Multi-Task Representation Learning

Mohamed-Rafik Bouguelia   Sepideh Pashami   Sławomir Nowaczyk*

Abstract

The majority of existing machine learning algorithms assume that training examples are already represented with sufficiently good features, in practice ones that are designed manually. This traditional way of preprocessing the data is not only tedious and time consuming, but also not sufficient to capture all the different aspects of the available information. With big data phenomenon, this issue is only going to grow, as the data is rarely collected and analyzed with a specific purpose in mind, and more often re-used for solving different problems. Moreover, the expert knowledge about the problem which allows them to come up with good representations does not necessarily generalize to other tasks. Therefore, much focus has been put on designing methods that can automatically learn features or representations of the data instead of learning from handcrafted features. However, a lot of this work used ad hoc methods and the theoretical understanding in this area is lacking.

1 Motivation

Representation learning is concerned with automatically transforming raw input data into representations or features that can be effectively exploited in machine learning tasks. Existing unsupervised approaches to representation learning such as [1, 2, 3, 4, 5] yield general features capturing dimensions of variation that may or may not be essential to a given task. On the other hand, supervised approaches to representation learning such as [6, 7, 8, 9, 10] can be overly specific as they allow to exclusively learn representations that help to discriminate among class labels related to a specific task. Nevertheless, such approaches have, especially recently, been extensively studied in the deep learning community [7, 11]. In this case, however, the learned representations cannot be directly applied to another task as they are explicitly tailored for a specific task.

The motivation for this research comes from our previous work on mapping of raw sensor data from Volvo trucks into low-dimensional representation, both in a supervised and unsupervised manner. Such a representation is needed for predictive maintenance solutions, as using the original raw data is not feasible. The overall goal is to extract general features which are suitable for more than one task, for example, estimating remaining useful life of several different components. Since those components can be related to different aspects of the truck operation, the representations that allow accurate predictions are related, but not necessarily the same. Achieving sufficient generality of the resulting features is not possible given current state of the knowledge in the field; more in-depth study of the underlying problem is needed before practical solutions can be developed.

We will contribute with extending the current representation learning methodology along two separate but interdependent directions. The first direction is considering training setup in which not one, but rather multiple related tasks are provided. This idea allows for a well-defined formalization of concepts such as complexity, diversity or incongruity among tasks to which the learned representation is expected to be applied to in the future. The second direction is aiming for a diverse set of representations, with clear and well-defined purposes and motives, instead of a single, all-encompassing one. Imposing such a meaningful structure onto the result allows for incremental generation and evaluation, as well as for explicit tradeoff between accuracy, generality and compression provided by the learned representation.

On the one hand, semi-supervised representation learning methods, like the one proposed in [12], aim to learn a representation based on few labeled data. However, the labels are still related to a single task and the result does not necessarily generalize well to multiple related ones. In parallel to representation

*Authors are with Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Sweden. Emails: mohbou@hh.se, seppas@hh.se and slanow@hh.se

learning, there has recently been considerable progress on the important problem of multi-task learning, which exploits similarities across several learning problems [13, 14]. The results show often significantly improve performance compared to learning each task independently. In the past few years different studies, such as [15, 16, 17, 18, 19], have advocated that a representation which captures properties that are invariant across tasks can significantly improve the performance. This generated an increasing interest in learning representations from multiple tasks, perhaps most noticeable in the computer vision domain. Nonetheless, despite the empirical success, formal justifications of why this works remain largely unexplored.

We formulate such multi-task representation learning problem as an extension of the classical supervised machine learning problem. The latter states that, given a set of training examples sampled according to an unknown underlying distribution $\mathcal{D}$, labelled by an unknown function $\mathcal{T}$, the goal is to find a hypothesis $\mathcal{H}$ that minimizes the probability of $\mathcal{H}$ and $\mathcal{T}$ disagreeing. The proposed extension states that, in addition to underlying distribution of data examples, now $\mathcal{D}_E$, there also exists an underlying unknown distribution over tasks, $\mathcal{D}_T$. A number of training tasks, sampled from this distribution, provide the (set of) labelling for the training data. The goal is to produce a group of representations with the lowest error on unseen data across all the expected tasks. A measure of the capacity (or expressive power) of a function family, such as the VC dimension [20], can be generalized to capture the essential complexity across multiple tasks. This way measures for approximating the true error, on unseen data across all the expected tasks, can be modeled depending on the a priori assumptions about the inherent difficulty of the particular problem instance. Transformations of raw input into features can be based on capturing different dimensions of variation in data, and either being essential for a particular class of tasks, or providing broad benefits by being generally useful for many different tasks. Starting from the establishment of theoretical foundations for this problem, the end-goal is to create an algorithm that efficiently produces such representations.

