer learning. Instead, there are a number of review papers about other aspects of transfer learning. For example, an early review about general forms of transfer learning that has been widely recognized is the paper by [92]. An update of such a general review has been presented by [130] emphasizing the distinction between homogeneous and heterogeneous transfer learning. In contrast, the review by [36] focused solely on heterogeneous transfer learning, while the review by [157] focused on homogeneous transfer learning touching also briefly on deep transfer learning. A further general review, however, limited to domain adaptation focusing on theoretical considerations, e.g., risk bounds and PAC (probably approximately correct) learning is from [58]. A similar theoretical review can be found in [146] also providing information about deep learning approaches. Finally, a non-comprehensive, very brief review of deep transfer learning methods has been presented in [118]. Neither of the latter three reviews has a focus on text applications.

We would like to highlight that most reviews about applications of transfer learning are for image analysis. For instance, the paper by [115] discussed transfer learning approaches for various image applications, including image classification, and action recognition, [94] discussed visual domain adaptation, [42] focused on emotion recognition, and [118] discussed computer vision and image classification. In addition, there are also reviews about transfer learning for further application areas such as activity recognition [30], reinforcement learning [119] and sentiment classification [3]. However, while the latter is based on text data, deep transfer learning models are not reviewed. In contrast, the review by [71] provides a brief survey of deep learning approaches for text data but with a sole focus on sentiment analysis.

In this paper, review deep transfer learning models with a focus on applications to text data. For completeness, we are also including a review of important definitions and previous classifications of general transfer learning methods. In Section 3, we discuss text data frequently used in studies when analyzing deep transfer learning methods. In Section 4, we introduce a visual taxonomy of deep transfer learning models for text applications and in Section 5 we provide a discussion thereof. This paper finishes with concluding remarks in Section 6.

2. Background of transfer learning

In this section, we provide some background information about transfer learning in general. Section 2.1 describes the underlying concept of transfer learning and provides examples related to the analysis of text data. Section 2.2 gives important definitions needed for transfer learning and discusses various special cases. In Section 2.3, we review previous categorizations of transfer learning, and in Section 2.4 we present a new nomenclature.

2.1. Motivation and underlying concept

Transfer learning is a general machine learning paradigm [136,113] that allows the transferring of knowledge from one domain (called source domain) to another domain (called target domain) allowing the data in the source and target to be different [92,109]. One advantage of transfer learning over other learning paradigms, e.g., supervised learning, is that transfer learning can deal with insufficient training data in the target domain [130] by exploiting information from a different, but related (source) domain to make predictions of labels of unseen target instances [154]. In general, it is a technique for

improving a learner, e.g., a classifier, by transferring information between two related domains [36]. Although, it is a challenge to design a system able to leverage information from one domain or task for another domain or tasks [27], the advantages of transfer learning are numerous. For instance, much less time is needed for training a new model and fewer records and data are required for the target domain [30]. In contrast to other machine learning algorithms that can learn a new task without any prior knowledge, transfer learning can ultimately boost predictive performance on a new target task by leveraging information gained from solving previous but related (source) tasks [9]. This is especially relevant when there is limited or no data available for a particular problem, but ample data is available for a related problem.

In this paper our focus is on analyzing text data. For this reason, we provide in the following two examples from this application domain to visualize the problem. In general, transfer learning finds widespread application in natural language processing [87]. An example for this is Named Entity Recognition (NER), where the aim is to identify an entity from a text into semantic types such as location, person, or organization. Among several kinds of data, electronic health records (eHR) provide informative textual information, because they contains detailed information about patients and their clinical history. However, getting labeled data is difficult in a clinical context. Also, there are privacy issues, which make it difficult to share data. In this scenario, it would be beneficial if one would train a classifier with large amounts of eHR data inside a hospital and then transfer learned information (instead of data) outside the hospital to train another classifier for a related task even when only a limited amount of data is available.

Another example is sentiment analysis, in which we classify reviews of a product, e.g., a laptop, into positive and negative sentiments. For such a classification task, one needs to gather many reviews of a product, and then train the classifier on these reviews. However, the process of labeling data can be extremely costly. In such a situation, one could apply transfer learning for adapting a classifier, e.g., trained on camera reviews, to classify the reviews about laptops.

In Fig. 1, we show a visualization of the general idea underlying transfer learning. Fig. 1 A shows the conventional setting of supervised learning where data from a domain is used to learn a model for making predictions as specified by a task. In contrast, transfer learning extends the concept of a domain and a task. Specifically, instead of having only one domain and one task, transfer learning distinguishes between a source domain and a target domain and a source task and a target task. From these a model is learned by leveraging information provided in the source domain and by optimizing the results of the target task. We would like to highlight that the transfer between the source and the target can be accomplished by a number of different approaches, as discussed in detail below. For this reason in Fig. 1 B there are two arrows from the source to the target; one connects the domains whereas the other connects the models. This means, one can either adjust the data or the model. Below we will formalize these approaches.

2.2. Definitions

In order to obtain a quantitative understanding of transfer learning, we need to review some definitions. The first definitions are about a domain and a task [92].

Definition 2.1 (*Domain*). A domain D is a tuple $D=\{(\mathscr{X}), P(X)\}$ where $\mathscr{X}$ is the set of all instances, X is an instance, i.e., $X \in \mathscr{X}$, and $P(X)$ the marginal probability distribution over all instances.

Definition 2.2 (*Task*). Given a domain, $D=\{(\mathscr{X}), P(X)\}$, a task T is given by the tuple $T=\{\mathscr{Y}, f\}$ where $\mathscr{Y}$ is the label space and f is a prediction function, i.e., $f : \mathscr{X} \rightarrow \mathscr{Y}$.

We would like to remark that the prediction function can not be observed, but the function is learned from training data. The prediction function assigns a label to a given instance and can be written as conditional probability distribution given by $P(Y|X)$. Thus T can be written as $T = \{\mathscr{Y}, P(Y|X)\}$ where $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$.

To illustrate the above definitions, let's consider the problem of review classifications where the task is to classify reviews into positive and negative sentiments. In this situation, $\mathscr{X}$ is the space of all word vectors, x_i is the ith instance corresponding to a review, X is a particular review sample, $\mathscr{Y}$ is the set of all labels which are positive and negative, Y is a particular label for particular review, and y_i is positive or negative.

Based on the definition of a domain and a task, we can now define transfer learning [92].

Definition 2.3 (*Transfer learning*). Given a source domain D_S, target domain D_T, source task T_S corresponds to D_S and target task T_T corresponds to D_T. Transfer learning improves the learning of the predictive function in the target f_T using the information in D_S and T_S where $D_S \neq D_T$ and/or $T_S \neq T_T$.

Based on the general definition of transfer learning, a number of important sub-cases can be distinguished. Since a domain is given by $D=\{(\mathscr{X}), P(X)\}$, $D_S \neq D_T$ implies that either $\mathscr{X}_S \neq \mathscr{X}_T$ or $P_S(X) \neq P_T(X)$. It is important to highlight that from $\mathscr{X}_S \neq \mathscr{X}_T$ follows $P_S(X) \neq P_T(X)$. Hence, both statements are not independent from each other. In contrast, $P_S(X) \neq P_T(X)$ does not follow $\mathscr{X}_S \neq \mathscr{X}_T$ but also $\mathscr{X}_S = \mathscr{X}_T$ is possible. In summary, whenever the source domain differs from the target domain then they have also a different marginal distribution, however, the converse is not true. In the literature, the case $P_S(X) \neq P_T(X)$ with $P_S(Y|X) = P_T(Y|X)$ is called covariate shift [94,57].

For instance, in our review classification example above, having two different but related domains could mean that the word-features may be different (e.g., the text in the source is in a different language from the text in the target), which means

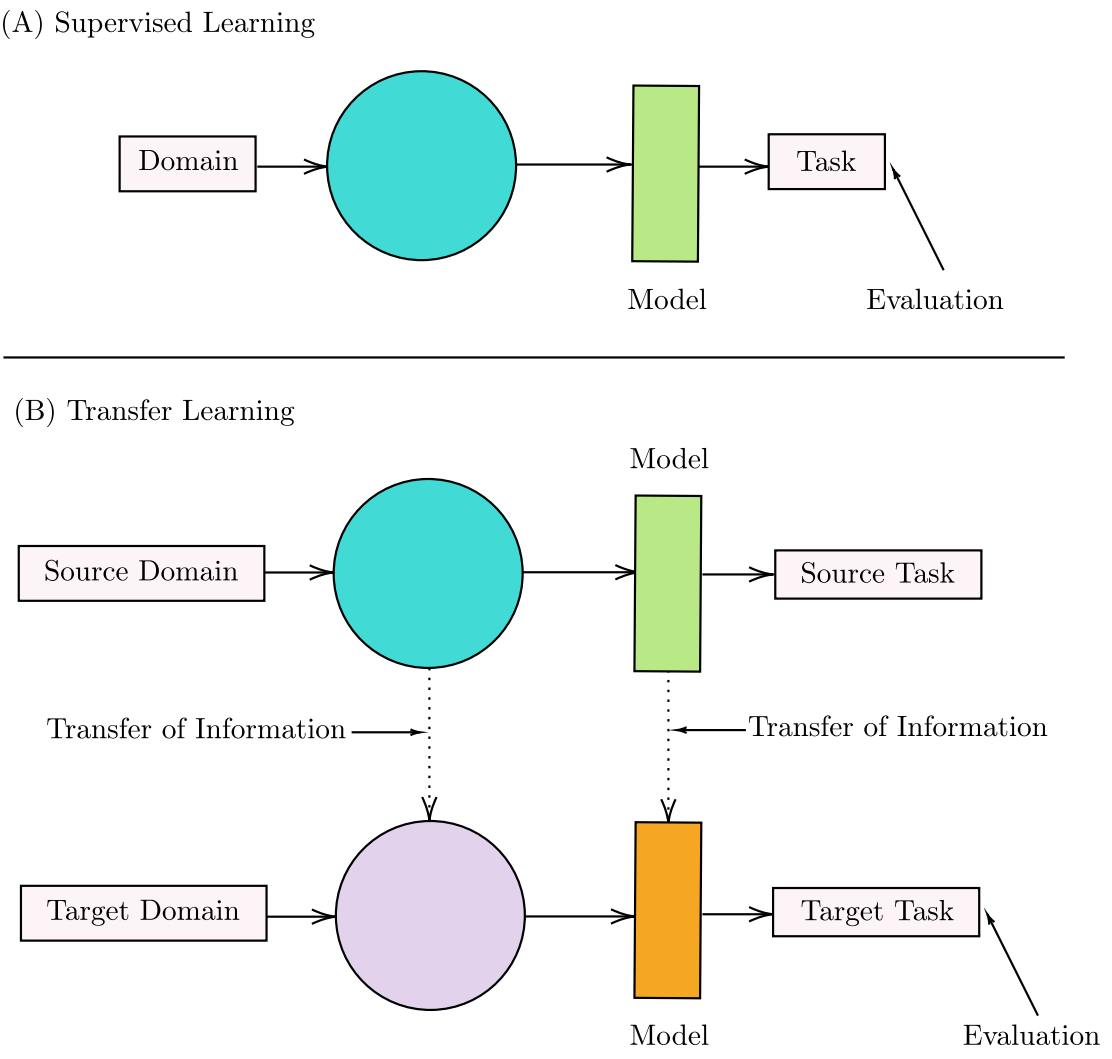

Fig. 1. Visualization of the conceptual idea of transfer learning. A: Traditional supervised learning model for learning a task. B: For transfer learning one needs to distinguish between a source domain and target domain, providing two independent sets of data, and a source task and target task. The purpose of the model learned from the source domain is to enhance the model learned from the target domain and only the performance of this model is of interest. This asymmetry is emphasized by indicating which task is evaluated.

that the topics are different. It could also mean that the marginal distribution is different (e.g., the topic in the source is different from the topic in the target, while the language of the two domains is the same).

Likewise, when the learning tasks are different, i.e., $T_S \neq T_T$ then this implies that either $\mathcal{Y}_S \neq \mathcal{Y}_T$ or $P_S(Y|X) \neq P_T(Y|X)$. Similar to the statements above, also these two conditions are not independent from each other. Specifically, from $\mathcal{Y}_S \neq \mathcal{Y}_T$ follows that also $P_S(Y|X) \neq P_T(Y|X)$ holds, however, from $P_S(Y|X) \neq P_T(Y|X)$ does not follow $\mathcal{Y}_S \neq \mathcal{Y}_T$. Another independent case is given by different prior distributions of the labels, i.e., $P_S(Y) \neq P_T(Y)$. In the literature, the case $P_S(Y) \neq P_T(Y)$ with $P_S(X|Y) = P_T(X|Y)$ is called prior shift and $P_S(Y|X) \neq P_T(Y|X)$ with $P_S(Y) = P_T(Y)$ concept shift [57].

We would like to remark that in the literature the relations discussed above between the different statements are omitted, e.g., [92,130]. Unfortunately, this gives the false impression that those conditions are all independent from each other forming individual cases. As seen above, this is not the case. In Table 1, we summarize the different cases discussed above that follow from the main cases $D_S \neq D_T$ and $T_S \neq T_T$.

In order to visualize the above cases, we discuss now some application examples thereof. For instance, feature divergence describes the situation when the marginal probability of the source domain is different from the target domain $P_S(X) \neq P_T(X)$.This is also known as feature mismatch or domain mismatch [130]. This issue arises when the words are used in one domain more than the other. This takes place because words could have a strong relationship with the domain topic. It may also take place when there are few features shared among the classes. Also, words may have different meanings in the domains. For instance, words like "blur," "fast," and "sharp" are used to describe electronics products, but they don't express

Table 1
A summary of different cases one can distinguish for transfer learning. The provided examples give descriptive instances for the review classification problem. TL: transfer learning.

Main case	Sub-case	Description	Example
D: $D_S \neq D_T$	1: $\mathscr{X}_S \neq \mathscr{X}_T \rightarrow$ $P_S(X) \neq P_T(X)$ Heterogeneous TL	The source and the target domain have a different feature space.	Source domain: Reviews classification about camera products in Germany language. Target domain: Reviews classification about laptops products in English language.
	2: $P_S(X) \neq P_T(X)$ & $\mathscr{X}_S = \mathscr{X}_T$ Homogeneous TL	The source and the target domain have different marginal distribution.	Source domain: Reviews classification about Toys products in English language. Target domain: Reviews classification about laptops products in English language.
T: $T_S \neq T_T$	1: $\mathscr{Y}_S \neq \mathscr{Y}_T \rightarrow$ $P_S(Y\|X) \neq P_T(Y\|X)$	The source and the target domain have different label space.	Source domain: Has two labels("Good", "Bad"). Target domain: Has four labels: "Good", "Perfect", "Disgusting", "Amazing".
	2: $P_S(Y\|X) \neq P_T(Y\|X)$ & $\mathscr{Y}_S = \mathscr{Y}_T$	The source and the target domain have different probability distribution.	Source domain: "small" means positive label Target Domain: "small" means negative label.
	3: $P(Y_S) \neq P(Y_T)$	The labels unbalanced between the source and target	Source domain: has 20 positive labels. Target domain: Has 70 positive labels.

a sensible opinion about books [102]. Another example of a feature mismatch could occur when a word has a negative meaning in one domain but a positive meaning in another. When describing a mobile phone, the word "tiny" has a positive sentiment, but when describing a hotel room, it has a negative sentiment [130]. Another issue may take place when domains have different feature space $\mathscr{X}_S \neq \mathscr{X}_T$. Consider that we have reviews of products written in German in the source and the target contains reviews written in English. Hence, the terms translated from the source document do not exactly represent the words used in the target. One example is the German word "betonen," which Google translator translates into "emphasize" in English; however, the target documents use the English word "highlight" [153]. The difficulty regarding transfer learning may arise when the distribution of labels in the source and the target are different, or when few labels are available in one class, which makes learning from existing data difficult. This problem could take place also if there is no label available in the class of interest in the source.

It is important to highlight that for all transfer learning scenarios above, the source and the target should be related to each other in some form in order to allow the successful transfer of information, because otherwise negative transfer learning may take place [108,130]. In general, negative transfer learning means that the information learned from the source domain has a negative effect on the target task.

For reasons of clarity, we would like to note that transfer learning is similar but different to other forms of learning including multitask learning. In multi-task learning, there is no significant difference between the domains, and the aim is to enhance the output of all of them. However, in transfer learning, which involves using source domain to enhance the output of a target, the target domain is more important than the source [148].

2.3. Categorizations of general transfer learning approaches

So far there is no unique categorization of transfer learning known but different suggestions have been proposed. In the following, we review three main categorizations which are based on learning paradigms [92], properties of the feature spaces [130] and solution-based approaches [92,130]. Based on these, we introduce a new nomenclature of transfer learning that provides a comprehensive categorization.

