Exploration Reduction by Selecting a Hierarchical Order of Implicit Author Demographic Characterizations

Chung-Chi Chen[1], Hen-Hsen Huang[2] and Hsin-Hsi Chen[3]

[1]AIST, Japan
[2]Institute of Information Science, Academia Sinica, Taiwan
[3]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan

Abstract

This paper focuses on the selection of hierarchical orders in multi-task architectures, a significant challenge in developing neural network architectures. We propose a systematic methodology based on the statistical results of the Apriori algorithm to arrange the order of co-training tasks. Our findings demonstrate that this approach can provide near-optimal performance, significantly reducing the exploration times in multi-task scenarios. The models developed using this methodology surpass state-of-the-art performances in flu vaccination intent prediction and music review sentiment analysis tasks, demonstrating its efficacy.

Keywords

Hierarchical order, demographic characterization, Exploration reduction,

1. Introduction

The development of neural network architectures frequently necessitates a significant degree of trial-and-error, in addition to substantial computational time. State-of-the-art models across various tasks often require thousands of GPU days for training [1]. The environmental and computational expenses associated with such complex models present serious challenges [2]. Therefore, developing guidelines to enhance performance and reduce iterative testing becomes a critical area of exploration.

A key area of computational intensity arises when probing the hierarchical order of a hierarchical multi-task architecture. Given four candidate tasks within an architecture, to obtain optimal performance, we would need to experiment with all 24 possible hierarchical orders. Previous works have demonstrated the efficacy of multi-task architectures in natural language processing (NLP) tasks [3, 4, 5, 6]. However, the literature remains sparse in providing insights on choosing the hierarchical order for these architectures. This paper attempts to bridge this gap, offering a systematic analysis for selecting an optimal hierarchical order for multi-task architectures.

In: M. Litvak, I.Rabaev, R. Campos, A. Jorge, A. Jatowt (eds.): Proceedings of the IACT'23 Workshop, Taipei, Taiwan, 27-July-2023

c.c.chen@acm.org (C. Chen); hhhuang@iis.sinica.edu.tw (H. Huang); hhchen@ntu.edu.tw (H. Chen)

Table 1
Statistics of Twitter FV.

Flu Vaccination		Region		Gender		Age	
Taken/Plan	502	Midwest	227	Female	835	<= 30	775
No Intent	832	Northeast	427	Male	499	> 30	559
		South	412				
		West	268				

The crux of multi-task architectures lies in sharing learned embeddings and information across tasks. We propose an approach based on the Apriori algorithm's statistical results [7] to organize the order of the co-training tasks. Our findings suggest that the hierarchical order proposed by our approach achieves near-optimal performance, significantly reducing the number of required exploration iterations.

The contributions of this paper are three-fold:

1. We highlight a crucial intersection between sustainable NLP and multi-task learning.
2. We propose an efficient method for arranging the hierarchical order in multi-task architecture, offering near-optimal performance with fewer explorations.
3. The models developed using our methodology outperform state-of-the-art performances in flu vaccination intent prediction [8] and music review sentiment analysis [9] tasks.

2. Related Work

The human learning process often involves sharing information or experience across tasks, an idea reflected in multi-task learning architecture. This concept has seen success in diverse applications, such as computer vision [10] and NLP [11]. However, decisions regarding what and how to share remain open questions. A comprehensive survey of multi-task learning is provided by Zhang and Yang [12]. The field typically bifurcates into hard-sharing and soft-sharing methods, as overviewed by Ruder [13]. In both cases, a majority of previous works have focused on information sharing within the encoder [14, 15, 16]. This paper instead offers a guideline for information sharing by designing a hierarchical architecture, particularly focusing on the learning order selection. Bidirectional Encoder Representations from Transformers (BERT) have revolutionized the NLP field [17]. Researchers have leveraged BERT and other pre-trained text-encoders to set new standards on several NLP tasks [18, 19, 20]. This work uses BERT as an encoder and delves further into the issue of hierarchical order selection.

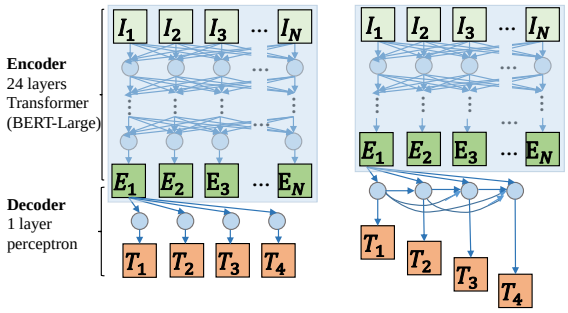
3. Datasets

Our experiments utilize two publicly available datasets. Each dataset contains four labels for a single input sample, therefore, we treat our task setting as a four-label classification problem.

