Research on Task Discovery for Transfer Learning in Deep Neural Networks

Arda Akdemir
University of Tokyo, Japan
aakdemir@hgc.jp

Abstract

Deep neural network based machine learning models are shown to perform poorly on unseen or out-of-domain examples by numerous recent studies. Transfer learning aims to avoid overfitting and to improve generalizability by leveraging the information obtained from multiple tasks. Yet, the benefits of transfer learning depend largely on task selection and finding the right method of sharing. In this thesis, we hypothesize that current deep neural network based transfer learning models do not achieve their fullest potential for various tasks and there are still many task combinations that will benefit from transfer learning that are not considered by the current models. To this end, we started our research by implementing a novel multi-task learner with relaxed annotated data requirements and obtained a performance improvement on two NLP tasks. We will further devise models to tackle tasks from multiple areas of machine learning, such as Bioinformatics and Computer Vision, in addition to NLP.

1 Introduction

Deep neural network based machine learning models have shown remarkable progress in the last decades across a wide range of tasks. The typical training regime uses a large amount of labeled data to get a general mapping of the elements in the input space to the label space, which is known as supervised learning. Yet, it is shown by numerous studies that these models suffer from overfitting and are sensitive to noise and examples that are not available in the training data (Jia and Liang, 2017; Belinkov et al., 2017). In addition, these models are usually trained from scratch for each new task where the weights of the models are initialized randomly. This approach does not follow the way humans learn new tasks, i.e. leveraging external world knowledge and information obtained from related tasks when learning a new task (Bruner, 1985; Hayes et al., 2002).

Transfer learning (TL) is a biologically motivated training paradigm that aims to mitigate the above mentioned real-world challenges of conventional supervised learning (Ruder, 2019). Signals in the training set of a source task are used as additional information for a given target task to enable better generalization. It is especially useful when the labeled data is limited for the target task and when the tasks are relatively similar (Collobert and Weston, 2008; Hashimoto et al., 2017; Ruder, 2019). Learning the structure among tasks is an essential first step to benefit most from transfer learning, and to this end Zamir et al. (2018) proposed a fully-computational framework to learn this structure in the Computer Vision domain. Straightforward application of transfer learning algorithms may lead to *catastrophic forgetting* where models forget the source task after being exposed to the target task. In addition, there is a lack of theoretical understanding of the task relationships, and as a result, tasks for transfer learning are usually determined with hindsight.

Multi-task learning (MTL) is a special case of transfer learning where multiple tasks are learned simultaneously. Caruana (1997) summarizes multi-task learning as leveraging information obtained from the training data of different tasks to improve generalization. It enables better generalization and lowers the annotated data requirements (Caruana, 1997; Maurer et al., 2016). Current multi-task learning systems typically use *hard-sharing*, where a low layer hidden representation is shared among all tasks to have an inductive bias (Collobert et al., 2011; Chu et al., 2015). It is recently shown that for dissimilar tasks *hard-sharing* may degrade the performance, which is also called *negative-transfer* (Yosinski et al., 2014). More sophisticated information sharing methodologies must be consid-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 33–41

ered in addition to finding useful task combinations, to make the most out of multi-task learning and to avoid *negative-transfer*.

The above findings and challenges motivate our research on transfer learning in deep neural networks. Specifically, we focus our research on investigating the task relations on the currently proposed models and on proposing new task combinations. Through our research, we plan to find answers to 'where to transfer from' (task selection), 'what to transfer' (datasets and data selection) and 'how to transfer' (pretraining and model architecture). Our main hypothesis is that, 1) neural network based transfer learning models improve over their single-task counterparts both in terms of generalizability and overall performance, 2) currently proposed transfer learning models do not achieve their fullest potential, and 3) there are many task combinations that will benefit from transfer learning. We will focus on the following research questions about transfer learning models throughout this thesis:

RQ1. How to optimize the model architecture and sharing methodology for a given task combination?

RQ2. What are some good auxiliary tasks to improve the perfomance of a target task?

RQ3. How to find useful pretraining schemes?