2.3.1. Transfer learning paradigms

According to [92], transfer learning can be categorized by the way of the learning: inductive learning, transductive learning, and unsupervised learning.

- In inductive transfer learning, source and the target tasks are different while the source and the target domains may or may not be different. Furthermore, at least some labeled target domain data are required.
- In transductive transfer learning, the source task and the target task are the same, however, the source domain and target domain are different from each other. Furthermore, no labeled data are available in the target domain while labeled data are available in the source domain (for a thorough discussion of transductive transfer learning see [85]).
- In unsupervised transfer learning, the source and target tasks are different but related. Because the focus is on related unsupervised learning tasks, e.g., clustering or dimension reduction, no labeled data are available in the source and target domains.

We would like to highlight that in the literature there is no unique terminology about the meaning of unsupervised transfer learning. While in [92] unsupervised transfer learning is the case of having no labeled source domain data and no labeled

target domain data, in [17] it is assumed that labeled source domain data are available but no labeled data for the target domain. Yet another notation is used in [30] by distinguishing between supervised or unsupervised and informed or uninformed. Specifically, the former relates to the presence or absence of labeled data in the source domain, while the latter refers to the presence or absence of labeled data in the target domain. Hence, unlabeled source and target domain data is referred to unsupervised uninformed transfer learning, whereas labeled source and unlabeled target data is supervised uninformed transfer learning.

We would also like to highlight that there is a similar confusion in the literature about the term semi-supervised transfer learning. In [21], semi-supervised transfer learning is the case of having labeled source data and no labeled target data. However, in [17] semi-supervised transfer learning is the case of having abundant labeled source data and limited labeled target data. Comparing this terminology with the one for unsupervised transfer learning discussed above one can see that there is even confusion between these main categories because in [17] having labeled source domain data and no labeled data for the target domain is called unsupervised transfer learning while the same case is called semi-supervised transfer learning by [21].

2.3.2. Homogeneous vs heterogeneous transfer learning

In addition to the above categorization one can distinguish homogeneous transfer learning and heterogeneous transfer learning [130,92]. Homogeneous transfer learning refers to the situation where the source domain and target domain have the same feature space $\mathcal{X}_S = \mathcal{X}_T$. In contrast, heterogeneous transfer learning refers to the scenario where the source domain and target domain have a different feature spaces $\mathcal{X}_S \neq \mathcal{X}_T$. With respect to our indicators given in Table 1 heterogeneous transfer learning corresponds to the case D1.

2.3.3. Solution-based distinctions

A third possible categorization can be given by distinguishing solution-based approaches that describe 'how to transfer'. Specifically, according to [92,130] these approaches can be distinguished as follows:

- instance-transfer
- feature-representation transfer
- parameter transfer
- relational-knowledge-transfer

Instance-transfer approaches are based on re-weighting of instances in the source domain to use them directly together with data from the target domain [21]. That means instance-transfer approaches do not distinguish between training in the source domain and the target domain but combine those data. In general, instances are weighted such that differences in the marginal distributions of source and target are minimized. Such approaches can only be used when $\mathcal{X}_S = \mathcal{X}_T$, hence, they can only be used for homogeneous transfer learning.

Feature-representation transfer approaches do not require the same feature space for source and target domain. Feature-based transfer learning methods build a new feature space in either of the following ways. Asymmetric approaches: They transform the source features to match the target features. Symmetric approaches: They learn a common latent feature space before transforming both the source and target features into a new feature representation.

The parameter transfer methods may be the most simple and intuitive approaches because they share parameters between source and target model. This enables a clear understanding of the transfer learning model.

Relational-knowledge-transfer methods transfer information based on a defined relationship between source and target.

2.3.4. Others

Finally, we would like to mention that the paper by [118] proposed a categorization specifically for deep transfer learning. Their categorization consists of the following four groups:

- instance-based deep transfer learning
- mapping-based deep transfer learning
- network-based deep transfer learning
- adversarial-based deep transfer learning

Instances-based approaches use instances from the source domain with the appropriate weight. Mapping-based deep transfer learning methods focus on mapping instances from two domains into a new data space of greater similarity. Network-based deep transfer learning methods work by reusing the pre-trained parameters of the source domain for the target domain. Adversarial-based approaches find transferable features that are compatible for two domains using adversarial technology.

As one can see, all four categories have a strong similarity to the solution-based transfer learning approaches discussed in Section 2.3.3, which have not been suggested for deep learning but general machine learning methods. This indicates that the above categorization is in fact not limited to deep learning models.

We would like to mention that there are further categorizations of transfer learning, e.g., [157,68]. However, all of these are similar to the above three main categorizations and do not lead to new systematics.

2.4. Comprehensive nomenclature of transfer learning

From the discussion of the different categorizations above, it becomes clear that none of these is complete but each addresses a specific aspect or provides a certain perspective on transfer learning. For this reason, in order to obtain a comprehensive and unique terminology for the various cases and perspectives one needs a different approaches.

It is important to realize that the three main categorization above are independent from each other. That means each describes cases that are not covered by the other two categorization. For this reason, we suggest to introduce a nomenclature of transfer learning that combines the main features of those three categorizations. Specifically, we suggest the following terminology:

$$\text{Terminology} : (\mathrm{A}).(\mathrm{B}).(\mathrm{C}) \tag{1}$$

with

$$\mathrm{C} = \{(\mathrm{C1}(i)) - (\mathrm{C2}(i))\}_i^S \tag{2}$$

for a multi-step learning procedure with S steps. That means, we suggest a nomenclature that is a combination of the following three components:

- A: probability space-based (depending on the properties of the different feature spaces and label spaces; see Table 1)
- B: solution-based (depending on the realization of the model; see Section 2.3.3)
- C1(i): source domain data for step i; see Fig. 2)
- C2(i): target domain data for step i; see Fig. 2)

Here

$$C = \{(\mathrm{C1}(1)) - (\mathrm{C2}(1))\}_i^S = \{(\mathrm{C1}(1)) - (\mathrm{C2}(1)), \ldots, (\mathrm{C1}(S)) - (\mathrm{C2}(S))\} \tag{3}$$

is a set whose components correspond to the pairs $(\mathrm{C1}(i)) - (\mathrm{C2}(i))$ for each step i characterizing the used data whereas S is the total number of steps of a learning procedure. We would like to note that for pure types of data a learning paradigm is entailed.

$$\text{unlabeled data} : D_u = \{(x_i)\}_i^{N_u} \rightarrow \text{ unsupervised learning} \tag{4}$$

$$\text{labeled data} : D_s = \{(x_i, y_i)\}_i^{N_u} \rightarrow \text{ supervised learning} \tag{5}$$

$$\text{partially labeled data} : D_{se} = D_u \cup D_s \rightarrow \text{ semi-supervised learning} \tag{6}$$

That means by specifying the type of data in a learning step, one specified the learning paradigm. Below we will see that the mixing/selecting of data for different learning steps makes this characterization step-dependent and, hence, a local property of a learning procedure. In contrast, we will see that (A) and (B) correspond to global properties of a transfer learning model.

Let's discuss the above nomenclature by starting with the data-dependent component. Since transfer learning requires two different domains, a source domain and a target domain, there are in total 9 different combinations of unlabeled data, labeled data and partially labeled data, as shown in Fig. 2. For instance, the case for unlabeled source data and labeled target data is called (unlabeled data)-(labeled data) transfer learning (an example thereof is BERT [37] - see Section 4.3.1), whereas the case for unlabeled source data and partially labeled target data is called (unlabeled data)-(partially labeled data) transfer learning. We would like to remark that the situation when labeled source data are available regardless of the type of target data and $P_S(X)\neg = P_T(X)$ (with $\mathscr{X}_S = \mathscr{X}_T$) holds, in the literature this is called domain adaptation [131,33] which is a form of transductive transfer learning [94]. In Fig. 2 domain adaptation is highlighted by the purple oval. Furthermore, the situation where we have unlabeled source data and labeled target data is in the literature called self-taught learning [106], a form of inductive transfer learning.

Reviewing the literature one finds that many of the currently used deep transfer learning models are multi-step procedures. That means instead of consisting of one step for learning the parameters of a model the learning is extended over several steps. Furthermore, not every step utilizes the same data but selected subsets of the available data. For this reason, in the above terminology we added information about step i of the model as index. For instance, Stacked Denoising Autoencoders (SDA) [45]) use in the first step all unlabeled data from the source domain and the target domain, while in the second step a classifier is trained using only the labeled data from the source domain (detailed about SDA are discussed in Section 4.1.1).

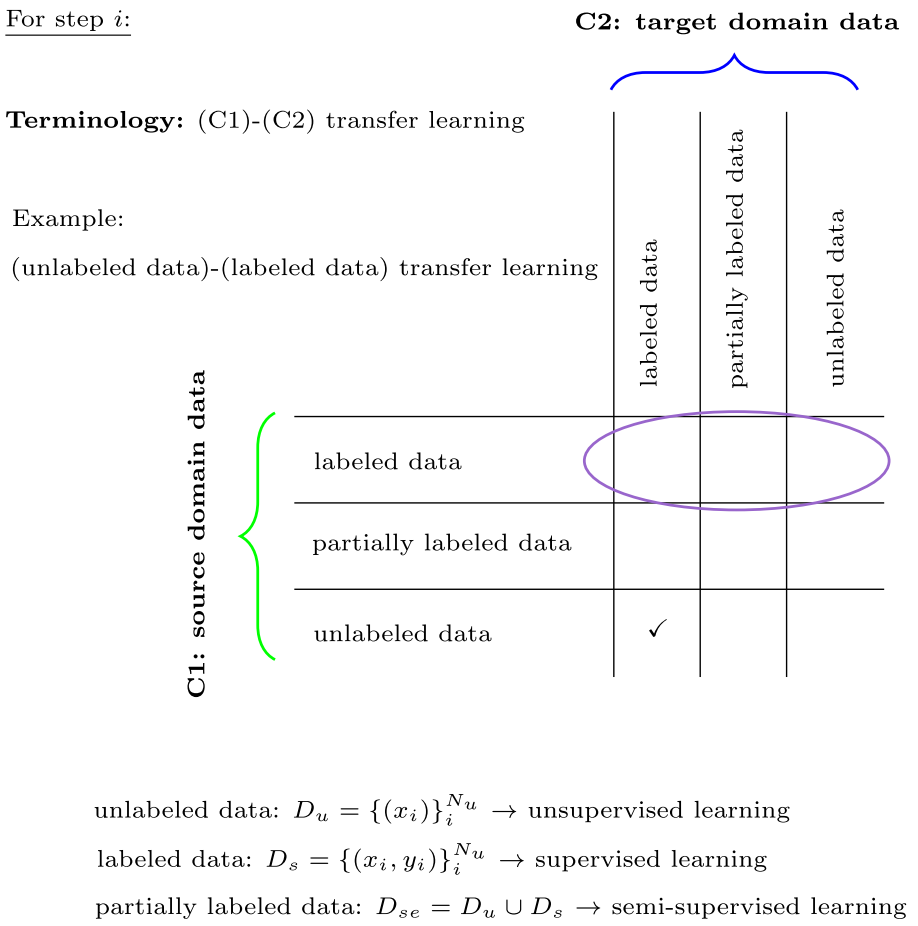

Fig. 2. Combinations between source domain data (C1) and target domain data (C2) for learning a transfer learning model. Depending on the type of data, a learning paradigm is entailed for step i of a multi-step learning model. The purple circle highlights the focus of domain adaptation.

Importantly, this behavior is not unique to SDA but can be observed through out the literature. However, such multi-step procedures lead to additional combinations that need to be considered because the data are not used in one specific way but source and target data can be combined or selected in various different ways for each learning step.

It is important to highlight that a multi-step procedure does no longer allow to conclude, e.g., from given source domain data to a learning paradigm. The reason for this, as discussed for SDA, is that while the source data may be labeled, these data do not have to be used in this form but a selection can be made, e.g., ignoring the labels. Of course this would not be sensible if a model would consist of a one-step procedure because this would limit the amount of information used for the learning of the model. However, for a multi-step procedure this is not the case because other learning steps can utilize the labeled data. Hence, multi-step procedures allow the selection and even mixing of data from different domains without losing information during the learning process. In terms of the notation of a transfer learning model, this complexity is reflected in the combinatorial form of our nomenclature, adding an index to the pairs of source and target data used in step *i*, i.e., $(\mathsf{C1}(i)) - (\mathsf{C2}(i))$ (see Eqn. 3). Conceptually, this means the characterization of the used data is a local property of a multi-step learning procedure because each step *i* can utilize different (combinations of) data.

In contrast to the characterization of the used data, the characterization of the probability spaces (A) and solution-based approach (B) are global properties. The reason for this is that the property of the underlying probability spaces cannot be changed nor effected by the number of learning steps of the model. Also the solution-based approach, e.g., via parameter transfer, is a global strategy defining how to transfer the knowledge from the source to the target. Overall, the combinations of (1) data, (2) properties of data and (3) model approaches, for various learning steps of a models lead to a combinatorial plurality of transfer learning. This underlines that transfer learning is a diverse family of learning models.

Table 2
An overview of text data used by studies analyzing deep transfer learning. All resources are publicly available.

Data set	Domain and Language	Description and Reference
Amazon product reviews [16]	Books, Electronics, Kitchen, DVDs, Videos	Consists of about 340,000 text reviews of different Amazon products. Each review is classified into positive or negative.
	English	[45,23,2,29,128,67,158,63,129,80,152,143,135,149]
Multi language products reviews [102]	Books, DVD, Music	Contains reviews written in four languages, and each language has 4000 reviews.
	English (EN), German (GE), French (FR), Japanese (JP)	[153,133]
Spam mail dataset [15]	Public(u), Private(u1), Private(u2), Private(u3)	The email spam data contain private inboxes and public inbox. Each private inbox consists of 1,250 spam and 1,250 non-spam emails, and the public inbox consists of 2,000 spam and 2,000 non-spam emails.
	English	[73]
20Newsgroup	computer(C), record(R), science(S), talk (T)	Contains approximately 20,000 news article on several subcategories.
	English	[73,29,128,129,25]
SemEval 2015 [99]	Restaurant, Laptop	Contains 1572 review sentences about restaurant and 1907 review sentences about laptop.
	English	[138,26]
Camera[54]	Camera	3770 camera reviews sentences.
	English	[138]
Movie1 [93]	Movie	Includes about 10662 positive and negative reviews about movies.
	English	[138]
Movie2 [116]	Movie	Collections of 9613 positive and negative reviews about movies
	English	[138]
Pathology dataset [134]	Ductal Carcinoma In- Situ(DCIS), Lobular Carcinoma In-Situ (LCIS), In- vasive Ductal Carcinoma (IDC), Atypical Lob- ular Hyperplasia (ALH)	Includes 96.6 k breast pathology reports collected from three hospitals representing aspects of breast disease.
	English	[147]
Yelp	Restaurants	Positive and negative review about overall restaurant.
	English	[147,26]
Hotel review [125]	Value, Room Quality, Check-in Service, Room Service, Cleanliness	Includes a total of around 200 k reviews collected from TripAdvisor.
	English	[147]
Hotel [69]	Reviews	Positive and negative hotel reviews.
	Chinese	[24]
BBN[84]	Sentiment	Contains 1200 sentences from social media posts.
	Arabic	[24]
AFPBB news	Politics, Environment-science-IT, Lifestyle, Sports	52,000 news documents from several categories.
	Japanese	[86]
Livedoor news	Topic news, IT-life-hack, livedoor-homme, sports-watch	Consists of 3000 livedoor news documents
	Japanese	[86]
CoNLL [112]	Organizations (ORG), Locations (LOC), Persons(PER), Miscellaneous (MISC)	Named entity recognition dataset includes 220 K news paper documents.
	English, German, Spanish	[107]
GermEval [11]	News	Named entity recognition dataset consists of 450 k tokens from Wikipedia articles.
	German	[107]
ONB [89]	Historical news	Named entity recognition dataset of Austrian newspaper texts from the Austrian National Library.
	German	[107]
LFT [89]	Historical news	Named entity recognition dataset of nwspaper corpus from Dr. Friedrich Temann Library.
	German	[107]
Amazon reviews[113]	Reviews	Electronics Positive and negative reviews collected by Stanford.
	English	[113,26]
Yelp review	Business Reviews	Positive and negative business reviews.
	English	[113,26]
Chinese medical NER (CM-NER) [127]	Cardiology, Respiratory, Neurology, Gastroenterology,	Named entity recognition corpus contains 1600 de-identified EHRs of hospital from four different specialties in four departments.
	Chinese	[127]

Table 2 (*continued*)

Data set	Domain and Language	Description and Reference
Twitter SemEval 2016 [88]	Review	Positive and negative Twitter review.
	English	[139]
Twitter SemEval 2018 [1]	Review	Positive and negative twitter review.
	English	[139,26]
Ren-CECps [103]	Anger, Expectation, Anxiety, Joy, Love, Hate, Sorrow, Surprise	Contains 1487 documents with each sentence labeled by a sentiment label and 8 emotion labels.
	Chinese	[139]
Chinese corpus[132]	Book, Computer, Hotel	Positive and negative reviews
	Chinese	[80]
Hotel	Reviews	dataset from Xiecheng website containing 2000 positive and 2000 negative samples.
	Chinese	[149]
The notebook [149]	Reviews	Contains 4000 negative and positive reviews collected from shopping website.
	Chinese	[149]
The Weibo [149]	Reviews	Contains 1 K negative and positive reviews collected from COAE 2015.
	Chinese	[149]
Technology product [149]	Reviews	Contains 8000 negative and positive reviews collected from COAE 2011.
	Chinese	[149]
Reuters multilingual dataset [7]	CCAT, C15, ECAT, E21, GCAT, M11	A cross-lingual data containing 11 000 articles from 6 Reuters news categories.
	English, German, French, Spanish, Italian	[152]
Imdb	Movies	Stands for internet movie database consisting of movies information.
	English	[113]
Standford	Movies	Contains 11,855 reviews.
	English	[113]
MIMIC-III	Health records data.	Contains data of hospital admission for adult patients including discharge summaries laboratory measurements, diagnostic codes, and medications.
	English	[140,62]
BioASQ3 dataset	Biomedical data.	Biomedical semantic indexing and question answering.
	English	[140]

3. Text data

Due to the fact that for any machine learning or artificial intelligence method, data assume a central role, in this section, we provide an overview of the text data used for studying deep transfer learning models. Specifically, Table 2 shows a detailed overview of the studied data. The table gives information about the name of the data set, domain, language, description, and studies that utilized the data for their analysis. It is important to note that the vast majority of the text data (18 out of 35) are in the English language. The other datasets are in the Chinese (7), German (6), Japanese (3); two datasets are in French and Spanish and one dataset is in Arabic and Italian. Among the selected articles, the most frequently used data by 14 studies is the Amazon data set. The Amazon data set was created by [16] and it includes reviews about 22 different products. However, four products (DVDs, Books, Kitchen, Electronics) were used in the selected studies. Other studies used publicly available datasets such as Reuters, Yelp review and Twitter SemEval. We would like to highlight that there are four data sets for named entity recognition (CoNLL, GermEval, ONB, LFT, and CM-NER). Furthermore, MIMIC-III data sets provide information about electronic Health Records (eHR).