The first dataset, denoted as Twitter FV, is sourced from Twitter [8]. The corresponding task involves predicting whether the author of a tweet has already received a flu vaccination or intends to do so. The second dataset, referred to as Amazon Sentiment, is collected from

Table 2
Statistics of Amazon Sentiment.

Sentiment		Region		Gender		Age	
Positive	27,035	Midwest	4,641	Female	6,665	<= 30	6,922
Negative	4,961	Northeast	9,822	Male	25,331	> 30	25,074
		South	8,956				
		West	8,577				



(a) Vanilla multi-task (b) Hierarchical multi-task

Figure 1: Structures of multi-task architectures. I_i and E_i denote the ith input token and the embedding of the ith input token, respectively. T_j denotes the prediction of Task j.

Amazon music reviews [9]. The task for this dataset is to label given reviews as either positive or negative. Huang and Paul [9] identified a correlation between the writer's demographic factors (region, gender, and age) and the tasks of flu vaccination intent and sentiment analysis. Accordingly, we incorporate these demographic features from their dataset as labels for auxiliary tasks. The statistics for both the Twitter FV and Amazon Sentiment datasets are detailed in Table 1 and Table 2, respectively.

4. Methods

4.1. Models

For individual task training performance testing, we adopt BERT-Large [17], a 24-layer Transformer [21]. We compare the standard multi-task architecture depicted in Figure 1 (a) with the hierarchical multi-task architecture shown in Figure 1 (b). We preprocess the input text using WordPiece [22] to obtain the input embedding, I. Post BERT-Large encoding, we acquire the token embedding $E \in \mathbb{R}^{1024}$. Following the classification task setup in previous work [17], we utilize the first token embedding of an input instance (E_1) to represent the encoded information. Subsequently, a one-layer perceptron is adopted for decoding and prediction. The Adam optimizer [23] is used for stochastic optimization, employing the cross-entropy loss function.

Table 3
Lift in each dataset.

	Region				Gender		Age	
	Midwest	Northeast	South	West	Female	Male	<= 30	> 30
Twitter FV - Taken	1.1824	0.8339	1.0578	1.0213	1.1393	0.7669	1.2344	1.1961
Sentiment - Positive	1.0203	0.9614	1.0133	1.0193	1.0203	0.9947	0.9901	1.0027

Table 4
I-Score in each dataset.

	Region	Gender	Age
Twitter FV - Taken	10.69%	18.62%	21.52%
Sentiment - Positive	2.29%	1.28%	0.63%

4.2. Hierarchical Order Selection Approach

In a hierarchical multi-task architecture, the optimal task order selection remains an open question. Exhaustively exploring all hierarchical orders to achieve the best performance is an obvious yet highly inefficient approach. This paper proposes a method based on the Apriori algorithm [7] for selecting the hierarchical order.

The Apriori algorithm is typically employed for association rule learning, with market basket analysis being a common application. Using this algorithm, given an item in a customer's basket (for instance, a bottle of milk), we can calculate the probability of another item (like cereal) also being included in the basket, based on previous transaction statistics. In our method, we treat each label as an individual item and compute the $Lift$ of a given label towards other labels as defined in Equation 1.

$$Lift(L_i, L_j) = \frac{Support(L_i \cap L_j)}{Support(L_i) \times Support(L_j)}, \tag{1}$$

Here, $Support(\cdot)$ denotes the frequency of the given label set in the dataset, and L denotes the label set. Table 3 presents the Lift between different labels in each dataset.

To estimate the informativeness of each auxiliary task, we further calculate the informativeness score (*I-Score*) using Equation 2.

$$\textit{I-Score}(Task_i|L_j) = \frac{\sum_{t=1}^{N} \mid Lift(L_t^{Task_i}, L_j) - 1 \mid}{N}, \tag{2}$$

In this equation, N denotes the number of labels in $Task_i$. The principle behind the *I-Score* is that whether the target label has a positive (> 1) or negative (< 1) correlation to L_j, the further it is from 1, the more information the target label provides. Table 4 displays the *I-Score* for each dataset.

We recommend arranging the auxiliary tasks in an ascending order of *I-Score*. The proposed approach's suggested hierarchical orders for both datasets are listed in Table 5. Consequently, given the order of auxiliary tasks, we only need to explore four hierarchical orders instead of probing all possible 24 combinations.

Table 5
Suggested hierarchical orders. fv, s, r, g, and a denote the flu vaccination intent detection, sentiment analysis, region, gender, and age tasks, respectively.

Twitter FV	Amazon Sentiment
fv-r-g-a	s-a-g-r
r-fv-g-a	a-s-g-r
r-g-fv-a	a-g-s-r
r-g-a-fv	a-g-r-s

Table 6
Experimental results of Twitter FV.