Representation learning has clearly demonstrated early success with deep learning in application areas such as computer vision, natural language processing and speech recognition. However, replicating those results in other domains has proven difficult, in part due to lack of sufficient theoretical foundations. It is well understood that the performance of machine learning methods is heavily dependent on the choice of the data representation (or features) on which they are applied. Unlike traditional feature engineering which requires labor-intensive effort, representation learning allows computers to autonomously create specific features which are appropriate for a particular problem. A good representation of data can also provide a substitute for storing raw data when dealing with big data in real-world applications. This research direction will have major impacts in various domains, as it enables building systems and algorithms that learn to perform new tasks based on experience gained from previous tasks. It will have major impact in a large number of application domains where machine learning is a key aspect; these includes text mining, patient healthcare data analysis, social network analysis, multiple object classification in computer vision, and predictive maintenance in the automotive industry, to mention but a few. We believe that it is a promising line of research, making progress towards real Artificial Intelligence.

2 Survey of the field and open challenges

Representation learning is challenging primarily for three general reasons: the immense space of possible solutions that should be considered; the difficulty in establishing a clear measurable objective for the learning process; and the insufficient understanding of how the properties of the problem instance match against the parameters of the representation learning process. This combination makes it hard to design efficient algorithms for determining which representation will ultimately be relevant for the expected distribution of tasks, as well as to propose a compelling theoretical foundation for such work.

The first of those main challenges is related to the generation of representations. Necessarily, for a given problem instance, a set of representations needs to be considered; such family of representations is usually generated by the same algorithm and corresponds to a certain family of functions. An important question for generating representations is that it is not clear whether there exists a single family of representations that is sufficient for all problems, or should multiple families be used, depending on the problem instance. Some comparative studies such as [21, 22, 23, 24] have been carried out for various application domains. However, so far, the properties that can influence the choice of one family of representations over another are still unknown [25]. Another aspect closely related to the generation of representations is about encouraging diversity among the set of representations. For example, it took quite some time after Breiman's 2001 paper [26] before modern ways to measure the diversity of trees within random forest were suggested. A similar development is needed for representation learning. A challenging aspect in this context is to balance specificity and diversity in an optimal way that leads to an improved accuracy. Promoting diversity among representations is very important, yet, it is not a well-studied aspect and there are no explicit metrics in the literature that allow capturing the diversity among a set of representations. To the best of our knowledge, the only directly relevant paper in this context is [27], which proposes a strategy to produce an ensemble of diverse representations specifically for the unsupervised case. This is done by controlling the trade-off between minimizing reconstruction error and maximizing diversity between reconstructions. A similar method has been applied in [28] for the problem of fall detection. However, both of those results are specific for autoencoders only and not directly applicable in a multi-task representation learning context.