4. Taxonomy of deep transfer learning models

In this section, we present a visual taxonomy of deep transfer learning models for applications to text data. The taxonomy is shown in Fig. 3. Its main branches are based on the categorization introduced in Section 2.4, i.e., they describe the data of the source domain (C1). For obtaining the remaining branches, we reviewed the literature and identified the dominating architectural principles of the neural networks. Those branches contain also information about distributional assumptions (see A in Section 2.4) and and solution-based approaches (see B in Section 2.4).

Overall, the taxonomy in Fig. 3 is a simplification of our nomenclature introduced in Section 2.4 and a reflection of the currently employed deep learning models and variations thereof. This enables a comprehensive overview of the contemporary literature. A discussion about the simplification is presented in Section 5.

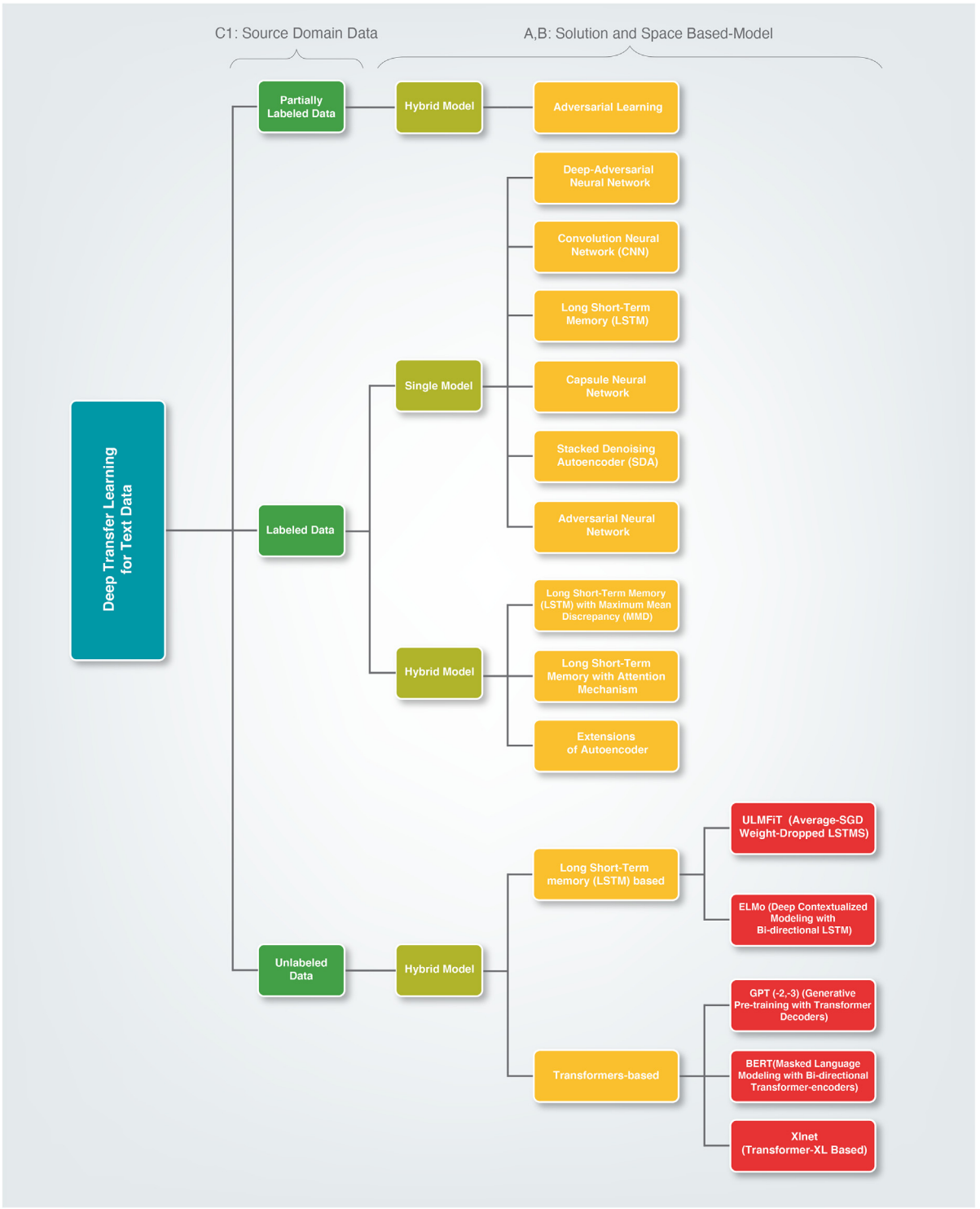

Fig. 3. Taxonomy of deep transfer learning for applications to text data. The two main branches of the taxonomy are based on the categorization introduced in Section 2.4, i.e., they describe the learning paradigm for the data of the source domain (C1). For the characteristics of the target domain (C2) the availability of labeled data is assumed enabling supervised learning of the target task.

4.1. Source domain: Labeled data

In this section, we discuss deep transfer learning approaches that are based on labeled data in the source domain, however the target domain can be labeled and unlabeled both. When both the source and target domains are labeled then such data is applied for multi-task learning where domain and targets are trained simultaneously. For the second case, when only source labels are available, models are applied for transfer learning as domain adaptations. Transfer learning based on labeled source data can be applied for both homogeneous and heterogeneous learning.

In fine-tuning or parameter sharing technique, a network is trained with a large amount of data for learning bias and weights parameters [118]. These weights can be then transferred to other networks to test or trained another model on similar data. Therefore, instead of starting from scratch, the network will use pre-trained weights.Training large models on large datasets need a lot of computing power [5]. Thus, convergence can be accelerated, and network generalization can be improved by training new models with pre-trained weights. Such methods are further subdivided into single and hybrid models. Fig. 4 shows deep transfer learning based on fine tuning. The network is trained with data from the source domain, and then the parameters are transferred into another network which is trained to predict the labels of the target domain.

4.1.1. Single model

Convolutional neural networks: A solution for feature divergence was proposed by [138] and a neural network model was build with two separate CNNs to jointly learn hidden feature representations. Convolutional neural networks learned whether the sentence includes a positive or negative domain sentiment while avoiding prediction for a large number of pivot features. The model was trained on source labeled data and fine-tuned with small number of labeled target data. In their analysis, they showed improvements over SCL and mSDA methods. The approach by [133] was proposed to address the cross-language features challenges by utilizing a parallel corpus. The source classifier was trained to label the parallel corpus, while the target classifier was trained on the labeled set. The paper by [113] discussed that the content of a neural network's embedding layer learned from one dataset can be used for another dataset. They also suggested if labeled data are available in the target dataset, the parameters can be fine-tuned. If labeled data is scarce, the parameters could be left frozen. A very deep convolutional neural network (VDCNN) was used in the paper of [86].In the first step, VDNN was trained on the source dataset. The model then trained on the target data using two ways. The first was to freeze the low layers and share the parameters of upper layers. The second was to share all layers without fixing any layers. The results showed that sharing all layers was more effective in performance than sharing part of them.

A deep transfer learning approach presented in the paper by [141] for Ninth Revision of International of Diseases (ICD-9) by using large number of (MIMIC) as a source dataset. The results indicated that deep transfer learning could improve the classification performance of the Ninth Revision of International of Diseases (ICD-9) of BioASQ3. Based on multi-layer convolutional neural network, [80] introduced transfer learning method based on CNN. The authors constructed a CNN model for extracting features from the source domain and to share the weights among the source and target domain. To train the labeled source dataset, the authors used a convolutional neural network with three convolutional layers and save the trained model structure as well as the weights of layers. When training the target domain dataset, the first three layers remain unchanged, and only the weights of the fully connected layer are fine-tuned with a small part of the labeled target data. The model was evaluated on Chinese and English sentiment and obtained comparable performance against several approaches such as DANN (domain-adversarial neural network) and SCL.

Long Short-Term Memory: In [62], a Long Short-Term Memory (LSTM) network has been extended to transfer learning. Specifically, a LSTM with 6 layers has been studied containing a token embedding layer, character embedding layer, character LSTM layer, token LSTM layer, fully connected layer and a sequence optimization layer. Transfer learning has been realized via parameter transfer that means different combinations of parameter freezing have been studied in a layer-wise fashion. The models used large source data (from MIMIC) and smaller (but still large) target data (from ib2). In [107], a bidirectional LSTM (BiLSTM) has been studied for named entity recognition. Also here parameter transfer has been used for realizing the knowledge transfer.

Capsule network: The model of a Capsule network (CapsNet) has been introduced by [111]. In contrast to CNNs based on scalar-valued feature extractors, capsule networks use vector-output capsules with dynamic routing, whereas a capsule consists in a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity [52]. While CapsNet has been introduced as a supervised learning model in [135] this model has been extended to transfer learning. Specifically, in [135] a deep transfer learning model has been introduced called TL-Capsule. The method consists of four layers: a convolutional layer, a primary capsule layer, a capsule compression layer, and a class capsule layer. The authors argue that capsule networks are able to capture the intrinsic spatial part-whole relations that constitute domain invariant knowledge which helps to transfer knowledge from the source to the target domain. TL-Capsule has been studied for three text classification tasks including cross-domain sentiment classification. As a result they outperformed 14 reference methods, including SCL [16] and DANN [44] (see below the discussion about Adversarial Neural Networks).

Another transfer learning model based on Capsule networks called TransCap was proposed by [26]. TransCap is based on an aspect routing approach allowing to generate sentence-level semantic features. Using TransCap, the transfer between document-level knowledge to aspect-level sentiment classification was studied for several different review classification tasks.

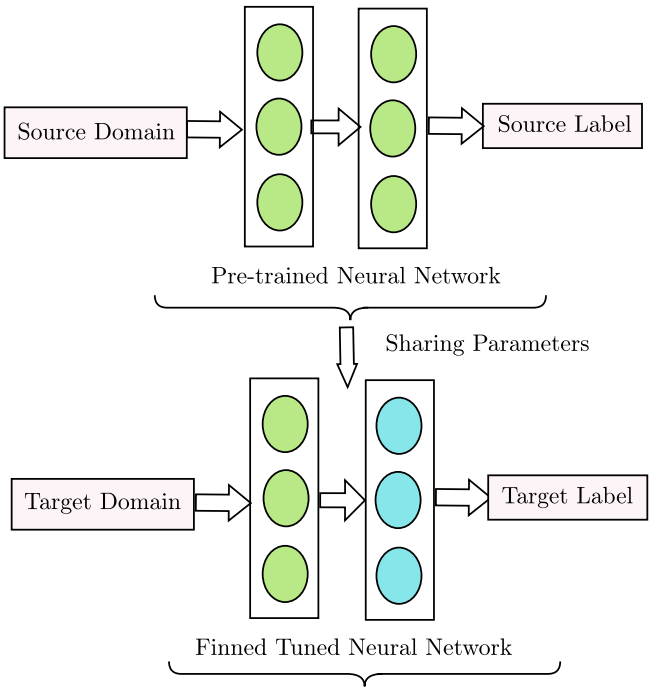

Fig. 4. Deep transfer learning based parameter sharing. The shared parameters are highlighted by the same color.

In Table 3, we show an overview of single models corresponding to the deep transfer learning methods discussed above. In this table, the name before the arrow describes the source domain and the name after the arrow describes the target domain. As one can see, most methods have been studied for the Amazon reviews performing sentiment classification. Furthermore, the error measures used for the avaluation are the accuracy, F1-score, precision and recall.

Autoencoder: In general, an Autoencoder consists of two parts [14,121]: An encoder and a decoder. The encoder maps the input data into a hidden representation, and the decoder tries to reconstruct the input data from the hidden representation. Formally, the encoder is a function $h(x)$ for input x whereas the decoder function results in a decoding given by $r(x) = g(h(x))$. The goal is to minimize the reconstruction error of the input x and the reconstructed input $r(x)$, i.e., $loss(x, r(x))$. As one can see, only unlabeled data are needed to training an Autoencoder. Once an Autoencoder has been trained, one can repeat the above procedure by stacking further Autoencoders while the corresponding Autoencoders are learned layer-by-layer [121]. The output of the hidden layers are frequently used to initialize either a supervised deep neural network or to feed a classifier in the form of a profile vector [49]. The latter allows to construct a new classifier with a deep network architecture.

In [45], a deep transfer learning approach based on **Stacked Denoising Autoencoder** (SDA) has been introduced for performing sentiment classification. For this analysis they used an extension of an Autoencoder called a Denoising Autoencoder (DAE). In contrast to an Autoencoder, a DAE uses a randomly corrupted instance $x\prime$ as input, instead of the uncorrupted input x, to learn a representation [121]. This makes it more difficult to learn the representation when the hidden layer is larger than the input layer because 'simply copying the data' is no longer possible.

The stacking of Denoising Autoencoders works in the same way as for stacking Autoencoders, i.e., the layers are learned in sequential order. This allows to create deep architectures. For transfer learning with labeled source data and unlabeled target data all unlabeled data from the source and the target are used for learning the Stacked Denoising Autoencoder. Finally, the output of the highest encoder layer is utilized as input for a Support Vector Machine (SVM). For training the SVM only the labeled data from the source are used.

It is important to note that the Stacked Denoising Autoencoders are trained with unlabeled data from the source and the target domain at the first step [45]. That means these data are combined into a single data set consisting only of unlabeled data. This step allows the SDA to learn a common invariant latent feature space. The learned features from the final layer are then used as input for learning the task of the source domain, e.g., for sentiment analysis, using only the labeled data from the source domain. For domain adaptation, the transfer loss is defined as the difference between the baseline in-domain error $e_b(T,T)$ and the transfer error $e(S,T)$. The following equation describes the transfer loss,

$$t(S,T) = e(S,T) - e_b(T,T). \tag{7}$$

In Eqn. 7, S and T denote the source and target respectively and $e(S,T)$ is the transfer error corresponding to the classification error of a classifier which is trained for data from the source domain and tested for data from the target domain. Also the baseline in-domain error $e_b(T,T)$ is a classification error of a classifier, however trained with labeled data from

Table 3
Single models for deep transfer learning. The column 'Technique' describes the used model, 'Reference' cites paper(s) that studied the model, 'Source → Target' provides information about the transferred domain, 'Performance' gives information about numerical results and 'Application' indicates the learned task.