Model	Flu Vaccination (fv)	
	Macro	Weight
NUFA+w [9]	85.41	**87.46**
Single-task BERT	86.51	87.29
Vanilla multi-task	85.83	86.75
Hierarchical multi-task (g-a-fv-r)	**86.66**	87.40

Table 7
Experimental results of Amazon Sentiment. * denotes the results are significantly better than the second-best result (ranking based on macro-averaged F1-score) at $p < 0.05$ using McNemar's test.

Model	Sentiment (s)	
	Macro	Weight
NUFA+w [9]	66.74	83.54
Single-task BERT	66.25	82.32
Vanilla multi-task	64.59	83.49
Hierarchical multi-task (a-r-s-g)	**70.43***	**85.21***

5. Experiments

Huang and Paul [9] divide the dataset into training and test sets randomly but do not specify the indices for splitting. Given Gorman and Bedrick's findings [24], single "standard split" results may not be reliable. Thus, we employ five-fold cross-validation to gauge each model's performance. To ensure reproducibility, the splitting indices are provided in the supplementary materials.[1] We report the average macro-F1 score. Since Huang and Paul [9] use the weighted F1-score on the Twitter FV and Amazon Sentiment datasets, we also report results using this metric.

5.1. Comparison with Baselines

In this section, we juxtapose the performance of the best hierarchical model against other baseline models. For the Twitter FV and Amazon Sentiment datasets, we utilize NUFA+w [9] as a baseline, a BiLSTM-based multi-task architecture. We also employ single-task BERT as

[1] http://explorationreduction.nlpfin.com/

Table 8
Performance of hierarchical multi-task architecture with suggested hierarchical orders. The bold results are not significantly different from the best performance. FV and S denote the flu vaccination intent detection and sentiment analysis tasks.

		Order	Macro F1	Rank
FV	Best	g-a-fv-r	86.66	1/24
	Suggested	fv-r-g-a	**86.00**	2/24
		r-g-fv-a	85.66	4/24
		r-g-a-fv	85.30	12/24
		r-fv-g-a	84.39	19/24
S	Best	a-r-s-g	70.43	1/24
	Suggested	a-g-s-r	**70.12**	2/24
		a-g-r-s	69.33	14/24
		s-a-g-r	69.29	16/24
		a-s-g-r	65.00	22/24

a robust baseline for all datasets. Table 6 and Table 7 present the experimental results, also detailing the hierarchical order. The parenthesized information denotes the hierarchical order from Task 1 to Task 4. We observe that the hierarchical architecture consistently outperforms across all datasets.

Interestingly, the vanilla multi-task architecture's performance trails behind that of single-task BERT in both the Twitter FV and Amazon Sentiment datasets. This finding suggests that when the encoder is merely fine-tuned without shared information between the task-specific components, some task-specific information may not be effectively learned by the models.

5.2. Performance with Suggested Orders

Table 8 displays the performance of the suggested orders. We find that the nearly optimal performance is achieved by exploring just four recommended hierarchical orders. Notably, this near-optimal performance does not significantly deviate from the best performance, achieved by probing all hierarchical orders. These results validate the efficacy of our proposed approach for hierarchical order selection and highlight the importance of prior knowledge about the dataset and labels. With our approach, attaining near-optimal performance requires exploring only a sixth of all possible permutations in four-task cases.

6. Discussion

6.1. Correlation between the Tasks

The performance of the hierarchical multi-task architecture is found to be on par with single-task BERT according to Table 6. However, Table 7 shows a significant difference between the performances of the hierarchical multi-task architecture and single-task BERT. We provide an in-depth analysis of this phenomenon in this section.

We execute ordinary least squares regression (OLS) on the task performances in both datasets.

Table 9
OLS Results. DV and IV denote the dependent and independent variables. std err denotes the standard error. t and P denote the t-value and p-value. The confidence level is set as 95%.

DV	IV	coef	std err	t	P > | t |
FV	Region	0.15	0.12	1.33	0.19
	Gender	0.04	0.06	0.79	0.43
	Age	0.00	0.09	-0.05	0.96
S	Region	0.99	0.04	26.91	0.00
	Gender	1.16	0.06	18.09	0.00
	Age	0.80	0.13	6.33	0.00

The OLS inputs consist of experimental results from all hierarchical orders, resulting in 24 distinct experimental results. Table 9 presents the statistics. We find that the performance of flu vaccination intent detection is not significantly correlated with the performance of all auxiliary tasks. This indicates that while demographic information proves useful in BiLSTM-based architectures [9], it may not be beneficial for the flu vaccination intent detection task in a BERT hierarchical multi-task architecture. Conversely, demographic information remains useful for music review sentiment analysis, as performance improvements in auxiliary tasks also enhance the sentiment analysis task's performance.