The first question aims to find the most useful architecture and the sharing methodology when the task combination is known/determined. Second is a higher-level research question to find useful task combinations and can be considered as the preliminary step for the first one. Finally, question three aims to find the right pretraining scheme to make the most ouf of transfer learning for a given set of target tasks. By combining these research questions, we aim to find the most useful multi-task learning setting for a given domain.

We started our research by analyzing the limitations of current supervised learning systems and showed the sensitivity of neural network based models to the changes in the domain (Akdemir et al., 2018). Next, we proposed a novel joint learning model that relaxes labeled data requirements for the Named Entity Recognition and Dependency Parsing tasks and showed improvements over the conventional methods. The results for the model are given in more detail in Section 4. We will further devise models to tackle tasks from multiple areas of machine learning, such as Bioinformatics and Computer Vision (CV) in addition to NLP. Specifically, we plan to focus on biomedical question answering and object detection tasks from Bioinformatics and CV areas, respectively. We motivate the choice of these two domains as follows: Transfer learning with ImageNet achieved a huge success, and almost all state-of-the-art models for downstream tasks in CV make use of transfer learning. The abundance of transfer learning based models makes CV a good test-domain for evaluating the contributions we will propose for different pretraining schemes for transfer learning. On the contrary, applications of transfer learning is scarce in Bioinformatics compared to CV and NLP. Hence, there should be various task combinations that can benefit from transfer learning in the Bioinformatics domain that were not investigated before. This motivated us to choose Bioinformatics as a target domain to find new task combinations.

The remainder of this paper is structured as follows. Section 2 gives a summary of the related work on transfer learning and multi-task learning. This is followed by the Research Plan, where we explain the methodology we will use regarding each research question. Finally, Section 4 describes the evaluation methods and datasets that will be used to assess the significance of our contributions regarding each research question.

2 Related Work

Our research is related to the works in the subtopics we summarize below.

2.1 Transfer Learning

We follow the taxonomy defined by Ruder (2019) to differentiate between transfer learning and multi-task learning. Specifically, transfer learning is an umbrella term for settings where information from a source task is leveraged to improve the performance of a target task. If the target and source are learned simultaneously, this methodology is defined as 'multi-task learning', whereas if we employ a sequential learning of each task, this is referred to as 'sequential transfer learning'. For instance, in the domain of reinforcement learning, Rusu et al. (2016) proposed 'progressive neural networks' which learn each task sequentially and fixes the parameters for the subsequent tasks. On the contrary, Hashimoto et al. (2017) proposed a joint many task model to simultaneously learn multiple NLP tasks.

In the area of Computer Vision, sequential trans-

fer learning unlocked many potentials. Models pretrained on ImageNet are finetuned on the target task datasests (Krizhevsky et al., 2012) to achieve state-of-the-art results. Similarly, Peters et al. (2018) showed pretrained models improve performance across a wide range of NLP tasks. Radford et al. (2018) and Devlin et al. (2019) pretrained models over huge unlabeled datasets and these models are successfully applied to many downstream NLP tasks. However, Mou et al. (2016) showed that transferability depends largely on the semantic relatedness of the tasks. Finding related tasks is a key factor to achieve better transfer learning models, but a thorough understanding of how to find the most useful pretraining task is still missing (Ruder, 2019).

Another key factor to improve transferability is the selection of relevant data. Recently, Ruder and Plank (2017) proposed learning a similarity metric over the training sets by using Bayesian Optimization for transfer learning. Their work is limited to a domain adaptation setting where the source tasks are the same as the target task but the domains of the datasets are different. We propose extending their method to avoid *negative-transfer* in various multi-task settings.

2.2 Multi-task Learning

Ruder (2017) gives a comprehensive overview of multi-task learning models, where they define two main categories based on the information sharing methodology: *hard-sharing* and *soft-sharing*. In *Hard-sharing*, models contain a low-level layer which is shared among all task-specific layers, whereas in *soft-sharing* each model has its own weight set and regularization is applied to force these weights to be similar across all models. *Soft-sharing* based models are shown to benefit from multi-task learning when applied to related tasks. Yet, the benefits of this method are unclear for loosely related tasks.