The second main challenge is related to the evaluation of representations. One aspect that makes it different from most machine learning problems (such as classification), is the difficulty in establishing a clearly defined objective. The standard way of performing this evaluation is to measure the representation or feature learning algorithm in terms of its usefulness with respect to a particular task [23, 24]. This is typically done at regular intervals (to enable early stopping), by evaluating the performance of a cheap classifier trained using the learned features. However, a first issue is that regularly alternating between learning features and training a classifier produces a substantial computational overhead. That raises the important question of how to balance between extraction and classification stages. One principled solution to this problem is to use a tree of classifiers as proposed in [29]. In this solution, test inputs traverse along individual paths, where each path extracts different features for inputs that benefit from them the most. Similar methods include those proposed in [30, 31]. However, not only is this problem NP-hard, but most importantly, such methods give an incomplete evaluation of the features. In particular, the extension to the case of multi-task representation learning is challenging. As indicated in [25], these issues strongly motivate the use of unsupervised evaluation measures. For example, for auto-encoders [32, 33, 34], the reconstruction error on the test data can readily be used as an evaluation measure. However, such a measure can be unreliable because systems that learn more features as the time goes tend to overfit and systematically produce a lower test reconstruction error. Besides the accuracy, either with respect to a particular task or to a group of tasks, however, one can imagine a number of other criteria for assessing the quality and usefulness of a representation. For example, data compression has been studied in [35, 36], with the aim to transform the data into a compact but expressive form. Finally, the generality of representations across tasks is an important aspect that needs to be taken into account in multi-task representation learning; however, there is not yet any formal definition of this property. In particular, the aspect of balancing all those important metrics, both when generating and evaluating a multi-task representation, is usually not considered.

Finally, the third main challenge is understanding properties of the problem instance from the perspective of representation learning. There exist very few theoretical studies which expose problem properties that are relevant in the context of a multi-task representation learning. For example, the recent work [15] establishes theoretical results about the benefit of learning a representation from multiple tasks compared to learning from each task separately, based on basic properties such as the sample size, the data dimensionality, and the number of tasks. However, this work is only demonstrated for the specific case of subspace learning (i.e., linear feature learning), and does not take into consideration other relevant properties, such as the complexity of tasks or the similarity between tasks. Most existing multi-task representation learning methods such as [16, 17, 18, 19] assume that an expert can determine which tasks are related, or they implicitly assume that the available tasks are related and can be readily used to perform a joint training. However, in real-word, this assumption may not be always satisfied. Learning common representations across many unrelated or dissimilar tasks can lead to poor representations that decrease the performance compared to learning representations from each task separately, as discussed in [37, 38]. Therefore, similarity among tasks is an important problem property that needs to be taken into consideration.

3 Research questions

The challenges discussed above are very broad, therefore, we propose the following five main research questions as a good starting point.

> *1. What problem instance properties require the generation of multiple diverse representations over generating a single multi-dimensional representation?*

Usual representation learning methods assume that the goal is learning a single low-dimensional representation of the data. However, in many cases it can be more beneficial to create a set of several independent representations. A set of multiple representations can always be seen as one higher-dimensional representation; yet, having multiple representations provides an explicit structure that can be exploited in various ways. In particular, if a set of representations is expected to be useful across a wide spectrum of tasks, such a structure offers a convenient way of expressing the trade-off between accuracy on each task and the diversity of representations within this set. In this context, the questions that need answering are when to learn multiple representations, how many of them should be created, and how to promote the diversity within the set of representations on several different levels (e.g., on the algorithm, the mapping and the data levels).

> *2. How to generate a family of representations which ensures a good coverage of the space of mapping functions that are appropriate for a given problem instance?*

In practice, the representations are necessarily generated according to some algorithm, which induces a particular family of representations. In order to establish the theoretical foundations for the multi-task representation learning problem, it is crucial that one can measure whether this family is expressive enough to provide sufficient coverage, appropriate for the given problem instance. The way of connecting the expressiveness of the family with the complexity of the expected tasks it should be used on is going to be an important contribution. An essential research question is: can existing concepts such as VC dimension be extended to capture this match? One idea is to use different subsets of training tasks to achieve high diversity of representations and to get a reliable estimation of the expected coverage that can be achieved. However, the actual concrete method for doing that needs to be developed. Further, how to generate a large set of non-redundant representations, from simple to more complex ones, which are able to approximate any and all of the expected future task?

> *3. Does identifying and grouping related tasks, followed by learning multiple representations separately for each group, lead to improved outcome?*

Existing results [15] establish the benefits of learning a representation from multiple related tasks compared to learning from each task separately. At the same time, it has been shown that trying to learn a common representation across unrelated or dissimilar tasks can decrease the performance [37, 38]. Therefore, a natural question is whether it is possible to automatically find an appropriate partitioning of tasks that leads to learning better representations from tasks within each group? Under what conditions are such a step necessary? Establishing that will require new advancements in determining how the similarity among tasks should be measured, and which properties of the problem instance affect it to a different degree.