6.2. Limitations

While our study offers initial insights into optimizing hierarchical order selection in multi-task architectures, it does have certain limitations. First, we did not analyze all multi-task setting datasets due to the sheer volume of possibilities. Another potential limitation is our omission of GPU cost calculations in our study. However, the inferred reduction in carbon dioxide emissions stemming from decreased exploratory iterations is an important point to note. Assuming identical datasets and models, our proposed method implies that nearly optimal performance can be achieved with only one-sixth of the carbon dioxide emissions associated with exhaustive exploration. Lastly, the performance correlation between tasks observed in our study could be context-specific. As demonstrated, the demographic information was useful in the context of music review sentiment analysis but not in the flu vaccination intent detection task. This finding suggests that not all information may be universally useful across different tasks in a hierarchical multi-task learning setup.

7. Conclusions

This paper presented a systematic analysis for selecting an optimal hierarchical order for multi-task architectures, addressing a gap in the literature. We proposed an approach based on the Apriori algorithm to organize task order and demonstrated that the resulting hierarchical order achieves near-optimal performance while considerably reducing the number of exploration iterations.

8. Acknowledgments

This research is supported by National Science and Technology Council, Taiwan, under grants 110-2221-E-002-128-MY3, 110-2634-F-002-050-, and 111-2634-F-002-023-. The work of Chung-Chi Chen was supported in part by JSPS KAKENHI Grant Number 23K16956.

References

[1] H. Liu, K. Simonyan, Y. Yang, DARTS: Differentiable architecture search, in: International Conference on Learning Representations, 2019.

[2] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019.

[3] X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4487–4496.

[4] R. Masumura, Y. Shinohara, R. Higashinaka, Y. Aono, Adversarial training for multi-task and multi-lingual joint modeling of utterance intent classification, in: EMNLP, Brussels, Belgium, 2018, pp. 633–639.

[5] V. Sanh, T. Wolf, S. Ruder, A hierarchical multi-task approach for learning embeddings from semantic tasks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 6949–6956.

[6] K. Hashimoto, C. Xiong, Y. Tsuruoka, R. Socher, A joint many-task model: Growing a neural network for multiple NLP tasks, in: EMNLP, Copenhagen, Denmark, 2017, pp. 1923–1933.

[7] R. Agrawal, R. Srikant, et al., Fast algorithms for mining association rules, in: Proc. 20th int. conf. very large data bases, VLDB, volume 1215, 1994, pp. 487–499.

[8] X. Huang, M. C. Smith, M. J. Paul, D. Ryzhkov, S. C. Quinn, D. A. Broniatowski, M. Dredze, Examining patterns of influenza vaccination in social media, in: Workshops at the Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[9] X. Huang, M. Paul, Neural user factor adaptation for text classification: Learning to generalize across author demographics, in: Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019), 2019, pp. 136–146.

[10] S. Liu, E. Johns, A. J. Davison, End-to-end multi-task learning with attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1871–1880.

[11] M. Luong, Q. V. Le, I. Sutskever, O. Vinyals, L. Kaiser, Multi-task sequence to sequence learning, in: Y. Bengio, Y. LeCun (Eds.), 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016.

[12] Y. Zhang, Q. Yang, A survey on multi-task learning, arXiv preprint arXiv:1707.08114 (2017).

[13] S. Ruder, An overview of multi-task learning in deep neural networks, arXiv preprint arXiv:1706.05098 (2017).
[14] S. Maharjan, J. Arevalo, M. Montes, F. A. González, T. Solorio, A multi-task approach to predict likability of books, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 1217–1227.
[15] R. Masumura, Y. Shinohara, R. Higashinaka, Y. Aono, Adversarial training for multi-task and multi-lingual joint modeling of utterance intent classification, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 633–639.
[16] M. S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, P. Bhattacharyya, Multi-task learning for multi-modal emotion recognition and sentiment analysis, 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), 2019, pp. 370–379.
[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL, 2019, pp. 4171–4186.
[18] C. Alberti, D. Andor, E. Pitler, J. Devlin, M. Collins, Synthetic QA corpora generation with roundtrip consistency, in: ACL, 2019, pp. 6168–6173.
[19] K. Clark, M.-T. Luong, U. Khandelwal, C. D. Manning, Q. V. Le, BAM! born-again multi-task networks for natural language understanding, in: ACL, 2019, pp. 5931–5937.
[20] J. Straková, M. Straka, J. Hajic, Neural architectures for nested NER through linearization, in: ACL, 2019, pp. 5326–5331.
[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
[22] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation, arXiv preprint arXiv:1609.08144 (2016).
[23] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[24] K. Gorman, S. Bedrick, We need to talk about standard splits, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019.