Long and Wang (2015) attempted to learn the information flow between task-specific models. Ruder (2017) showed the effect of applying regularization to the network weights to generalize better. Using a more sophisticated approach to control the information flow and applying additional regularization terms on the network weights are promising ways to obtain improvements over the current models. Zhang et al. (2018) proposed learning the most suitable model for a given multi-task setting using the previous results obtained for various (S, M) pairs where S is a set of tasks and M is the learning model. They find the best candidate covariance matrix which represents the task relations to estimate the relative error for a new multi-task setting and show the effectiveness of their approach. One drawback of these approaches is that they focus only on learning the task-relatedness between tasks and ignore the architectural variations. Meyerson and Miikkulainen (2019) showed that architectures can also be decomposed to allow sharing of various sub-modules for a set of tasks. Yet, more research is necessary to find out the best method of sharing and the best architecture for a given multi-task setting.

3 Research Plan

In this section, we restate the research questions and explain the approach we are planning to take.

RQ1: How to optimize the model architecture and sharing methodology for a given task combination?

Currently proposed multi-task learners mostly use hard sharing, where models share a common low-level layer, and task-specific sharing methods are not analyzed for many task combinations (Collobert et al., 2011; Søgaard and Goldberg, 2016; Hashimoto et al., 2017). Following Long and Wang (2015), we plan to use learnable parameters to control the information flow between each task-specific model. Learning joint label embeddings for disparate label classes (Augenstein et al., 2018) is another promising approach that goes beyond hard-sharing. Specifically, we will apply this method to leverage our previously proposed joint learner for Dependency Parsing and Named Entity Recognition. Part-of-speech tags strongly correlate with named entities and dependencies (Hashimoto et al., 2017; Akdemir and Güngör, 2019b). Thus, we argue that learning joint label embeddings of these tasks can help to further capture the relations between them.

RQ2: What are some good auxiliary tasks to improve the performance of a target task?

Regarding this research question, we will fix a target task and try to improve the performance by 1) incorporating a transfer learning framework and 2) applying a more sophisticated data selection mechanism. To better understand the task relations (where to transfer from), we will compare the performance on a fixed target task using several auxiliary tasks

obtained through different task selection mechanisms. Lee et al. (2019) proposed pretraining the BERT model in the biomedical domain and apply the model to make predictions in several different downstream tasks in Bioinfomatics such as gene-disease relation extraction and biological named entity recognition. We argue that their approach can be combined with multi-task learning to further leverage the information available in the dataset of each task. Specifically, we claim that biological named entity recognition can be used as an auxiliary task to improve the performance of biological question answering systems. Our preliminary results are given in Section 4. The biological named entity dataset consists of several types of entities (genes, chemicals and disease mentions) and each type can be considered as a different task. We will use these set of tasks to compare the performance of the task selection mechanisms.

Deciding which data are useful (what to transfer), in addition to finding promising task combinations, is another key factor to increase transferability (Ruder and Plank, 2017). However, many of the current multi-task models use all the available data for all tasks (Long and Wang, 2015; Hashimoto et al., 2017; Lee et al., 2019). To this end, we will apply the previously proposed data selection mechanisms on our new task combinations to find the most useful and relevant examples from each dataset to improve the transferability and to avoid negative-transfer. Previous work on data selection successfully showed that using a Bayesian suite for deciding which data to use for multi-task learning brings significant improvements (Ruder and Plank, 2017). This motivated us to incorporate similar data selection mechanisms to further improve the performance of transfer learning models. We will compare several data selection mechanisms by fixing the model to be used and the task combination.

RQ3: How to find useful pretraining schemes?

The standard approach in sequential transfer learning is to pretrain a model using an objective that is relevant to and useful for the target task. In NLP, the prevailing method is to train a language model using the next sentence prediction and masked token prediction objectives over huge unlabeled datasets, e.g. the BERT model (Devlin et al., 2019). The pretrained models are usually fine-tuned on task-specific datasets, yet the characteristics of the downstream task are usually not considered during the pretraining process. Regarding this research question, our main goal is to find task-specific pretraining schemes and to compare the performance with fine tuned models that are not pretrained considering the downstream task (Lee et al., 2019).