Technique	Reference	Source → Target		Performance	Application
Convolution Neural Network	[138]	Movie1→ Laptop Movie1→ Restaurant Movie1→ Camera Camera→ Restaurant Camera→ Laptop Camera→ Movie1 Camera→ Movie2 Restaurant→ Camera	Restaurant→ Laptop Restaurant→ Movie1 Restaurant→ Movie2 Laptop→ Camera Laptop→ Restaurant Laptop→ Movie1 Laptop→ Movie2	Accuracy: 78.7%	sentiment classification
	[133]	EN-Books→ FR-Music EN-Books→ FR-DVDs EN-Books→ GE-Music EN-Books→ GE-DVDs EN-Books→ JP-Music EN-Books→ JP-DVDs EN-DVDs→ FR-Music EN-DVDs→ FR-Books EN-DVDs→ GE-Music	EN-DVDs→ GE-Books EN-DVDs→ JP-Music EN-DVDs→ JP-Books EN-Music→ FR-DVDs EN-Music→ FR-Books EN-Music→ GE-DVDs EN-Music→ GE-Books EN-Music→ JP-DVDs EN-Music→ JP-Books	Accuracy: 81.08%	sentiment classification
	[86]	AFABB → livedoor		Precision: 94% Recall: 94% F1: 94%	text categorization
	[113]	Amazon→ Movie YELP → Movie IMDb → Movie Amazon→ Stanford	YELP→ Stanford IMDb→ Stanford Amazon→ Movie	Accuracy: 82.72%	sentiment classification
	[80]	Book → Hotel Book → Computer Hotel → Book	Hotel → Computer Computer → Book Computer → Hotel	Accuracy: 80.72 % Precision: 81.61 % Recall: 79.29 % F1: 80.42 %	sentiment classification
	[141]	BioASQ3 → MIMIC-III		F1: 48.3 % Precision: 37.1 % Recall: 42.0 %	text categorization
Long Short-Term Memory	[62]	MIMIC → i2b2 2014	MIMIC → i2b2 2016	F1: 97.97%	text categorization
	[107]	CoNLL→ GermEval CoNLL→ LFT CoNLL→ ONB	GermEval→ CoNLL GermEval→ LFT GermEval→ ONB	Accuracy: 75.7%	named entity recognition
Capsule Neural Network	[135]	Reuters single label → Reuters Multi label		precision: 87.4%	text categorization
	[26]	Yelp → SemEval Amazon → SemEval	Twitter → SemEval	Accuracy: 76.6% F1: 70.5%	sentiment classification

the target domain and tested on the target data. Interestingly, it has been found that for a large number of distinct domains, the mean of transfer loss is not informative [45]. For this reason, two new metrics have been proposed for measuring domain adaptation by transfer ratio (Q) and In-domain ratio (I):

$$Q = 1/n \sum_{(S,T)} \frac{e(S,T)}{e_b(T,T)} \tag{8}$$

Here n is number of pairs, i.e., (S,T) where, $S \neq T$ and

$$I = 1/m \sum_{S} \frac{e(T,T)}{e_b(T,T)}. \tag{9}$$

In Eqn. 9, m is the total number of source domains.

Although it has been shown that this method clearly outperforms other transfer learning methods, such as SCL (Structural Correspondence Learning) [17], SFA (Spectral Feature Alignment) [91], and MCT (Multi-label Consensus Training) [64], a major disadvantage of this approach is not to consider the mismatch between the distribution of the source and target

domain. This can lead to a distribution shift between the source and the target domain resulting in problems for domain adaptation giving a poor performance of the model [27]. Another disadvantage of the model is its high computational cost due to its iterative numerical optimization [23].

In Fig. 5, we show the SDA transfer learning model used by [45] from [121] and DAE [122].

In order to improve the above model, in [23] an improved approach consisting of **marginalized Stacked Denoising Autoencoder** (mSDA) has been proposed. In this approach, a linear denoiser is used as basic building block allowing random feature corruptions to be marginalized out. Theoretically, this implies that a model is trained with infinite many corrupted samples for which even a closed-form solution is presented. Therefore the optimization can be performed in a non-iterative way allowing to speed-up the training considerably. Application of mSDA for classifying Amazon reviews showed that the resulting performance is comparable to SDA but much faster.

A method applicable when the feature space of the source and target are different, i.e., $\mathcal{X}_S \neg = \mathcal{X}_T$, has been introduced in [153]. The model, called **Hybrid Heterogeneous Transfer Learning** (HHTL) learns three different mappings: Two homogeneous feature mappings from each unlabeled source and unlabeled target data using mSDA. In addition, they learn a heterogeneous mapping between these features allowing to cross source and target instances. The latter mapping minimizes the difference between homogeneous source features and heterogeneous target features. As a classifier, they train a SVM based on the transformed labeled source data by concatenating also intermediate layers of the homogeneous features. The motivation for the HHTL model was to reduce the bias, e.g., from instance shift or feature mismatch, occurring due to cross-domain variations [153]. HHTL was evaluated for the Amazon review dataset, where English reviews were used as the labeled source domain data and three other languages French (FR), German (GE), and Japanese (JP)) were used as the unlabeled target domain data. Overall, HHTL improved compared to other methods, e.g., mSDA.

For improving mSDA in the case when only unlabeled target data are available, in [29] a regularized version has been suggested. For avoiding overfitting the authors utilize a method by [43] that regularizes intermediate layers with the prediction task. Comparison with mSDA showed an improved performance for the Amazon review data set.

In Table 4, we summarize deep transfer learning methods based on Autoencoder. The information provided is similar to Table 3. As one can see, all studies used the Amazon review data. However, the performance varies between the approaches. We would like to note that all studies applied SDA on sentiment classification. In addition, Autoencoder was applied for news

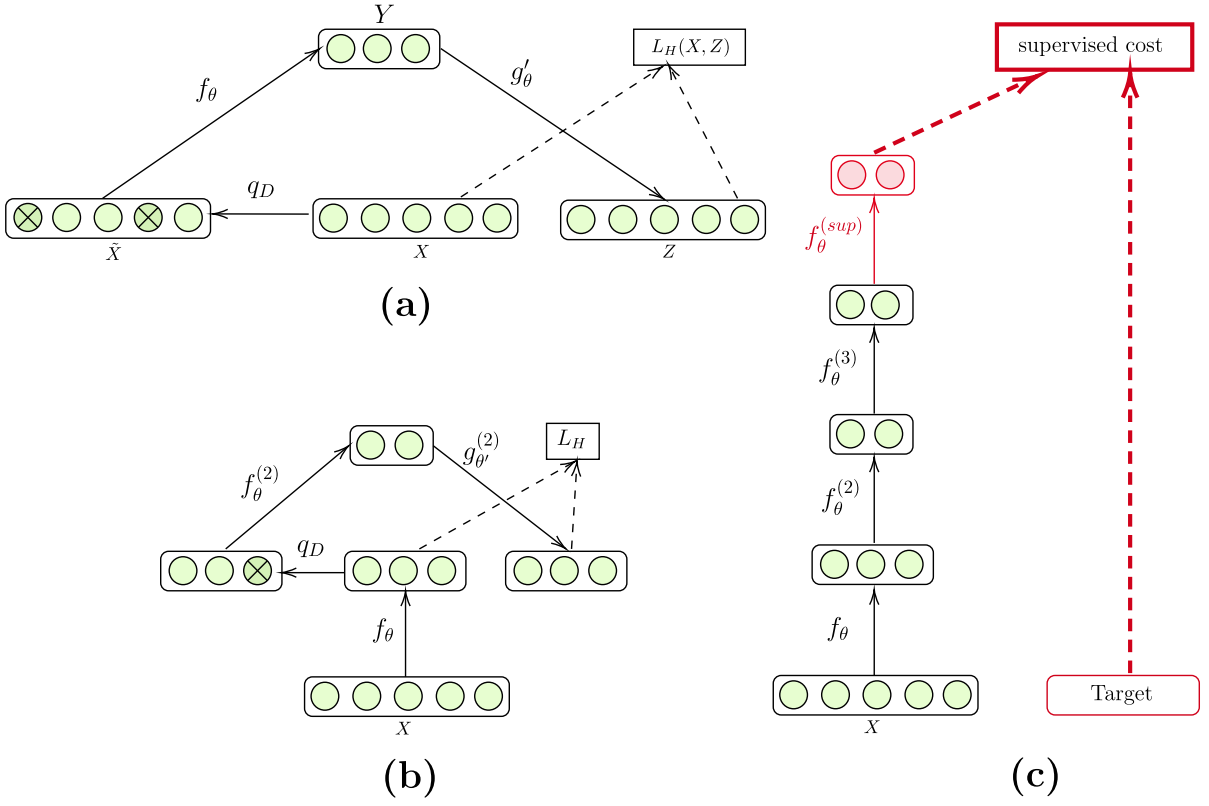

Fig. 5. (a) Denoising Autoencoder with single layer. (b) Two layers stacked Denoising Autoencoder. (c) Fine-tuning of the deep learning model as discussed by [122].

Table 4
Stacked Autoencoder for deep transfer leaning. The column 'Technique' describes the used model, 'Reference' cites paper(s) that stududied the model, 'Source → Target' provides information about the transferred domain, 'Performance' gives information about numerical results and 'Application' indicates the learned task.

Technique	Reference	Source → Target		Performance	Application
Autoencoder	[45]	Kitchen → Books Kitchen → Electronics Kitchen → DVDs Electronics → Books Electronics → Kitchen Electronics → DVDs	Books → Kitchen Books → DVDs Books → Electronics DVDs → Books DVDs → Electronics DVDs → Kitchen	Transfer error: 21:3%	sentiment classification
	[23]	Kitchen → Books Kitchen → Electronics Kitchen → DVDs Electronics → Books Electronics → Kitchen Electronics → DVDs	Books → Kitchen Books → DVDs Books → Electronics DVDs → Books DVDs → Electronics DVDs → Kitchen	Transfer error: 11:5%	sentiment classification
	[153]	EN-Books → FR-Music EN-Books → FR-DVDs EN-Books → GE-Music EN-Books → GE-DVDs EN-Books → JP-Music EN-Books → JP-DVDs EN-DVDs → FR-Music EN-DVDs → FR-Books EN-DVDs → GE-Music	EN-DVDs → GE-Books EN-DVDs → JP-Music EN-DVDs → JP-Books EN-Music → FR-DVDs EN-Music → FR-Books EN-Music → GE-DVDs EN-Music → GE-Books EN-Music → JP-DVDs EN-Music → JP-Books	Accuracy: 74.1%	cross-language sentiment classification
	[29]	Kitchen → Books Kitchen → Electronics Kitchen → DVDs Electronics → Books Electronics → Kitchen Electronics → DVDs	Books → Kitchen Books → DVDs Books → Electronics DVDs → Books DVDs → Electronics DVDs → Kitchen	Accuracy: 81.32%	sentiment classification
Extensions of Autoencoder	[128]	Kitchen → Books Kitchen → Electronics Kitchen → DVDs Electronics → Books Electronics → Kitchen Electronics → DVDs	Books → Kitchen Books → DVDs Books → Electronics DVDs → Books DVDs → Electronics DVDs → Kitchen	Accuracy: 67.47%	sentiment classification
		Computer → Recording Computer → Science Computer → Talk	Recording → Science Recording → Talk Science → Talk	Accuracy: 78.26%	text categorization
	[158]	Kitchen → Books Kitchen → Electronics Kitchen → DVDs Electronics → Books Electronics → Kitchen Electronics → DVDs	Books → Kitchen Books → DVDs Books → Electronics DVDs → Books DVDs → Electronics DVDs → Kitchen	Accuracy: 78.0%	sentiment classification
	[151]	EN-Reuters → FR-Reuters EN-Reuters → GE-DVDs	EN-Reuters → It-Music EN-Reuters → SP-DVDs	Average accuracy: 79.8%	sentiment classification
	[74]	Kitchen → Books Kitchen → Electronics Kitchen → DVDs Electronics → Books Electronics → Kitchen Electronics → DVDs	Books → Kitchen Books → DVDs Books → Electronics DVDs → Books DVDs → Electronics DVDs → Kitchen	Accuracy: 85.99%	sentiment classification
		Comp → Rec Comp → Sci Comp → talk	Rec → Sci Rec → talk Sci → talk	Accuracy: 94.55 %	text categorization
		Public → User1 Public → User2 Public → User3	User1 → Public User2 → Public User3 → Public	Accuracy: 90.64%	spam classification

categorization and for spam classification. Autoencoder was also proposed to solve cross language features as it was applied for cross language features classification. The error metrics used by the studies were: accuracy, precision, recall, F1-score and transfer loss. Transfer loss was only used by two studies.

Adversarial Neural Network: Recently, adversarial learning gained a lot of attention in transfer learning [95,4]. In [44], a model has been introduced based on adversarial learning. Specifically, the model jointly learns a feature representation and two discriminative classifiers, one for class label prediction and one for predicting the domain of instances. The underlying idea is to learn a representation that cannot discriminate between the origin of the instances yet minimizes the risk of the labeled source data motivated by the domain adaptation theory by [10]. Importantly, for accomplishing this, the model uses the labeled source data and the unlabeled target data simultaneously. That means all parameters are learned in one step. The architecture of their model, called **Deep-Adversarial Neural Network** (DANN), consists of three main components: deep feature extractor (green), label predictor (blue) and domain classifier (red). Fig. 6 shows an overview of these components.

The DANN consists of two classifiers. The first classifier is called a discriminative classifier which is trained to predict task-specific class labels. The second classifier is called a domain classifier. This classifier is trained to predict whether an instance comes from the source domain or the target domain. For the training process, the input maps to the hidden layers. This operates as feature extractor $G(.;\theta_f)$. The second component of the network, label predictor $G(.;\theta_y)$, discriminate class labels using feature extractor, $G(.;\theta_f)$, for an efficient domain adaptation the model is trained in an adversarial manner where the extracted features $G(.;\theta_f)$ of input data from the two domains are trained in such a way that $G(.;\theta_d)$ cannot recognize the domain of the extracted features.

Applications of DANN to 12 different sentiment classification tasks (based on Amazon reviews) showed that overall the model achieves good classification accuracy compared to benchmark models. Interestingly, it has been shown that using a Marginalized Stacked Denoising Autoencoders (mSDA, see Section 4.1.1) as representation learner together with a (shallow) DANN allows to further improve the performance.

The paper by [24] proposed an adversarial approach for cross-language features challenges. The approach transferres information learned from rich-source language data to low-source language. The network uses two classifiers: a sentiment classifier and a language adversarial classifier. Both classifiers use input from a shared feature extractor. This way hidden representations are learned. The analyses were conducted on English as a rich-resource language and Chinese and Arabic as low-resource language targets.

We just briefly want to mention that there are also adversarial learning approaches that transfer from more than one source domain such as the one suggested in the paper of [150].

Table 5 shows analyses of deep transfer learning methods based on adversarial learning. The information provided is similar to Table 3. The majority of studies are about sentiment classification of Amazon reviews. Interestingly, all studies used the accuracy as performance measure.

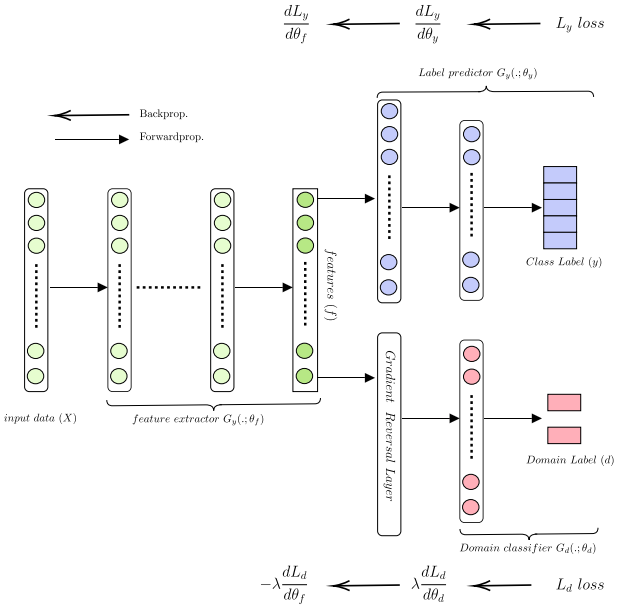

Fig. 6. Architecture of the Deep-Adversarial Neural Network (DANN). The DANN consists of three main components: deep feature extractor (green), label predictor (blue) and domain classifier (red).

Table 5
Adversarial models for deep transfer leaning. The column 'Technique' describes the used model, 'Reference' cites paper(s) that studied the model, 'Source → Target' provides information about the transferred domain, 'Performance' gives information about numerical results and 'Application' indicates the learned task.