> *4. How to evaluate representations in a multi-task setting based on several aspects such as the compactness of representations, the accuracy relative to each individual task, and the generality of across all tasks?*

Being able to evaluate representations is absolutely crucial, however, today in many cases it is done in an ad hoc manner. There is a need to define measures based on a well justified theoretical description of the problem at hand. In this context, an important question is, how can one establish conditions of whether a representation or set of representations is "good enough" for any given set of tasks? In particular, a good representation can be defined as a one which leads to a low prediction error (i.e., high accuracy) over the whole population of data and expected tasks (according to the unknown underlying distribution). Hence, given reasonable assumptions, which evaluation measures can be proven to be good approximations of the true error in multi-task representation learning? Moreover, as the data compression plays a non-negligible role in the context of representation learning, it is important to select a reasonably small number of representations that are common across tasks, while improving the accuracy relative to learning each task independently. Therefore, another question relates to the number (or ratio) of representations that are "sufficient" to improve the accuracy, across a set of tasks. More generally, how can we quantify the expected benefit of representation learning from multiple tasks? Finally, how to balance all these different aspects of the evaluation: the data compression (or compactness of representations), the accuracy relative to each task, and the generality of representations across tasks?

> *5. How to define the complexity of tasks (in addition to other problem instance properties) to address all the above questions in a principled way?*

In order to answer all the above research questions in a principled way, problem instance properties that potentially influence the choice of algorithms, representation families and quality measures need to be defined and formalized. The success of a set of representations for a given problem instance is related to the difficulty or complexity of tasks that one needs to deal with. In this context, the complexity of tasks is the most important property, which leads to an important question: what would be a good measure for the complexity of tasks? What is the equivalent of VC dimension for a family of tasks? Necessarily, such a measure needs to consider the similarity between tasks. Even though each task within a training set can be simple, if these tasks are very different, one may need a quite diverse and expressive family of representations. On the contrary, a much simpler family of representations can be sufficient for a set of very individually difficult but overall similar tasks. Task complexity is of course only one, even if arguably the most important, property of the problem that needs to be studied. Other examples include the amount of noise, the size (and overlap) of data available for each task as well as the dimensionality and the heterogeneity of the data.

4 Methods, approaches and ideas

Defining relevant properties of the problem. The starting point is to define the most relevant properties that can be used to describe a problem instance, for example to investigate a measure for estimating the complexity of expected tasks based on the available training tasks. One possibility is generalizing the VC dimension [20], which is related to the inherent complexity of a space, for a set of tasks. The goal is to capture how difficult are the tasks we are expecting to have to deal with in the future. Another is a measure for modeling the similarity between tasks, possibly modeled either based on a direct comparison between parameters learned from the different tasks, or based on how well the parameters learned from one task, perform other tasks. Based on such properties one can describe the problem instance, together with additional basic features such as the number of tasks, the dimensionality of the original data representation, the level of noise, and the data size per task. Those properties can lead to an upper bound on the performance of representations across tasks. For example, the PAC learning framework [39] (Probably Approximately Correct learning) enables mathematical analysis of machine learning which stipulates that with high probability, a learned hypothesis (e.g., a classification model) will have low generalization error for a given classification task. PAC learning does not take into consideration the possible existence of multiple related tasks, nor does it concern itself with data representations.

Generating representations. A strategy for generating adequate families of representations can be based on efficient methods for generating large sets of representations that are able to approximate or represent any function of certain properties. For example, the field of Functional Data Analysis primarily focuses on smooth functions, which is probably too broad for our needs. The coverage of the space of functions by the generated representations will be measured in order to ensure a trade-off between the "exploration of all possible representations" and the "exploitation of the best generated representations". In order to produce more useful representations, new methods for determining which tasks are related and therefore can be automatically grouped together based on the similarity, are needed.