Curriculum learning aims to find a good ordering of the training samples to go beyond random sampling (Bengio et al., 2009). The training samples are ordered according to their difficulties using prior knowledge. Recently, Jiang et al. (2015) proposed self paced curriculum learning which tries to learn this ordering dynamically during training to mitigate the drawbacks of defining static difficulties for training samples using external knowledge. Following this idea of changing the difficulty of the training samples (Bengio et al., 2009; Kumar et al., 2010; Jiang et al., 2015; Liang et al., 2016), we propose using 'adaptive masking' for pretraining language models. The standard approach for pretraining with masked language modeling involves predicting the distribution of a randomly masked word using its context (Devlin et al., 2019). Each masked word can be considered as an instance of a cloze test which is frequently used to assess the linguistic skills in humans. In a cloze test, students are expected to understand the context to fill in the masked word. Randomly selecting which words to mask causes the difficulty of each instance to change randomly as well. We propose adaptively changing the difficulty of the next training instance by observing the performance of the model. In this context, we define difficulty as the amount of contextual information necessary to select the most probable word, whereas Bengio et al. (2009) defined difficulty as the inverse of the frequency of each masked token regardless of their contexts. Table 1 illustrates why going beyond random masking is a promising method to improve the learning process. For the first example, the model (or the person tested) must predict 'school' from the context which includes the word 'students'. In the second example, the model must comprehend the overall negative meaning to predict 'low' instead of 'high'. [1] The idea can be extended easily to other domains of machine learning such as object detection where 'difficult words' are replaced with 'difficult objects'.

[1] The examples were taken from intermediate and advanced level cloze grammar tests from the englishlearner website: https://www.englishlearner.com/tests

Difficulty	Sentence
Intermediate	Two students from Cologne, Germany, ages 17 and 18, are accused of plotting an attack at their **school** on November 20.
Advanced	**Low** levels of **literacy** have a damaging impact on almost every aspect of adult life.

Table 1: Two example sentences for the masked language modeling task. The underlined tokens are the originally masked ones in the reference tests. Tokens that are more challenging to predict are shown in bold.

4 Evaluation

In order to evaluate the significance of our contributions, we will do evaluations for each research question separately. Below we give the evaluation methodology, together with example tasks and the related datasets that will be used for each research question.

4.1 RQ1.

We will compare our proposed methodology with the previously proposed multi-task learners and the state-of-the-art single-task learners in the same setting. We proposed a novel multi-task learning framework to improve the performance of the target task, Named Entity Recognition, using the information obtained from the auxiliary task, Dependency Parsing, for the Turkish language. Dependency Parsing is chosen as the auxiliary task following the previous work that showed the importance of dependencies for the Named Entity Recognition task, for morphologically rich languages, e.g the Turkish language (Güngör et al., 2018; Straka et al., 2019; Akdemir and Güngör, 2019a). The results in Table 2 show that our proposed model (Model 2) achieves an absolute 2.45% F-1 score overall improvement over the conventional joint learning model (Model 1). The conventional model requires a single dataset annotated with labels for both tasks, which is a delimiting constraint for less resourced languages. Instead, we proposed using separate datasets for each task (Akdemir and Güngör, 2019b) which allows the model to be trained on a larger dataset.

Next, we proposed a hierarchical multi-task learning framework (Akdemir et al., 2020) that builds on our previous work mentioned above. In this framework, each task-specific component is implemented following the state-of-the-art models and experiments are conducted using different sharing methodologies to find the most useful setting for this task combination. We followed Qi et al. (2018) and Lample et al. (2016) to implement a Highway Long Short Term Memory (H-LSTM) based dependency parser and a BiLSTM Conditional Random Fields based named entity recognizer. In addition, we used BERT subword contextual embeddings as the common low-level layer shared by the task-specific components. This framework achieved absolute improvements of **18.86%** and **4.61%** F-1 over our previously proposed model for DEP and NER tasks respectively. In addition, the framework showed absolute improvements of **1.44%** and **0.13%** F-1 over the state-of-the-art models for the Turkish language for DEP and NER tasks respectively. The details about the implementation and the experiments conducted are given in (Akdemir et al., 2020).