Technique	Reference	Source → Target		Performance	Application
Domain Adversarial Neural Network	[44]	Kitchen → Books	Books → Kitchen		
		Kitchen → Electronics	Books → DVDs		
		Kitchen → DVDs	Books → Electronics	Accuracy: 82%	sentiment classification
		Electronics → Books	DVDs → Books		
		Electronics → Kitchen	DVDs → Electronics		
		Electronics → DVDs	DVDs → Kitchen		
	[24]	English → Chinese		Accuracy: 58.7%	sentiment classification
		English → Arabic		Accuracy: 75.6 %	sentiment classification
	[150]	(Kitchen + Dvds + Electronics)→ Books (Kitchen + books + Dvds)→ Electronics (Kitchen + Electronics + books)→ DVDs (Electronics + books + Dvds)→ Kitchen		Accuracy: 81.32 %	sentiment classification
Adversarial Learning	[147]	Value→ Restaurant	Cleanliness→ Restaurant		
		Room→ Restaurant	Checkin→ Restaurant	Accuracy: 87.3%	sentiment classification
		Service→ Restaurant			
		DCIS→ LCIS	IDC → DCIS		
		DCIS→ IDC	IDC → LCIS		
		DCIS→ ALH	IDC → ALH	Accuracy: 91.2	Text categorization
		LCIS→ DCIS	ALH → DCIS		
		LCIS→ IDC	ALH → IDC		
		LCIS→ ALH	ALH → LCIS		
	[67]	Kitchen→ Books	Books→ Kitchen		
		Kitchen→ Electronics	Books→ DVDs		
		Kitchen→ DVDs	Books→ Electronics	Accuracy: 85.40%	sentiment classification
		Electronics→ Books	DVDs→ Books		
		Electronics→ Kitchen	DVDs→ Electronics		
		Electronics→ DVDs	DVDs→ Kitchen		
	[66]	Kitchen→ Books	Books→ Kitchen		
		Kitchen→ Electronics	Books→ DVDs		
		Kitchen→ DVDs	Books→ Electronics	Accuracy: 86.6 %	sentiment classification
		Electronics→ Books	DVDs→ Books		
		Electronics→ Kitchen	DVDs→ Electronics		
		Electronics→ DVDs	DVDs→ Kitchen		

4.1.2. Hybrid model

In this section, we discuss transfer learning models that either combine different models or utilize an additional mechanism to form a new model. Hence, such models are no longer a single model by a hybrid model.

Long Short-Term Memory with label-aware maximum mean discrepancy: The paper by [127] introduced a label-aware double transfer learning framework (La-DTL) and applied it to Chinese medical Named Entity Recognition. La-DTL is based on a text representation learned by Bi-LSTM networks and maximum mean discrepancy (MMD). In general, MMD is a distance measure widely used in machine learning, including transfer learning, to measure the distributional distance between the source domain and the target domain [47]. Formally, it is a two-sample statistical hypothesis test which tests the equality of two distributions based on the observed samples. It measures the difference of the mean values of a smooth function from the samples drawn from the two domains. The authors evaluated the approach on a 10 K corpus collected from a Chinese hospital from four departments: Neurology, Cardiology, Respiratory, Gastroenterology. The proposed approach combines Bi-LSTM with Conditional Random Fields (CRF) and utilizes a label-aware maximum mean discrepancy

metric (La-MMD) for sharing feature representations of hidden layers (H) and a $L2$ constraint on the CRF layers for parameter transfer. In this model, the input data is converted into a sequence of embedding vectors and sent to the Bi-LSTM to obtain the contextual information sequentially in both directions. The Bi-LSTM encodes the information into fixed-length two–hidden vector layers (H) each for source and target domain. The hidden vector layers are connected with domain constrained Conditional Random Fields (CRF) layers of the source and target domain. In general, the CRF is a probabilistic modeling framework that aims to predict the label sequence of input data [59]. The hidden layers of the source and target domain are optimized by label-aware maximum mean discrepancy (La-MMD) to reduce the domain discrepancy for transferable feature representations. However, the parameter sharing of the CRF layer may not yield optimal results for transfer learning when the source and target data have too diverse distributions. In such a case, reducing the Kullback–Leibler (KL) divergence $D_{KL}(P_S(y|H)||P_T(y|H))$ (y is a sequence of labels, H is a hidden vector sequence) is not manageable. Therefore the shared parameters of the CRF layers are optimized by reducing the upper bound ($L2$ constraint) of the KL-divergence. The training of La-DTL is performed with a mini-batch that includes training data from both domains.

Long Short-Term Memory with attention mechanism: Finally, a number of methods were introduced using LSTM networks with an attention mechanism. In general, an attention mechanism maps important features from the input sentence and assigns higher weights to those features.

An attention mechanism has been effectively applied in various applications such as relation classification [155] and sentiment classification [126]. A deep transfer learning model was suggested by [139] to improve the performance of multi-label emotion classification. The model consisted of two LSTM layers named shared Bi-LSTM and target-specific Bi-LSTM. The shared Bi-LSTM layer extracts the shared features between source and target for sentiment and emotion classifications. The target-specific Bi-LSTM extracts the specific emotions which are specific for the emotion classification task. The model consisted of two base models, which are attention-based Bi-LSTM for the source domain for sentiment classification, another Bi-LSTM is for the target domain for emotion classification. The attention mechanism added in Bi-LSTM assign weights, α_s and α_t, that pays more attention to general sentiments. The target input layer is connected with the source-domain Bi-LSTM layer for a shared representation, extracting shared sentiment features for sentiment and emotion classification tasks. The hidden representation of shared space and the target space assign higher weights to frequently occurring words related to sentiments and lower weights to less frequent words for emotions. Therefore allowing the model to pay attention to both emotions and sentiments, the authors proposed a dual attention transfer approach for computing attention-specific weights ($\alpha_s = f(h_s, z)^s$) in a shared space and emotion-specific weights ($\alpha_t = f(h_t, z, \alpha_s) : h_t = \{h_1, h_2 \dots h_n\}$ is a hidden state vector and z is the summary vector of the final hidden state) in target space by sending attention weights of the shared domain (α_s) as inputs to compute attention weights of the target domain. The model's training is performed by alternating an optimization approach by using mini-batches of source data and target data that update parameters of the source domain Bi-LSTM section and all parameters of the whole model alternatively.

Another approach based on an attention mechanism was proposed by [144]. The attention mechanism helped the model to share sentences and aspects for better transferring information across the domains. The approach consists of two attention networks, one is used to classify common features, and the other is used to extract information from aspects. Extensive experiments showed that the model had better performance compared to other methods including DANN and mSDA. The interactive attention transfer network (IATN) consists of S-net and A-net LSTM networks for sentences and aspects. The S-net and A-net, first, obtain word and aspect embeddings from a pre-trained model utilized as input to LSTMs, which transforms sentence and aspect embeddings to hidden semantic state layers (h_s) and aspect hidden states (h_a) for all aspects in a sentence. In the layer for interactive representation between sentence and aspect, a non-linear pooling method is applied for reducing the feature space and preserving crucial features. The pooling layer computes sentence pooling (h_s^p) and aspect pooling (h_a^p) vectors. In order to consider the effect of aspects on sentence for representing final sentiment features, the interactive word sentence attention weight vector is defined as the function of encoded features of sentences and pooling layer of aspect layer, i.e., $\alpha_i = f(h_s^i, h_a^p)$. Similarly, an aspect attention vector is a function of encoded features of aspects and a pooling layer of a hidden representation of a sentence, i.e., $\beta_i = f(h_a^i, h_s^p)$. The final sentence and aspect representations are defined as, $S_r = \sum_{i=1}^{n+1} \alpha_i h_s^i, A_r = \sum_{i=1}^{n+1} \beta_i h_a^i$. The final layers of IATK are domain classifier and sentiment classifier that are connected by a shared feature space. The objective function for model training to optimize parameters for individual attention learning contains domain classification loss (negative), sentiment classification loss and optimizes both loss functions simultaneously for interactive attention learning.

In Table 6, an overview of hybrid models is shown discussed in this section. The information provided is similar to Table 3.

Extensions of Autoencoder: In this section, we discuss models that combine an autoencoder either with other models or mechanisms. In [128] a Deep Nonlinear Feature Coding (DNFC) model was proposed that has two advantages over the mSDA. First, it conducts a minimization of domain divergence by using the Maximum Mean Discrepancy (MMD) [47] and, second, it exploits the nonlinearlity of data by kernelization. This is done by adding an mDA encoder at the first layer of DAE (if one compares with weiss uses two matrix Q1 and Q2, whereas Q2 uses the MMD measure). The learning process measures the discrepancy between the distribution of the source and the target domain at each layer of the deep feature space and minimizes the distance of MMD during the learning.

Another hybrid method based on Autoencoders was proposed in [158]. The idea of their approach is to combine Structural Correspondence Learning (SCL) and Autoencoders. SCL was introduced by [17] for predicting pivot features in the source and

Table 6
Hybrid models for deep transfer learning. The column 'Technique' describes the used model, 'Reference' cites paper(s) that studied the model, 'Source → Target' provides information about the transferred domain, 'Performance' gives information about numerical results and 'Application' indicates the learned task.

Technique	Reference	Source → Target		Performance	Application
Long Short-Term Memory with maximum mean discrepancy (MMD)	[127]	Cardiology→ Respiratory	Neurology → Respiratory		
		Cardiology → Neurology	Neurology → Cardiology	Average accuracy: 71.15%	entity recognition
		Cardiology → Gastroenterology	Neurology → Gastroenterology		
		Respiratory → Neurology	Gastroenterology → Respiratory		
		Respiratory → Cardiology	Gastroenterology → Cardiology		
		Respiratory → Gastroenterology	Gastroenterology → Neurology		
Long Short-Term Memory with attention	[139]	SemEval 2016 → SemEval 2018		Accuracy: 58.3% F1: 54.4%	sentiment classification
	‚144]	Kitchen→ Books	Books→ Kitchen		
		Kitchen→ Electronics	Books→ DVDs		
		Kitchen→ DVDs	Books→ Electronics	Average accuracy: 85.9%	sentiment classification
		Electronics→ Books	DVDs→ Books		
		Electronics→ Kitchen	DVDs→ Electronics		
		Electronics→ DVDs	DVDs→ Kitchen		

the target domain. The purpose of SCL is to identify similarities between features of different domains by modeling their correlations with pivot features. The pivot features are the main features that commonly occur in the source and target domain and exhibit importance for a task. SCL works by first obtaining the pivot features and then utilizing them to find common, low-dimensional features. In contrast, to other methods based on Autoencoders, in [158] Autoencoders do no directly receive input from the available instances but receive a low dimensional representation of non-pivot features learned by SCL.

The model by [151] is for heterogeneous transfer learning. Specifically, the authors introduce a model called Deep semantic mapping model for Heterogeneous multimedia Transfer Learning (DHTL). For minimizing cross language variations between domains, DHTL integrates a deep neural network with a canonical correlation analysis (CCA) in each layer. Canonical correlation analysis (CCA) is a common statistical technique for determining maximally correlated linear projections of two random vectors [8]. CCA is used in DHTL to maximize correlation in order to optimize a commonly correlated features in the source and target domain. The model consists of several layers-based auto-encoders with CCA to train domain-specific and representation-shared networks at the same time. First, the network is per-trained on the co-occurrence data using shared semantic mapping. Then, in backpropagation, the top layer correlation matching between domains is used to fine-tune the entire network to obtain the unified deep semantic mapping. Experiments show that the proposed model outperforms a number of existing models in terms of classification accuracy.

In [73,74], several shortcomings of MMD have been highlighted, including:

1. The kernel-based MMD identifies only a local generalization but is ineffective for global nonlinearities.
2. Predefined kernels are not optimal for maximizing the two-sample matching power of MMD.
3. The scaling of MMD is non-efficient for large data sets.
4. Large domain discrepancy when the model is estimating weights of target features which are not existed in the original (source) feature representation.

In general, a good representation disentangles the factors of variation between the domains and preserves information about the data [13]. That means for case 4 that the disentanglement of hidden factors of variation can increase the cross-domain distribution discrepancy. Hence, invariant factors learned by deep learning model would reduce the domain discrepancy. Overall, this would lead the target error becoming statistically unbounded.

In [74] a generalized framework for domain adaptation has been proposed to jointly learn the transferable representation and classifer. The model consists of two components, a Transfer Denoising Autoencoder (TDA) and a Transfer Deep Network (TDN) (see below). Both models utilize a multi-kernel MMD method (MK-MMD), a nonparametric test statistic performing a two-sample comparison in linear time using a B-test [74], for avoiding the problems of MMD discussed above. This way multiple TDAs can be stacked to achieve a deep network. Overall, the TDA is used as an unsupervised pre-training step of the TDN. The TDN is a multilayer perceptron regularized by the MK-MMD on all hidden layers. This performs a supervised fine-tuning step of the resulting model. Numerical analysis showed that TDN outperformed other state-of-the-art methods in sentiment classifications, email spam filtering and newsgroup categorization.

4.2. Source domain: Partially labeled data

In this section, we discuss deep transfer learning approaches that are based on partially labeled data in the source domain, i.e., the source domain contains labeled and unlabeled data.

4.2.1. Hybrid model

Adversarial learning: In order to tackle cross-aspect and cross-domain adaptation, the study by [147] introduced an Aspect-augmented Adversarial Network (AAN). ANN consists of five components. The first component is a sentence embedding that uses a CNN to obtain sentence-level vector embeddings by minimizing reconstruction loss. The second component is relevance prediction, which predicts the relevance of each sentence's relevance to an aspect by minimizing the error between labeled relevance and predicted relevance. The third component is document encoding, which generates a document feature vector by summarizing each sentence's relevance score with the sentence embedding. The document feature vector is connected with the fourth component, the transformation layer. Together with a regularization term the transformation layer maps the document features to domain invariant features. Finally, the transformation layer is the input to the last components, the label predictor and domain classifier. The domain classifier is a feed forward network that functions as the adversary for domain invariance. Overall, a joint objective function is used to learn the parameters of the Aspect-augmented Adversarial Network by combining the individual optimization functions of the five components, i.e., word reconstruction, relevance label optimization, transformation layer regularization, source class label loss, and domain adversary minimization.

An attention-based extension of Adversarial networks was proposed by [67]. This study introduced an Adversarial Memory Network (AMN), which can capture pivots of features shared in the source and target domain. An attention mechanism discovers important features from the input sentence and assigns higher weights to these features. Attention mechanisms have been effectively applied in various applications such as relation classification and sentiment classification [126]. The success behind the attention mechanism is that a low-level position has its importance for a high-level representation [155]. Overall, the approach in [67] consists of two networks sharing parameters. The first network is used for sentiment classification and the second network for domain classification. Both networks are trained jointly to select features. Later, [66] improved the AMN model by using a hierarchical attention network to capture the pivot and non-pivot features.

4.3. Source domain: Unlabeled data

In this section, we discuss deep transfer learning approaches that are based on unlabeled data in the source domain.

In general, unlabeled data imply that unsupervised learning methods need to be used. Compared with labeled data, unlabeled data provide less information about the underlying distribution from which the data are drawn because the labels are missing. This translates directly to the learning paradigm making unsupervised learning less powerful than supervised learning. Interestingly, self-supervised learning provides a strategy for softening this restriction [78]. Specifically, self-supervised learning assigns, based on some rule, labels to the unlabeled data. This allows to employ supervised learning methods for dealing with unlabeled data. It is clear that not every 'rule' leads to sensible labels. For instance, methods like BERT and GPT (see below for details) utilize the sequential character of text data for predicting the 'next word' in a sentence. This corresponds to a masking of tokens [78]. Importantly, self-supervised learning methods still belong to an unsupervised source domain since the labels are not naturally given.

Self-supervised learning provides two approaches that can be used for the pretraining of a model when only unlabeled data are given; Generative learning and discriminative learning. In generative learning, a probability distribution of possible outputs is generated based on the input [83]. For example, when a masked token is given, the model would try to generate all possible tokens that could fit in the masked position based on the dataset, which can be very time consuming. In contrast, in descriminative learning, the models learn to distinguish between the created labels or classes of the input [78]. An example would be learning to generate separate embeddings for each token in the search space.

In this section, hybrid deep transfer learning approaches based on unlabeled source data are discussed. In the following, we distinguish between two main types of models: LSTM-based and Transformer-based language models which are currently the dominating models.

4.3.1. Hybrid model

Linguistic Feature Modelling. Currently, there is much interest in building generalized linguistic models that are able to represent features and grammars of a given natural language. A linguistic model or a language model (LM) is defined as a probabilistic model that predicts the next token of a given sequence of tokens [46]. This corresponds to learning a conditional probability distribution corresponding to [78]

$$p(x_i|x_1.x_2,\ldots.,x_{i-1}) \tag{10}$$

for a sequence $x_1, x_2, \ldots., x_i$.

This is achieved by pretraining a model to predict the next token given a previous sequence of tokens or masked token given a windowed sequence of tokens. Generally, to pretrain linguistic models, substantially large text corpora is required. The language model can then be used as the source model for transfer learning in multiple NLP target tasks, with little to no finetuning.