Encouraging diversity among representations. The generation of representations should be directed by an evaluation process which allows selecting, among all the possible representations, the ones that fulfill a range of assessment criteria, in particular, preservation of diversity among the generated representations and generality of the representations across tasks. First, on the algorithm level: if representations are created by sufficiently different algorithms, they are likely to be different. This can be done, for example, by explicitly controlling the bias and variance for a family of algorithms. Second, on the mapping level: representations are functions from one feature space to another, and those functions can be compared based on their mathematical properties, as defined in either Hilbert or Banach spaces. Third, on the example level: one can measure how do the relative positions change for the data points in the training sets. All of these can be done in either supervised and unsupervised manner, or as a combination of both approaches.

Algorithm for multi-task representation learning. The final goal, clearly, is an efficient algorithm which benefits from the results of the other work packages. The algorithm takes as input a dataset and a set of tasks, and produces as output a set of representations that are expected to generalize well across unseen tasks.

5 Preliminary results

This idea builds on our previous work of evaluating several approaches for both supervised and unsupervised mapping of raw sensor data from Volvo trucks into low-dimensional representation. Such a representation is needed for predictive maintenance solution, as using the original raw data is not feasible. The overall goal is not to find the best low-dimensional representation tailored to a very specific task, but rather to identify the method for learning a widely applicable representation.

For example, general low-dimensional representations of the data are calculated to find various truck configuration. Data originates from 79974 unique Volvo trucks and is recorded during a full year. The data of a single truck is represented with a bivariate histogram, where the axes correspond to a pair of sensors: turbocharger speed vs boost pressure. Each task describes various truck configurations, e.g., engine, gearbox, country of operation or brand, while the bivariate histograms describe the usage of the truck. We have performed a comparison of techniques based on t-distributed stochastic neighbor embedding (t-SNE) and convolutional autoencoders (CAE) in a supervised fashion over 74 different 1-vs-Rest tasks using random forest. The results show that t-SNE is most effective for 2D and 3D, while CAE could be recommended for 10D representations. Fine-tuning of the results shows slight improvement using low-dimensional representation in comparing to the original data representation.

6 References

[1] Coates, A., Lee, H., & Ng, A. Y. (2010). An analysis of single-layer networks in unsupervised feature learning. Ann Arbor, 1001(48109), 2.

[2] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

[3] Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. ICML Unsupervised and Transfer Learning, 27, 17-36.

[4] Bengio, Y., Yao, L., Alain, G., & Vincent, P. (2013). Generalized denoising auto-encoders as generative models. In Advances in Neural Information Processing Systems.

[5] Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., & Brox, T. (2014). Discriminative unsupervised feature learning with convolutional neural networks. In Advances in Neural Information Processing Systems.

[6] Zhuang, F., Cheng, X., Luo, P., Pan, S. J., & He, Q. (2015, July). Supervised Representation Learning: Transfer Learning with Deep Autoencoders. In IJCAI.

[7] Bengio, Y. (2013, July). Deep learning of representations: Looking forward. In International Conference on Statistical Language and Speech Processing (pp. 1-37). Springer Berlin Heidelberg.

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[9] Zou, F., Wang, Y., Yang, Y., Zhou, K., Chen, Y., & Song, J. (2015). Supervised feature learning via l 2-norm regularized logistic regression for 3d object recognition. Neurocomputing, 151, 603-611.

[10] Fan, H., Cao, Z., Jiang, Y., Yin, Q., & Doudou, C. (2014). Learning deep face representation. arXiv preprint arXiv:1403.2802.

[11] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

[12] Banijamali, E., & Ghodsi, A. (2016). Semi-Supervised Representation Learning based on Probabilistic Labeling. arXiv preprint arXiv:1605.03072.

[13] Evgeniou, T., Micchelli, C. A., & Pontil, M. (2005). Learning multiple tasks with kernel methods. Journal of Machine Learning Research, 6(Apr), 615-637.

[14] Evgeniou, T., & Pontil, M. (2004, August). Regularized multi–task learning. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 109-117). ACM.

[15] Maurer, A., Pontil, M., & Romera-Paredes, B. (2016). The benefit of multitask representation learning. Journal of Machine Learning Research, 17(81), 1-32.

[16] Gong, P., Zhou, J., Fan, W., & Ye, J. (2014, August). Efficient multi-task feature learning with calibration. In Proceedings of the 20th ACM SIGKDD (pp. 761-770).