	Model 1	Model 2
PER	84.50	86.48
LOC	81.97	86.36
ORG	78.34	78.63
Overall	82.11	**84.56**

Table 2: Results comparing the proposed model (Model 2) with the conventional joint learner (Model 1). All results are given in percentage (%) F-1.

We will further test the validity of our hypothesis on other less resourced morphologically rich languages such as the Czech Language (Demir and Özgür, 2014).

Dataset. To test our hierarchical multi-task learner on the Czech Language, we will use the 'Czech Named Entity Corpus 2.0' (Ševčíková et al., 2007) for the NER task and the PDT-UD treebank (Hajič et al., 2017) of the 'CoNLL 2018 Shared Task' (Zeman et al., 2018) for Dependency Parsing task. The NER dataset contains 8,993 sentences with 35,220 entities and uses a two-level named entity classification. For our purposes it is sufficient to use the first level classes (10 classes) as the named entity labels, referred as *supertypes*. PDT-UD contains 87,913 sentences obtained mainly from newswire.

4.2 RQ2.

To evaluate the significance of the contributions we make regarding **RQ2**, we will fix a target task and compare the performance using the newly proposed auxiliary task(s). As mentioned in Section 3, an example target task is biomedical question answering. We argue that detecting and categorizing diseases and biological entities is an important first step to answer biological questions. In addition, the effect of applying data selection will be evaluated by fixing a deep learning model for the object detection task. It was chosen because there are numerous models already proposed for multi-task object detection which allows us to clearly assess the significance of our contributions.

Dataset. We use the BC2GM (Smith et al., 2008), BC4CHEMD (Krallinger et al., 2017), and BC5CDR (Li et al., 2016) datasets for biological named entity recognition which contain gene entities, chemical entities and disease mentions respectively. To test our claim, we use the BioASQ dataset (Tsatsaronis et al., 2015) used during the biomedical question answering competition which contains yes-no, factoid and list type questions.

The preliminary results we obtained for Biological Question Answering task can be seen on Table 3. [2] We started with BERT (Devlin et al., 2019) embeddings and obtained improvements through 1) transfer learning on the biomedical abstracts from PubMed, 2) pretraining the question answering module on the Squad question answering dataset and 3) training a multi-task learning model for all question types. Step 3 is our contribution and has not been employed before, to the best of our knowledge. We aim to show further improvements by incorporating multi-task learning of biological named entities.

	BioAsq-6b - Factoid		
Model	LAcc	SAcc	MRR
BERT (baseline)	0.24	0.35	0.28
+TL on PubMed	0.32	0.50	0.39
+pretraining on Squad	0.39	0.58	0.47
+MTL of all questions	**0.42**	**0.61**	**0.49**

Table 3: Initial results on Biological Question Answering-6 factoid type questions.

For multi-task object detection from different domains, we will use the Office-Caltech (Gong et al., 2012) dataset, which is the standard benchmark for transfer learning in Computer Vision. The Office dataset contains images from three different domains; Amazon, Webcam and DSLR, containing 31 categories. Caltech dataset is the 10 overlapping categories from the Caltech-256 dataset (Griffin et al., 2007).

[2]LAcc,SAcc and MRR are abbreviations for Lenient Accuracy, Strict Accuracy and Mean Reciprocal Rank, respectively.

4.3 RQ3.

We will evaluate our newly proposed pretraining schemes both performance-wise and resource-wise. We choose the standard pretraining objective of BERT (Devlin et al., 2019) as the baseline and we will train the same model using our newly proposed 'adaptive masking'.

Dataset. We will use the unlabeled Wikipedia articles in English for pretraining the model using both pretraining tasks. Next, we will evaluate the performance of the system on the benchmark 'The Stanford Question Answering Dataset', SQuAD 2.0, which contains over 150,000 answerable and unanswerable questions. We choose question answering as the downstream task, as it was used as the downstream task to evaluate the performance of BERT (Devlin et al., 2019) .