In general, language models belong to the class of sequential transfer learning (STL), which is a branch of inductive learning paradigm [110]. Importantly, the models are trained in sequence and separately for the source and target task, hence the name STL. That means, no mixing or selecting of the data is involved. The aim is to transfer information. i.e., feature representations (in ELMo) or parameters (in BERT, GPT, etc.) from the source model to the target model, and thereby enhance the performance of the target task. In the **pretraining stage**, a linguistic model is trained in self-supervised manner with copious amounts of unlabelled data while in the **fine-tuning/adaptation stage**, the language model is fine-tuned to the target task.

Recurrent Neural Network based models

Recurrent Neural Network (RNN) based language models that transfer parameters or feature representations to a target tasks were some of the earliest language models used. These models are based on different implementations of RNNs.

ELMo: The ELMo (Embeddings from Language Models) model is a bidirectional language model (bi-LM) based on a 2-layer bidirectional LSTM [96], introduced by Allen Institute of AI. The model is pretrained with *One Billion Words Benchmark* dataset by Google [22]. In contrast to other language models discussed later, ELMo is a feature-representation transfer model, and as such does not modify any of the pretrained parameters in the fine tuning stage. Instead, internal states of the two LSTM layers are linearly combined to derive contextual character-wise embeddings for target tokens, and a secondary neural network is trained with the embeddings to apply for the target task. As a result, ELMo can be either conveniently integrated into an existing system for providing character embeddings, or used as the primary model with an additional layer for fine-tuning [97]. This is possible because, while ELMo produces embeddings, these are different to earlier text embedding approaches such as word2vec or GloVe, as the ELMo embeddings are extracted from a language model, pretrained bidirectionally [96].

ULMFiT: Universal Language Model Fine-Tuning (ULMFiT) is a generalized transfer learning approach proposed by [53]. The model uses averaged stochastic gradient descent weight-dropped LSTMs (AWD-LSTM) [81]. AWD-LSTM is a simple 3-layer LSTM that uses *DropConnect* [124], a regularization approach where the neural networks set weights of inputs to zero, as opposed to *dropout* where the output activation is set to zero. DropConnect reduce over-fitting by randomizing which inputs are used within the LSTM units, without disrupting output hidden states. These LSTMs also use Non-monotonically triggered average stochastic gradient descent (NT-ASGD) as the optimization method. The AWD-LSTM is pretrained in an unsupervised manner with the WikiText-103 corpus [82] consisting of $28,595$ general Wikipedia text articles. The fine-tuning of the target task is performed in a supervised manner using discriminative fine-tuning and slanted triangular learning rates. The learning model has been evaluated for sentiment analysis and text classification.

Transfer Learning based Emotion Recognition in Conversations (TL-ERC) [51] is a very recent deep transfer learning model for emotion mining by considering contextual information. The model uses generative pretraining to build an end-to-end conversational model that can be fine-tuned to identify emotions in a dialog. This compensates for the scarcity of labelled data of the task, and achieves better validation results surpassing previous models such as c-LSTM + Att [100], Memnet [117] and CMN [50]. Interestingly, this is achieved with significantly less epochs, e.g., compared to DialogueRNN [77].

The base architecture used by TL-ERC is a seq2seq Hierarchical Recurrent Encoder-Decoder (HRED) [114], a classic deep learning architecture used, e.g., for building dialog models [51]. The source architecture of TL-ERC consists of three sub-modules: an encoder RNN for encoding the input sentence, a context RNN to model context of the conversation (shared within the dialog model), and a decoder RNN to output the response to the input sentence. The source context model parameters and sentence encodings (HRED and BERT-based - see below) are then transferred to the target task. The target model is fine-tuned using a small labelled dataset with conversational sentences as input and emotions as labels. The authors highlight that using BERT-based encodings performs better in the target model compared to HRED encodings. The source task uses the Cornell movie dialog corpus [35] and the Ubuntu dialog corpus [75] for pretraining and the target task is fine-tuned for three labelled datasets, namely, IEMOCAP [20] (labelled for anger, happiness, sadness, neutral, excitement, and frustration, SEMAINE [79] (labelled for valence, arousal, power, and expectancy) and DailyDialog [65] (labelled for anger, happiness, sadness, surprise, fear, disgust and no-emotion).

Transformer-based models Transformers are sequential transduction models introduced primarily for machine translation tasks [120]. The underlying architecture uses attention to keep track of long-chain dependencies in text without the need for using Recurrent Neural Networks. The original Transformer model consists of 6 stacked encoder-decoder blocks with each encoder and decoder containing a feed-forward neural network (FFNN) and multi-head self-attention mechanism as illustrated in Fig. 7A and Fig. 7B. This differs from Autoencoder, which contain a single neural network performing both encoding and decoding. As mentioned above, Transformers were specifically designed for translation tasks requiring a mapping between two entirely different feature spaces, e.g., from one language to another. The architecture has since been used as stacked encoder-decoder pairs, stacked encoders-only and stacked decoders-only configurations, as best applicable for the intended task. The number of stacked encoder/decoder blocks can vary from $N = 12$ (BERT architecture) to $N = 24$ (XLNet architecture).

BERT: Bidirectional Encoder Representations from Transformers (BERT) is an auto-encoding language model trained using stacked encoder blocks from Transformers (Fig. 7 A) with a masked language modeling (MLM) to learn embeddings bidirectionally [37]. The model was introduced by Google and is pretrained on large unsupervised text corpora using two

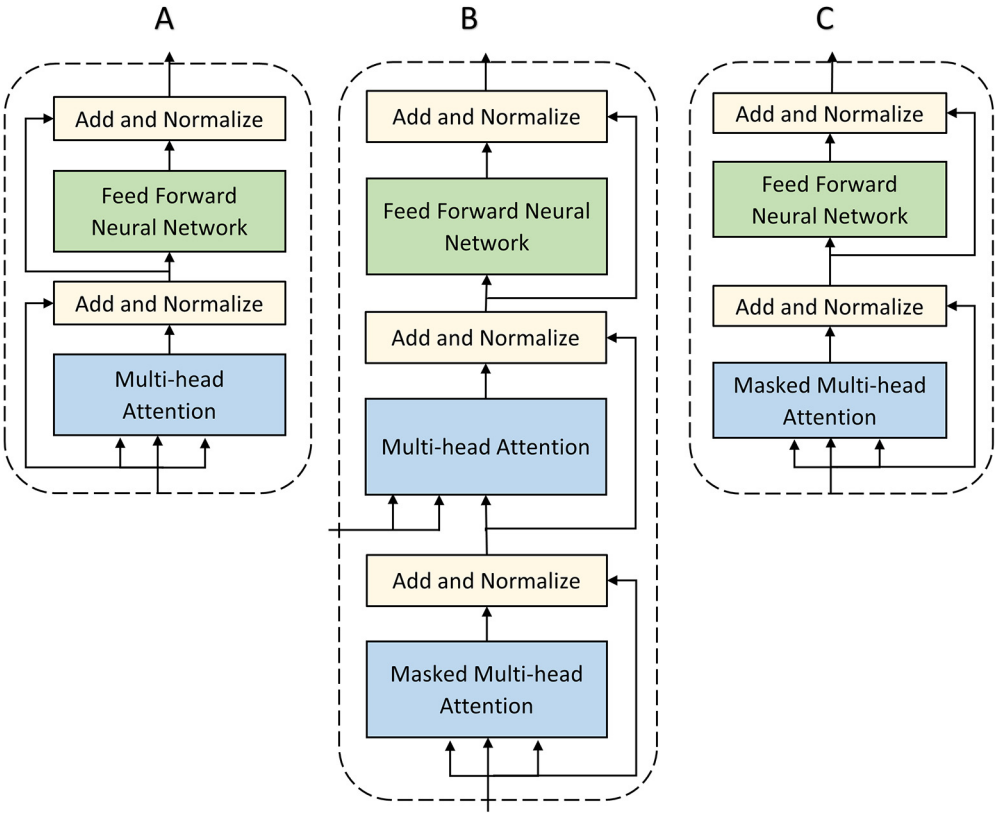

Fig. 7. A) Original Transformer-Encoder Block. B) Original Transformer-Decoder Block. C) Modified Transformer-Decoder Block For Generative Pretraining (GPT). Each block consists of its own feed forward neural network and multi-head attention mechanism. The original transformer by [120] consists of N stacked encoder-decoder blocks (A-B), whereas BERT consists of N stacked transformer-encoders (A) only. The GPT/GPT-2/GPT-3 models use N stacked modified transformer-decoders (C) with only masked-multi head attention as opposed to the original transformer-decoder (B), which use two different atten.tion mechanisms.

self-supervision tasks: 1) By using MLM which corrupts the input text sequences and attempts to predict the original text sequence, 2) By attempting to predict the next sentence for a given sequence of words (sentence). The datasets used for the above pretraining consist of text passages from BooksCorpus with 800 M words [156] and a filtered version of English Wikipedia with 2500 M words. Training the model with the aforementioned data results in a contextualized embedding model that can be later used in target tasks fine-tuned with task-specific supervised data. During fine-tuning, the source model is used with its learned weights, and an additional dense output layer is added to fine-tune the source model for the target task.

BERT was the first transformer-based language model to be introduced by Google, and was available open source triggering a mass interest among research communities due to its versatility to be easily adapted to target tasks. There have been multiple variations of the BERT model published since, including RoBERTa which is a Robust BERT approach introduced by [72]; mBERT that is a multilingual BERT model by Google [98]; and BioBERT which is a model retrained using PubMed and PubMed Central corpora adapted specifically to biomedical domain [61].

GPT, GPT-2 and GPT-3: Generative Pre-training (GPT) based language models are a new generation of language models introduced by OpenAI [104]. In contrast to BERT, which uses transformer-encoders, GPT uses the modified instance of transformer-decoder introduced by [70] which is illustrated in Fig. 7 C. The first GPT model was pre-trained unidirectionally using the BookCorpus dataset [156] containing over 7000 unpublished books, by windowing such that each token's right sided neighbours are masked (auto-regressive modelling). The pre-training objective is to model the distribution $p(x_i|x_{i-w},\ldots,x_{i-2},x_{i-1})$ where $X=\{x_1,x_2,\ldots,x_i\}$ is the complete token sequence, x_i is the i^{th} token, w is the window size and $X_w=\{x_{i-w},\ldots,x_{i-2},x_{i-1},x_i\}$ are tokens within the window w. Once pretrained the GPT language model can be fine-tuned for multiple target tasks such as natural language inference, question answering, semantic analysis or classification by using a labeled target dataset.

The GPT-2 is the next improvement in generative pretraining modelling, using the same decoder architecture as GPT with adjustments to the normalization [105]. However, the difference between GPT and GPT-2 is in the amount of unsupervised data used for the initial pretraining of the model. GPT-2 is trained on WebText dataset uniquely composed for GPT-2, con-

taining over 8 million documents scraped from web pages covering a wide variety of subjects. Interestingly, the authors state that the GPT-2 model requires no fine-tuning to transfer to a target task.

GPT-3 is the latest advancement in generative pretraining based language modelling which has been introduced as a task-agnostic language model, i.e., a language model that can be used for a multiple number of target tasks and natural language understanding tasks without the need for fine-tuning, trained with labelled target data [19]. The GPT-3 language model is pretrained with over 300 billion tokens, using multiple text corpora such as CommonCrawl, Wikipedia, WebText2. Overall, the model performed outstandingly in several natural language understanding tasks.

Both the GPT-2 and GPT-3 language models by OpenAI have shown impressive capability to model natural languages and related tasks such as text summarizing, reading comprehension, translation, question answering, natural language inference, and modelling long-range text dependencies [105,19], paving the way to phenomenal progress in artificially intelligent language inference and generation. GPT-3 has even shown promising performance in AI tasks that involve sentence/paragraph completion, common sense reasoning, news article text generation and several other synthetic and qualitative tasks involving natural language understanding [19]. For this reason, the GPT-3 model has recently attracted much limelight in the transfer learning in natural language processing community.

Transformer-XL and XLNet: Transformer-XL (Extra Large) was introduced by Google as an improvement to ordinary transformers introduced earlier in [120], such that much longer contextual dependencies can be modelled for a machine translation tasks. In order to accomplish this, Transformer-XL uses a recurrence mechanism and a new positional encoding system that allows the model to represent dependencies 80% longer than RNNs and 450% longer than simple transformers [34]. The Transformer-XL processes text in fixed segments, similar to transformers, however, the recurrence mechanism provides hidden information from the two previous segments to the current one, with relational positions.

Based on this new transformer model, a new language model called XLNet is introduced by Google and Carnegie Mellon University [137]. The authors describe their new modelling approach as Permutation Language Modelling (PLM), which combines the advantages of masked language modelling (MLM) and generative pretraining (GPT), but counterbalances their respective drawbacks. Generally, the generative pretraining objective uses auto-regressive modelling which restricts the ability of a LM to learn deep bidirectional context. Whereas masked language modelling (used in BERT), prohibits the LM from learning dependencies with respect to the masked/corrupted tokens, leading to a possible pretraining to fine-tuning discrepancy. To counteract, the PLM approach, XLNet uses all possible permutations of a given token sequence, each time predicting the next masked token. Since the sequence is permuted multiple times, the learning is bidirectional and no tokens are permanently masked reducing discrepancies.

The XLnet model is trained using a large set of sub-word tokens with over 30 billion words extracted from Wikipedia, BooksCorpus, Giga5, ClueWeb, and Common Crawl datasets. The pre-trained XLnet model can be fine-tuned for multiple target tasks with labeled target data and has shown to achieve superior results, e.g., compared to BERT.

ELECTRA: In contrast to the models discussed above, which use generative self-supervised learning for the pre-training, ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately), introduced as a collaboration between Google and Stanford University, uses a discriminative pre-training approach. According to [28], this method requires substantially less data and computational resources than XLNet and RoBERTa to perform in similar capacity as the latter models. Analysis showed that when using only 1/4th of the data ELECTRA performs similar to XLNet and RoBERTa but outperformed both models when using a similar amount of training data. The training approach introduced is called replaced token detection, where the input sentences are corrupted similar to BERT. However, instead of masking tokens, here a token is replaced with a new token generated using a small generator network similar to a masked language model (MLM). The pre-training task focuses on classifying tokens into two classes, i.e., tokens that are replaced and tokens that remain unchanged.

The general ELECTRA model used for the majority of evaluations is said to have been pre-trained with the same BookCorpus and Wikipedia datasets that have been used in BERT pre-training. However, the authors state that for the large ELECTRA model, pre-training was done using the same datasets as XLNet, which contains over 30 billion tokens [137].

5. Discussion

Despite the fact that research on transfer learning started already in the 1970s [18] becoming interesting for the wider machine learning community in the early 1990s [101], deep learning approaches for transfer learning with application on text data is a relatively new field. Among the first publications in this area, is the paper of [142], which used a restricted Boltzmann machine to discover hierarchical features for document classification. Since then a vast number of publications appeared making it difficult for obtaining a comprehensive and systematic overview of the field.

An additional issue results from the fact that so far there is no unique terminology used consistently in the literature. Instead, many different terms and characterization can be found to describe various forms of deep transfer learning models. For traditional approaches, which are not based on deep learning, the situation is similar. We think this is partially related to the interdisciplinary nature of the field where individual areas developed their own terminology independently from other areas. The latter seems also a reflection of a limited communication and exchange of information between the various application domains. However, even more severe is the lack of a unique terminology of general (including deep learning) transfer

learning models. While the review by [92] set the standard for the field, they did not solve this issue. Instead, the paper discussed different categories of 'What to transfer', 'How to transfer' and 'When to transfer' (see Section 2.3.1, 2.3.3 for a discussion) providing very useful specifications of general transfer learning. However, these specifications are not sufficient to give terminological clarity and guidance as can be seen from the confusion in the literature in recent years. Nevertheless, essentially all reviews that appeared after the paper of [92] either copy their specifications or are heavily derivative thereof including the paper of [128,157].

We think the root cause for these issues is the combinatorial nature of transfer learning. That means transfer learning is not a monolithic approach with a limited number of configurations for a limited number of situations but, instead, a diverse family of models. Due to this diversity, transfer learning defies a simple categorization. Specifically, for transfer learning one needs to specify (1) the data (for instance labeled source data and labeled target data), (2) properties of the data (for instance marginal distributions of the source and target) and (3) a model approach (for instance parameter transfer) for learning a task. Importantly, these three components are not alternatives but each one needs to be specified for a given transfer learning problem. Hence, defining transfer learning means to specify all combinations thereof. By making simplified assumptions about the different possible components (e.g., 9 different combinations of data (see Fig. 2), 4 different distributional properties of the data (see Table 1 different solution-based approaches (see Section 2.3.3) one obtains 144 different transfer learning categories.