[17] Zhao, H., Stretcu, O., Negrinho, R., Smola, A., & Gordon, G. (2017). Efficient Multi-task Feature and Relationship Learning. arXiv preprint arXiv:1702.04423.

[18] Argyriou, A., Evgeniou, T., & Pontil, M. (2008). Convex multi-task feature learning. Machine Learning, 73(3).

[19] Argyriou, A., Evgeniou, T., & Pontil, M. (2007). Multi-task feature learning. Advances in neural information processing systems, 19, 41.

[20] Vapnik, V. N., & Chervonenkis, A. Y. (2015). On the uniform convergence of relative frequencies of events to their probabilities. In Measures of Complexity (pp. 11-30). Springer International Publishing.

[21] Renshaw, D., Kamper, H., Jansen, A., & Goldwater, S. (2015). A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge. In INTERSPEECH (pp. 3199-3203).

[22] Cruz-Roa, A., Arevalo, J., Basavanhally, A., Madabhushi, A., & Gonzlez, F. (2015, January). A comparative evaluation of supervised and unsupervised representation learning approaches for anaplastic medulloblastoma differentiation. In Tenth International Symposium on Medical Information Processing and Analysis (pp. 92870G-92870G). International Society for Optics and Photonics.

[23] Tokarczyk, P., Montoya, J., & Schindler, K. (2012, July). An evaluation of feature learning methods for high resolution image classification. In ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 22nd ISPRS Congress, Melbourne, Australia.

[24] Shao, L., Cai, Z., Liu, L., & Lu, K. (2017). Performance evaluation of deep feature learning for RGB-D image/video classification. Information Sciences, 385, 266-283.

[25] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1798-1828.

[26] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[27] Reeve, H. W., & Brown, G. (2015). Modular Autoencoders for Ensemble Feature Extraction. NIPS 2015 Workshop on Feature Extraction: Modern Questions and Challenges. JMLR W&CP, volume 44, 2015.

[28] Khan, S. S., & Taati, B. (2016). Detecting Unseen Falls from Wearable Devices using Channel-wise Ensemble of Autoencoders. arXiv preprint arXiv:1610.03761.

[29] Xu, Z. E., Kusner, M. J., Weinberger, K. Q., & Chen, M. (2013, June). Cost-Sensitive Tree of Classifiers. In ICML (1) (pp. 133-141).

[30] Khan, S. H., Bennamoun, M., Sohel, F., & Togneri, R. (2015). Cost sensitive learning of deep feature representations from imbalanced data. arXiv preprint arXiv:1508.03422.

[31] Xu, Z. E., Kusner, M. J., Huang, G., & Weinberger, K. Q. (2013). Anytime Representation Learning. In ICML (3) (pp. 1076-1084).

[32] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11(Dec), 3371-3408.

[33] Masci, J., Meier, U., Cirean, D., & Schmidhuber, J. (2011, June). Stacked convolutional auto-encoders for hierarchical feature extraction. In International Conference on Artificial Neural Networks (pp. 52-59).

[34] Rifai, S., Vincent, P., Muller, X., Glorot, X., & Bengio, Y. (2011). Contractive auto-encoders: Explicit invariance during feature extraction. In Proceedings of the 28th International Conference on Machine Learning (ICML-11).

[35] Gregor, K., & LeCun, Y. (2011). Learning representations by maximizing compression. arXiv preprint arXiv:1108.1169.

[36] Gregor, K., Besse, F., Rezende, D. J., Danihelka, I., & Wierstra, D. (2016). Towards conceptual compression. In Advances In Neural Information Processing Systems.

[37] Kang, Z., Grauman, K., & Sha, F. (2011). Learning with whom to share in multi-task feature learning. In Proceedings of the 28th International Conference on Machine Learning (ICML-11) (pp. 521-528).

[38] Kumar, A., & Daume III, H. (2012). Learning task grouping and overlap in multi-task learning. arXiv preprint arXiv:1206.6417.

[39] Long, P. M. (1995). On the sample complexity of PAC learning half-spaces against the uniform distribution. IEEE Transactions on Neural Networks, 6(6), 1556-1559.