5 Summary

Transfer learning is a promising area of research for deep neural network based machine learning models. It helps achieve better generalization and utilization of the training datasets. In this paper, we pointed out the current key challenges and unsolved problems: 1) Going beyond the conventional way of hard-sharing in multi-task learning and finding the most useful architecture for a given setting, 2) Finding good auxiliary tasks in a multi-task setting for a specific target task, and 3) Finding useful pretraining schemes. Our research aims to apply the current work on transfer learning to new tasks and also find novel methods to obtain better multi-task learning models.

References

Arda Akdemir and Tunga Güngör. 2019a. A Detailed Analysis and Improvement of Feature-Based Named Entity Recognition for Turkish. In *International Conference on Speech and Computer*, pages 9–19. Springer.

Arda Akdemir and Tunga Güngör. 2019b. Joint Learning of Named Entity Recognition and Dependency

Parsing using Separate Datasets. *Computación y Sistemas*, 23(3).

Arda Akdemir, Ali Hürriyetoğlu, Erdem Yörük, Burak Gürel, Çağri Yoltar, and Deniz Yüret. 2018. Towards generalizable place name recognition systems: analysis and enhancement of NER systems on English News from India. In *Proceedings of the 12th Workshop on Geographic Information Retrieval*, page 8. ACM.

Arda Akdemir, Tetsuo Shibuya, and Tunga Gungor. 2020. Hierarchical multi task learning with subword contextual embeddings for languages with rich morphology.

Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. 2018. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1896–1906.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM.

Jerome Bruner. 1985. Child's talk: Learning to use language. *Child Language Teaching and Therapy*, 1(1):111–114.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Xiao Chu, Wanli Ouyang, Wei Yang, and Xiaogang Wang. 2015. Multi-task recurrent neural network for immediacy prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 3352–3360.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.

Hakan Demir and Arzucan Özgür. 2014. Improving named entity recognition for morphologically rich languages using word embeddings. In *2014 13th International Conference on Machine Learning and Applications*, pages 117–122. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE.

Gregory Griffin, Alex Holub, and Pietro Perona. 2007. Caltech-256 Object Category Dataset. *CalTech Report*.

Onur Güngör, Suzan Üsküdarli, and Tunga Güngör. 2018. Improving named entity recognition by jointly learning to disambiguate morphological tags. *arXiv preprint arXiv:1807.06683*.

Jan Hajič, Eva Hajičová, Marie Mikulová, and Jiří Mírovský. 2017. Prague dependency treebank. In *Handbook of Linguistic Annotation*, pages 555–594. Springer.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933.

Steven C Hayes, Dermot Barnes-Holmes, and Bryan Roche. 2002. Relational frame theory: A précis. In *Relational frame theory*, pages 141–154. Springer.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. 2015. Self-paced curriculum learning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, et al. 2017. Overview of the BioCreative VI chemical-protein interaction Track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

M Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Junwei Liang, Lu Jiang, Deyu Meng, and Alexander G Hauptmann. 2016. Learning to detect concepts from webly-labeled video data. In *IJCAI*, pages 1746–1752.

Mingsheng Long and Jianmin Wang. 2015. Learning multiple tasks with deep relationship networks. *arXiv preprint arXiv:1506.02117*, 2.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. 2016. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884.

Elliot Meyerson and Risto Miikkulainen. 2019. Modular Universal Reparameterization: Deep Multi-task Learning Across Diverse Domains. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7901–7912. Curran Associates, Inc.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How Transferable are Neural Networks in NLP Applications? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 479–489.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D Manning. 2018. Universal dependency parsing from scratch. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, National University of Ireland, Galway.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian Optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Magda Ševčíková, Zdeněk Žabokrtský, and Oldřich Krůza. 2007. Named entities in Czech: annotating data and developing NE tagger. In *International Conference on Text, Speech and Dialogue*, pages 188–195. Springer.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of BioCreative II gene mention recognition. *Genome biology*, 9(2):S2.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235.

Milan Straka, Jana Straková, and Jan Hajič. 2019. Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing and NER. In *International Conference on Text, Speech, and Dialogue*, pages 137–150. Springer.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task

transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. Conll 2018 sharedtask: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.

Yu Zhang, Ying Wei, and Qiang Yang. 2018. Learning to multitask. In *Advances in Neural Information Processing Systems*, pages 5771–5782.