Unfortunately, there is an additional layer of complexity that needs to be considered. While classic machine learning approaches are typically one-step methods, deep transfer learning models are often multi-step procedures using a mixture or a selection of source and target data for specific steps (see Section 4.1.1 and [45]). Hence, such multi-step procedures lead to additional combinations that need to be considered because the data are not used in one specific way but source and target can be combined or selectively combined in various different ways in each learning step. Overall, this results in many more categories of transfer learning than 144, which makes it clear why so far no comprehensive terminology exists that would assign simple names to individual configurations. Also, such a number of categories makes it impractical to visualize a comprehensive taxonomy, e.g., via a hierarchical tree, because hundreds of nested branches do not provide a simple overview.

In order to overcome those problems, we suggest a nomenclature (see Section 2.4) that on the one hand maintains the combinatorial structure of transfer learning and accommodates on the other hand the same time multi-step procedures. In order to demonstrate the utility of our nomenclature, we show in Fig. 8 two examples. The left figure shows a Stacked Denoising Autoencoder (see Section 4.1.1 and the paper of [45]). A SDA consists of two-learning steps. For step 1 unlabeled source and unlabeled target data are used for learning a new feature representation, whereas for step 2 the labeled target data are used for training the classifier. The figure on the right-hand-side shows BERT [37]. Also, BERT consists of two-learning steps. In step 1 only the unlabeled source data are used wheres in step 2 only the labeled target data are used for training the classifier. These two examples demonstrate also that transfer learning does not require to use the source and target data sequentially, although there are sequential transfer learning methods as discussed in Section 4.3.1 doing exactly this, but there are also other models that utilize a mixture of the source and target data over multiple steps. Overall, this leads to a combinatorial plurality making the categorization of transfer learning models a complex task by itself.

For reasons of clarity, we would like to emphasize that the visual taxonomy shown in Fig. 3 is a simplification of our nomenclature. The reason for this is twofold. First, as discussed above, the total number of possible combinations that define transfer learning models is very large. Second, not every combination that is theoretically feasible is equally frequent in the literature. For instance, we are not aware of any transfer learning model that is based on partially labeled source and partially labeled target data. This implies that many of the possible combinations are either sparsely populated by published articles or even empty. Hence, even if it would be feasible to visualize all combinations resulting from our nomenclature this would not be informative with respect to the current literature.

In order to improve the communication among the communities studying transfer learning we suggest that future publications adopt our nomenclature. This will remove ambiguity and enhance the exchange of crucial information for training transfer learning models.

Furthermore, in order to enhance future studies about deep transfer learning we provided in this paper also an overview of text data resources frequently used as benchmark data. From surveying the literature, we observed that most of the review dataset are publicly available including: Amazon product reviews, Multi-Domain Sentiment Dataset (https://www.cs.jhu.edu/ mdredze/ datasets/sentiment/index2.html), or crosss language Amazon products (https://webis.de/data/webis-cls-10.html). Furthermore, data about spam emails (http://www.ecmlpkdd2006. org/challenge.html#download) and News groups (http://qwone.com/ jason/20Newsgroups/) are available. From Table 2, one can see that the majority of the benchmark data used in the literature for studying deep transfer learning models are for product reviews, e.g., about books, cameras and laptops, or restaurants and hotels. It is interesting to note that there are only three datasets related to biomedical texts (e.g. MIMIC-III).

Regarding the evaluation of the studied models, all have been assessed for a supervised learning task performing a classification. In case of unlabeled target data, labeled test data have been used sampled from the target domain. Various types of error measures for classification have been used, including accuracy, F1-score, precision and recall [39], whereas the accuracy is the by far most frequently used measure. It is interesting to note that the standard error, estimating the variability of the mean of an error measure, is rarely reported despite that fact that it provides important information about the effect of changing training and test data. For this reason, it is not always clear if the reported error measures provide sensible esti-

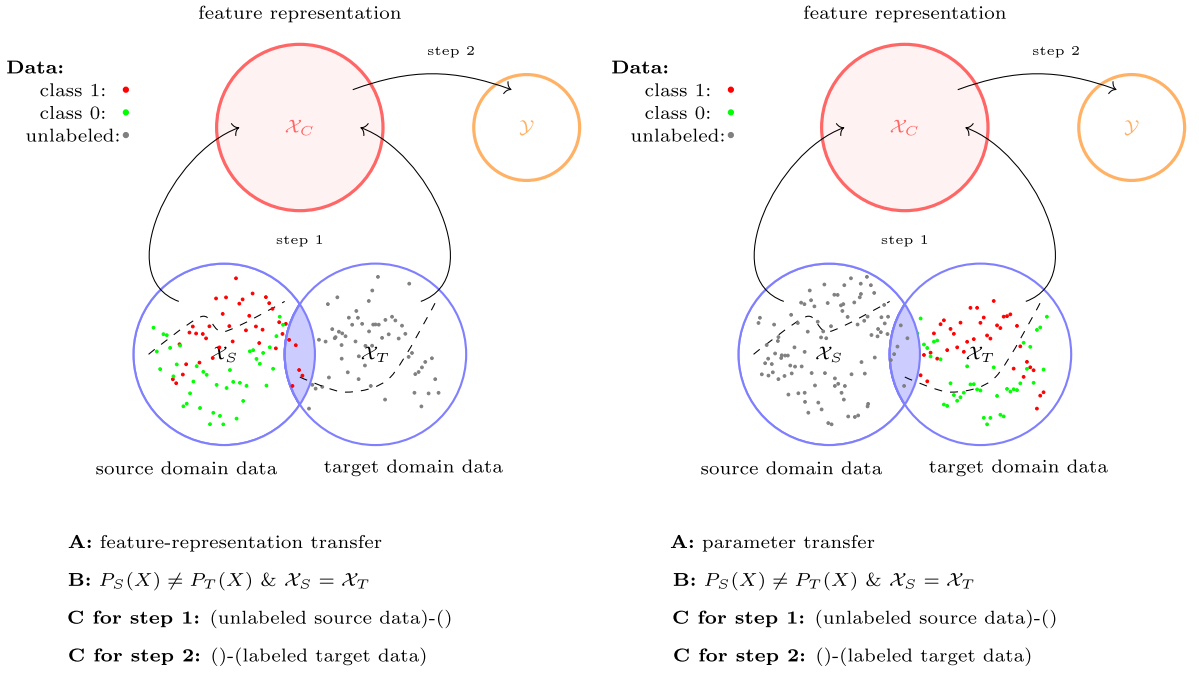

Fig. 8. Visualization of the combinatoral plurality offered by transfer learning with respect to the utilization of source and target data over multiple learning steps of a model. Left: Stacked Denoising Autoencoder (SDA). Right: Bidirectional Encoder Representations from Transformers (BERT). The components of the nomenclature of each model, i.e., A, B and C, are discussed in Section 2.4.

mates of the generalization error of the model [38,90]. Similarly, sample size considerations are rarely addressed and learning curves are largely neglected [6]. Overall, there is a need for more statistical considerations to demonstrate that the obtained results are statistically robust.

5.1. Future directions

Based on our discussion above, we identified a number of extensions that could be studied to further enhance our understanding of deep transfer learning. First, transfer learning could be studied in a broader range of text applications. For example, spam filtering, named entity recognition and part-of-speech (POS) are areas that remain so far understudied. Also more complex situations could be studied to demonstrate the utility of transfer learning, e.g., fraud detection and social media applications [41]. A related issue is that existing methods need to be evaluated not only for product reviews but also for related fields. This could include biological and medical reviews about medications or treatments. For instance, from our discussion above in Section 4.1 one can see that Autoencoder and Adversarial learning have been always evaluated on the Amazon review datasets. Hence, at this point it is unclear if those methods generalize well to other review topics.

Second, surveying the literature revealed that there is a variety of different ways to learn from source and target data to establish a deep transfer learning method. Specifically, SDA [45] and BERT [37] are two-step procedures. However, while BERT uses in the first step only the source data and in the second step only the target data, SDA uses in the first step a mixture of source and target data and in the second step only the labeled source data. In contrast, the adversarial learning model DANN [44] is a one-step learning method. It is interesting to note that so far neither the number of steps of a learning model nor the mixing of source and target data for the corresponding steps has been systematically investigated. This is important because the characteristics of a given data set entails a learning paradigm. Since transfer learning has the freedom to select, or mix, source and target data over multiple steps the learning paradigm of the corresponding steps can change. Overall, this makes transfer learning very flexible and it would be interesting to study what combination leads to optimal results.

Third, instead of learning from just one source domain one could extend transfer learning to include several source domains. Specifically, [55] argued that this should enhance the performance of models because information can be accumulated over a number of different source domains. While there are already a few approaches that accommodate more than one source domain, e.g., [149,150], such studies correspond to the minority.

Forth, for testing deep transfer learning methods it would be useful if a benchmark database or data repository would exist providing big text data for a variety of different application domains. This would allow to avoid the time-consuming and potentially costly process of data collection and data curation. This would also enable the standardized comparison among different models. Our section about frequently used text data (see Section 3) could provide a starting point for such an initiative.

Fifth, further theoretical studies would be helpful to provide a theoretical underpinning for general transfer learning and for transfer learning in specific situations. This could not only lead to a better foundation but might also inspire the development of new models with better practical performance.

6. Conclusion

In this article, we provided a comprehensive review of deep transfer learning models for analyzing text data. Our paper has three main objectives. First, we review an existing terminology and categorizations of transfer learning. Based on this, we introduced a new nomenclature allowing the unequivocal description of transfer learning models. Importantly, the nomenclature reflects the combinatorial plurality of transfer learning in an implicit form leading to a compact formulation. Second, we introduced a taxonomy of deep transfer learning models for applications to text data amenable for a visualization. This taxonomy combines key information expressed by our nomenclature but simplifies the combinatorial plurality offered by transfer learning in order to enable its practical utility. As a result, the taxonomy provides a comprehensive overview of the currently studied deep transfer learning models emphasizing architectural principles. Third, we provided an overview of useful resources of text data that have been used as benchmark data for studying text applications. Finally, we discussed a number of extensions that could be studied to further enhance our understanding of deep transfer learning.

Author Contributions

FES conceived the study. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

MD thanks the Austrian Science Funds for supporting this work (project P30031).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Abdul-Mageed, L. Emonet Ungar, Fine-grained emotion detection with gated recurrent neural networks, in: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), 2017, pp. 718–728.
[2] Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., and Marchand, M. Domain-Adversarial Neural Networks..
[3] T. Al-Moslmi, N. Omar, S. Abdullah, M. Albared, Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review, IEEE Access 5 (2017) 16173–16192.
[4] Alam, F., Joty, S., and Imran, M. Domain adaptation with adversarial training and graph embeddings. ACL 2018–56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) 1 (2018), 1077–1087..
[5] M.Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M.S. Nasrin, M. Hasan, B.C. Van Essen, A.A. Awwal, V.K. Asari, A state-of-the-art survey on deep learning theory and architectures, Electronics 8 (3) (2019) 292.
[6] S.-I. Amari, A universal theorem on learning curves, Neural networks 6 (2) (1993) 161–166.
[7] M.R. Amini, N. Usunier, C. Goutte, Learning from multiple partially observed views-an application to multilingual text categorization, Advances in Neural Information Processing Systems (2009) 28–36.
[8] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: International conference on machine learning (2013), PMLR, 2013, pp. 1247–1255.
[9] M.T. Bahadori, Y. Liu, D. Zhang, A general framework for scalable transductive transfer learning, Knowledge and Information Systems 38 (1) (2014) 61–83.
[10] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J.W. Vaughan, A theory of learning from different domains, Machine learning 79 (1) (2010) 151–175.
[11] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, Analysis of representations for domain adaptation, Advances in Neural Information Processing Systems 19 (2006) 137–144.
[12] Y. Bengio, Deep Learning of Representations for Unsupervised and Transfer Learning, JMLR: Workshop and Conference Proceedings 7 (2012) 1–20.
[13] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1798–1828.
[14] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, et al, Greedy layer-wise training of deep networks, Advances in Neural Information Processing Systems 19 (2007) 153.
[15] Bickel, S. Ecml-pkdd discovery challenge 2006 overview. In ECML-PKDD Discovery Challenge Workshop (2006), pp. 1–9..
[16] Blitzer, J., Dredze, M., and Pereira, F. ACL07 Biographies, Bollywood, Boom-boxes and Blenders. Association for Computational Linguistics - ACL 2007, June (2007), 440–447..
[17] J. Blitzer, R. McDonald, F. Pereira, Domain adaptation with structural correspondence learning, in: Proceedings of the 2006 conference on empirical methods in natural language processing, 2006, pp. 120–128.
[18] S. Bozinovski, Reminder of the first paper on transfer learning in neural networks, 1976, Informatica 44 (2020) 3.
[19] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)..
[20] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language Resources and Evaluation 42 (4) (2008) 335–359.

[21] R. Chattopadhyay, Q. Sun, W. Fan, I. Davidson, S. Panchanathan, J. Ye, Multisource domain adaptation and its application to early detection of fatigue, ACM Transactions on Knowledge Discovery from Data (TKDD) 6 (4) (2012) 1–26.
[22] Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. One billion word benchmark for measuring progress in statistical language modeling. Tech. rep., Google, 2013..
[23] Chen, M., Xu, Z., Weinberger, K.Q., and Sha, F. Marginalized denoising autoencoders for domain adaptation. Proceedings of the 29th International Conference on Machine Learning, ICML 2012 1 (2012), 767–774..
[24] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, K. Weinberger, Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification, Transactions of the Association for Computational Linguistics 6 (2018) 557–570.
[25] Y. Chen, S. Song, S. Li, L. Yang, C. Wu, Domain space transfer extreme learning machine for domain adaptation, IEEE Transactions on Cybernetics 49 (5) (2019) 1909–1922.
[26] Z. Chen, T. Qian, Transfer capsule network for aspect level sentiment classification, in: ACL 2019–57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2020, pp. 547–556.
[27] S.D.L.I.D. Chopra, Deep Learning for Domain Adaptation by Interpolating between Domains, in: International Conference on Machine Learning (ICML), Workshop on Representation Learning, 2013.
[28] Clark, K., Luong, M.-T., Le, Q.V., and Manning, C.D. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)..
[29] Clinchant, S., Csurka, G., and Chidlovskii, B. A domain adaptation regularization for denoising autoencoders. 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers (2016), 26–31..
[30] D. Cook, K.D. Feuz, N.C. Krishnan, Transfer learning for activity recognition: A survey, Knowledge and Information Systems 36 (3) (2013) 537–556.
[31] G. Csurka, A comprehensive survey on domain adaptation for visual applications, Advances in Computer Vision and Pattern Recognition (2017) 1–35, 9783319583464.
[32] W. Dai, O. Jin, G.-R. Xue, Q. Yang, Y. Yu, Eigentransfer: a unified framework for transfer learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 193–200.
[33] Dai, W., Xue, G.-R., Yang, Q., and Yu, Y. Transferring naive bayes classifiers for text classification. In AAAI (2007), vol. 7, pp. 540–545..
[34] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019)..
[35] Danescu-Niculescu-Mizil, C., and Lee, L. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. arXiv preprint arXiv:1106.3077 (2011)..
[36] O. Day, T.M. Khoshgoftaar, A survey on heterogeneous transfer learning, Journal of Big Data 4 (2017) 1.
[37] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)..
[38] F. Emmert-Streib, M. Dehmer, Evaluation of regression models: Model assessment, model selection and generalization error, Machine Learning and Knowledge Extraction 1 (1) (2019) 521–551.
[39] F. Emmert-Streib, S. Moutari, M. Dehmer, A comprehensive survey of error measures for evaluating binary decision making in data science, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2019) e1303.
[40] Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., and Dehmer, M. An Introductory Review of Deep Learning for Prediction Models With Big Data. Frontiers in Artificial Intelligence 3, February (2020), 1–23..
[41] F. Emmert-Streib, O. Yli-Harja, M. Dehmer, Utilizing social media data for psychoanalysis to study human personality, Frontiers in Psychology 10 (2019) 2596.
[42] K. Feng, T. Chaspari, A review of generalizable transfer learning in automatic emotion recognition, Frontiers in Computer Science 2 (2020) 9.
[43] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: International conference on machine learning, PMLR, 2015, pp. 1180–1189.
[44] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. The Journal of Machine Learning Research 17, 1 (2016), 2096–2030..
[45] Glorot, X., Bordes, A., and Bengio, Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. Proceedings of the 28th International Conference on Machine Learning, ICML 2011, 1 (2011), 513–520..
[46] J.M. Gomez-Perez, R. Denaux, A. Garcia-Silva, Understanding word embeddings and language models, in: A Practical Guide to Hybrid Natural Language Processing, Springer, 2020, pp. 17–31.
[47] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, A. Smola, A kernel method for the two-sample-problem, Advances in Neural Information Processing Systems 19 (2006) 513–520.
[48] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew, Deep learning for visual understanding: A review, Neurocomputing 187 (2016) 27–48.
[49] I. Guyon, G. Dror, V. Lemaire, G. Taylor, D.W. Aha, Unsupervised and transfer learning challenge, in: Proceedings of the International Joint Conference on Neural Networks, 2012, pp. 793–800.
[50] Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.-P., and Zimmermann, R. Conversational memory network for emotion recognition in dyadic dialogue videos. In Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting (2018), vol. 2018, NIH Public Access, p. 2122..
[51] D. Hazarika, S. Poria, R. Zimmermann, R. Mihalcea, Conversational transfer learning for emotion recognition, Information Fusion 65 (2021) 1–12.
[52] G.E. Hinton, A. Krizhevsky, S.D. Wang, Transforming auto-encoders, in: International conference on artificial neural networks, Springer, 2011, pp. 44–51.
[53] Howard, J., and Ruder, S. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018)..
[54] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177.
[55] Huang, X., Rao, Y., Xie, H., Wong, T.-L., and Wang, F.L. Cross-domain sentiment classification via topic-related tradaboost. In Proceedings of the AAAI Conference on Artificial Intelligence (2017), vol. 31..
[56] M.S. Iqbal, B. Luo, T. Khan, R. Mehmood, M. Sadiq, Heterogeneous transfer learning techniques for machine learning, Iran Journal of Computer Science 1 (1) (2018) 31–46.
[57] Kouw, W.M., and Loog, M. An introduction to domain adaptation and transfer learning. arXiv preprint arXiv:1812.11806 (2018)..
[58] W.M. Kouw, M. Loog, A review of domain adaptation without target labels, IEEE transactions on pattern analysis and machine intelligence (2019).
[59] Lafferty, J., McCallum, A., and Pereira, F.C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data..
[60] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[61] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.
[62] Lee, J.Y., Dernoncourt, F., and Szolovits, P. Transfer learning for named-entity recognition with neural networks. arXiv preprint arXiv:1705.06273 (2017)..
[63] Li, H., Parikh, N.A., and He, L. A novel transfer learning approach to enhance deep neural network classification of brain functional connectomes. Frontiers in Neuroscience 12, JUL (2018), 1–12..
[64] Li, S., and Zong, C. Multi-domain adaptation for sentiment classification: Using multiple classifier combining methods. In 2008 International Conference on Natural Language Processing and Knowledge Engineering (2008), IEEE, pp. 1–8..
[65] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. Dailydialog: A manually labelled multi-turn dialogue dataset. arXiv preprint arXiv:1710.03957 (2017)..

[66] Li, Z., Wei, Y., Zhang, Y., and Yang, Q. Hierarchical attention transfer network for cross-domain sentiment classification. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, January (2018), 5852–5859..
[67] Z. Li, Y. Zhang, Y. Wei, Y. Wu, Q. Yang, End-to-end adversarial memory network for cross-domain sentiment classification, IJCAI International Joint Conference on Artificial Intelligence (2017) 2237–2243.
[68] H. Liang, W. Fu, F. Yi, A survey of recent advances in transfer learning, in: 2019 IEEE 19th International Conference on Communication Technology (ICCT), IEEE, 2019, pp. 1516–1523.
[69] Lin, Y., Lei, H., Wu, J., and Li, X. An empirical study on sentiment classification of chinese review using word embedding. arXiv preprint arXiv:1511.01665 (2015)..
[70] Liu, P.J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., and Shazeer, N. Generating wikipedia by summarizing long sequences. arXiv preprint arXiv:1801.10198 (2018)..
[71] R. Liu, Y. Shi, C. Ji, M. Jia, A survey of sentiment analysis based on transfer learning, IEEE Access 7 (2019) 85401–85412.
[72] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)..
[73] Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In International conference on machine learning (2015), PMLR, pp. 97–105..
[74] M. Long, J. Wang, Y. Cao, J. Sun, P.S. Yu, Deep learning of transferable representation for scalable domain adaptation, IEEE Transactions on Knowledge and Data Engineering 28 (8) (2016) 2027–2040.
[75] Lowe, R., Pow, N., Serban, I., and Pineau, J. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. arXiv preprint arXiv:1506.08909 (2015)..
[76] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, G. Zhang, Transfer learning using computational intelligence: A survey, Knowledge-Based Systems 80 (2015) 14–23.
[77] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 6818–6825.
[78] Mao, H.H. A survey on self-supervised pre-training for sequential transfer learning in neural networks. arXiv preprint arXiv:2007.00800 (2020)..
[79] G. McKeown, M. Valstar, R. Cowie, M. Pantic, M. Schroder, The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent, IEEE transactions on affective computing 3 (1) (2011) 5–17.
[80] J. Meng, Y. Long, Y. Yu, D. Zhao, S. Liu, Cross-domain text sentiment analysis based on CNN_FT method, Information (Switzerland) 10 (2019) 5.
[81] Merity, S., Keskar, N.S., and Socher, R. Regularizing and optimizing lstm language models. arXiv preprint arXiv:1708.02182 (2017)..
[82] Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843 (2016)..
[83] Mesnil, G., Mikolov, T., Ranzato, M., and Bengio, Y. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. arXiv preprint arXiv:1412.5335 (2014)..
[84] S. Mohammad, M. Salameh, S. Kiritchenko, Sentiment lexicons for arabic social media, in: Proceedings of the tenth international conference on language resources and evaluation (LREC'16), 2016, pp. 33–37.
[85] A. Moreo, A. Esuli, F. Sebastiani, Lost in transduction: Transductive transfer learning in text classification, ACM Transactions on Knowledge Discovery from Data (TKDD) 16 (1) (2021) 1–21.
[86] S. Moriya, C. Shibata, PTransfer Learning Method for Very Deep CNN for Text Classification and Methods for its Evaluation, Proceedings - International Computer Software and Applications Conference 2 (2018) 153–158.
[87] Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L., and Jin, Z. How transferable are neural networks in NLP applications? EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings (2016), 479–489..
[88] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., and Stoyanov, V. Semeval-2016 task 4: Sentiment analysis in twitter. arXiv preprint arXiv:1912.01973 (2019)..
[89] C. Neudecker, An open corpus for named entity recognition in historic newspapers, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 4348–4352.
[90] Nicholson, A.M. Generalization error estimates and training data valuation. PhD thesis, California Institute of Technology, 2002..
[91] S.J. Pan, X. Ni, J.-T. Sun, Q. Yang, Z. Chen, Cross-domain sentiment classification via spectral feature alignment, in: Proceedings of the 19th international conference on World wide web, 2010, pp. 751–760.
[92] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22 (10) (2009) 1345–1359.
[93] Pang, B., and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. arXiv preprint cs/0506075 (2005)..
[94] V.M. Patel, R. Gopalan, R. Li, R. Chellappa, Visual domain adaptation: A survey of recent advances, IEEE signal processing magazine 32 (3) (2015) 53–69.
[95] Z. Pei, Z. Cao, M. Long, J. Wang, Multi-adversarial domain adaptation, in: 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 2018, pp. 3934–3941.
[96] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)..
[97] Peters, M.E., Ruder, S., and Smith, N.A. To tune or not to tune? adapting pretrained representations to diverse tasks. arXiv preprint arXiv:1903.05987 (2019)..
[98] Pires, T., Schlinger, E., and Garrette, D. How multilingual is multilingual bert? arXiv preprint arXiv:1906.01502 (2019)..
[99] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos, Semeval-2015 task 12: Aspect based sentiment analysis, in: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), 2015, pp. 486–495.
[100] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, L.-P. Morency, Multi-level multiple attentions for contextual multimodal sentiment analysis, in: 2017 IEEE International Conference on Data Mining (ICDM), IEEE, 2017, pp. 1033–1038.
[101] L.Y. Pratt, J. Mostow, C.A. Kamm, A.A. Kamm, Direct transfer of learned information among neural networks, Aaai 91 (1991) 584–589.
[102] P. Prettenhofer, B. Stein, Cross-lingual adaptation using structural correspondence learning, ACM Transactions on Intelligent Systems and Technology (TIST) 3 (1) (2011) 1–22.
[103] C. Quan, F. Ren, Sentence emotion analysis and recognition based on emotion words using ren-cecps, International Journal of Advanced Intelligence 2 (1) (2010) 105–117.
[104] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training..
[105] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI blog 1 (8) (2019) 9.
[106] R. Raina, A. Battle, H. Lee, B. Packer, A.Y. Ng, Self-taught learning: transfer learning from unlabeled data, in: Proceedings of the 24th international conference on Machine learning, 2007, pp. 759–766.
[107] Riedl, M., and Padó, S. A named entity recognition shootout for German. ACL 2018–56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) 2 (2018), 120–125..
[108] Rosenstein, M.T., Marx, Z., Kaelbling, L.P., and Dietterich, T.G. To transfer or not to transfer. In NIPS 2005 workshop on transfer learning (2005), vol. 898, pp. 1–4..
[109] Ruder, S. An Overview of Multi-Task Learning in Deep Neural Networks..
[110] Ruder, S. Neural Transfer Learning for Natural Language Processing. PhD thesis, NATIONAL UNIVERSITY OF IRELAND, GALWAY, 2019..
[111] Sabour, S., Frosst, N., and Hinton, G.E. Dynamic routing between capsules. arXiv preprint arXiv:1710.09829 (2017)..

[112] Sang, E.F., and De Meulder, F. Introduction to the conll-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050 (2003)..
[113] T. Semwal, G. Mathur, P. Yenigalla, S.B. Nair, A practitioners' guide to transfer learning for text classification using convolutional neural networks, SIAM International Conference on Data Mining, SDM 2018 (2018) 513–521.
[114] Serban, I., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the AAAI Conference on Artificial Intelligence (2016), vol. 30..
[115] L. Shao, F. Zhu, X. Li, Transfer learning for visual categorization: A survey, IEEE transactions on neural networks and learning systems 26 (5) (2014) 1019–1034.
[116] R. Socher, A. Perelygin, J. Wu, J. Chuang, C.D. Manning, A.Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1631–1642.
[117] Sukhbaatar, S., Szlam, A., Weston, J., and Fergus, R. End-to-end memory networks. arXiv preprint arXiv:1503.08895 (2015)..
[118] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. A survey on deep transfer learning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11141 LNCS (2018), 270–279..
[119] M.E. Taylor, P. Stone, Transfer learning for reinforcement learning domains: A survey, Journal of Machine Learning Research 10 (2009) 7.
[120] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)..
[121] Vincent, P., and Larochelle, H. Extracting and Composing Robust Features with Denoising.pdf. 1096–1103..
[122] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.A. Manzagol, Stacked denoising autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion, Journal of Machine Learning Research 11 (2010) 3371–3408.
[123] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep Learning for Computer Vision: A Brief Review, Computational Intelligence and Neuroscience 2018 (2018).
[124] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, R. Fergus, Regularization of neural networks using dropconnect, in: International conference on machine learning, PMLR, 2013, pp. 1058–1066.
[125] H. Wang, Y. Lu, C. Zhai, Latent aspect rating analysis without aspect keyword supervision, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 618–626.
[126] Y. Wang, M. Huang, X. Zhu, L. Zhao, Attention-based lstm for aspect-level sentiment classification, in: Proceedings of the 2016 conference on empirical methods in natural language processing, 2016, pp. 606–615.
[127] Z. Wang, Y. Qu, L. Chen, J. Shen, W. Zhang, S. Zhang, Y. Gao, G. Gu, K. Chen, Y. Yu, Label-Aware double transfer learning for cross-specialty medical named entity recognition, in: NAACL HLT 2018–2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1 i, 2018, pp. 1–15.
[128] Wei, P., Ke, Y., and Goh, C.K. Deep nonlinear feature coding for unsupervised domain adaptation. IJCAI International Joint Conference on Artificial Intelligence 2016-Janua (2016), 2189–2195..
[129] P. Wei, Y. Ke, C.K. Goh, Feature analysis of marginalized stacked denoising autoenconder for unsupervised domain adaptation, IEEE transactions on neural networks and learning systems 30 (5) (2018) 1321–1334.
[130] K. Weiss, T.M. Khoshgoftaar, D.D. Wang, A survey of transfer learning, vol. 3, Springer International Publishing, 2016.
[131] G. Wilson, D.J. Cook, A survey of unsupervised deep domain adaptation, ACM Transactions on Intelligent Systems and Technology (TIST) 11 (5) (2020) 1–46.
[132] Q. Wu, S. Tan, A two-stage framework for cross-domain sentiment classification, Expert Systems with Applications 38 (11) (2011) 14269–14275.
[133] Xu, R., and Yang, Y. Cross-lingual distillation for text classification. ACL 2017–55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) 1 (2017), 1415–1425..
[134] A. Yala, R. Barzilay, L. Salama, M. Griffin, G. Sollender, A. Bardia, C. Lehman, J.M. Buckley, S.B. Coopey, F. Polubriaginof, et al, Using machine learning to parse breast pathology reports, Breast cancer research and treatment 161 (2) (2017) 203–211.
[135] M. Yang, W. Zhao, L. Chen, Q. Qu, Z. Zhao, Y. Shen, Investigating the transferring capability of capsule networks for text classification, Neural Networks 118 (2019) 247–261.
[136] Q. Yang, Y. Zhang, W. Dai, S.J. Pan, Transfer learning, Cambridge University Press, 2020.
[137] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237 (2019)..
[138] J. Yu, J. Jiang, Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification, in: EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings, 2016, pp. 236–246.
[139] Yu, J., Marujo, L., Jiang, J., Karuturi, P., and Brendel, W. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. ACL..
[140] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, J. Wang, Automatic icd-9 coding via deep transfer learning, Neurocomputing 324 (2018) 43–50.
[141] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, J. Wang, Automatic ICD-9 coding via deep transfer learning, Neurocomputing 324 (2019) 43–50.
[142] Zhang, J. Deep transfer learning via restricted boltzmann machine for document classification. In 2011 10th International Conference on Machine Learning and Applications and Workshops (2011), vol. 1, IEEE, pp. 323–326..
[143] K. Zhang, H. Zhang, Q. Liu, H. Zhao, H. Zhu, E. Chen, Interactive Attention Transfer Network for Cross-Domain Sentiment Classification, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 5773–5780.
[144] K. Zhang, H. Zhang, Q. Liu, H. Zhao, H. Zhu, E. Chen, Interactive attention transfer network for cross-domain sentiment classification, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 5773–5780.
[145] Zhang, L. Transfer Adaptation Learning: A Decade Survey. 1–21..
[146] Zhang, L., and Gao, X. Transfer adaptation learning: A decade survey. arXiv preprint arXiv:1903.04687 (2019)..
[147] Y. Zhang, R. Barzilay, T. Jaakkola, Aspect-augmented Adversarial Networks for Domain Adaptation, Transactions of the Association for Computational Linguistics 5 (2017) 515–528.
[148] Zhang, Y., and Yang, Q. A Survey on Multi-Task Learning. 1–20..
[149] Zhao, C., Wang, S., and Li, D. Multi-source domain adaptation with joint learning for cross-domain sentiment classification. Knowledge-Based Systems 191, xxxx (2020), 105254..
[150] H. Zhao, S. Zhang, G. Wu, J.P. Costeira, J.M. Moura, G.J. Gordon, Adversarial multiple source domain adaptation, in: Advances in Neural Information Processing Systems 2018-Decem NeurIPS, 2018, pp. 8559–8570.
[151] L. Zhao, Z. Chen, L.T. Yang, M.J. Deen, Z.J. Wang, Deep semantic mapping for heterogeneous multimedia transfer learning using co-occurrence data, ACM Transactions on Multimedia Computing, Communications and Applications 15 (1s) (2019) 1–21.
[152] Zhao, S., Li, B., Yue, X., Gu, Y., Xu, P., Hu, R., Chai, H., and Keutzer, K. Multi-source Domain Adaptation for Semantic Segmentation. 1–14..
[153] J.T. Zhou, S.J. Pan, I.W. Tsang, Y. Yan, Hybrid heterogeneous transfer learning through deep learning, Proceedings of the National Conference on Artificial Intelligence 3 (2014) 2213–2219.
[154] Zhou, J.T., Xu, X., Pan, S.J., Tsang, I.W., Qin, Z., and Goh, R.S.M. Transfer hashing with privileged information. IJCAI International Joint Conference on Artificial Intelligence 2016-Janua (2016), 2414–2420..
[155] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 207–212.
[156] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 19–27.

[157] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A Comprehensive Survey on Transfer Learning, Proceedings of the IEEE (2020) 1–31.
[158] Y. Ziser, R. Reichart, Neural structural correspondence learning for domain adaptation, CoNLL 2017–21st Conference on Computational Natural Language Learning, Proceedings, CoNLL (2017) 400–410.