ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096604

MATCHING-BASED TERM SEMANTICS PRE-TRAINING FOR SPOKEN PATIENT QUERY UNDERSTANDING

Zefa Hu[1,2]*, Xiuyi Chen*[1,2]*, Haoran Wu*[1,2]*, Minglun Han*[1,2]*, Ziyi Ni*[1,2]*, Jing Shi*[2]*, Shuang Xu*[2]*, Bo Xu*[1,2]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2]Institute of Automation, Chinese Academy of Sciences, Beijing, China
{huzefa2018, chenxiuyi2017, wuhaoran2018, hanminglun2018, niziyi2021}@ia.ac.cn
{shijing2014, shuang.xu, xubo}@ia.ac.cn

ABSTRACT

Medical Slot Filling (MSF) task aims to convert medical queries into structured information, playing an essential role in diagnosis dialogue systems. However, the lack of sufficient term semantics learning makes existing approaches hard to capture semantically identical but colloquial expressions of terms in medical conversations. In this work, we formalize MSF into a matching problem and propose a Term Semantics Pre-trained Matching Network (TSPMN) that takes both terms and queries as input to model their semantic interaction. To learn term semantics better, we further design two self-supervised objectives, including Contrastive Term Discrimination (CTD) and Matching-based Mask Term Modeling (MMTM). CTD determines whether it is the masked term in the dialogue for each given term, while MMTM directly predicts the masked ones. Experimental results on two Chinese benchmarks show that TSPMN outperforms strong baselines, especially in few-shot settings[1].

***Index Terms*—** Medical Dialogues, Spoken Language Understanding, Slot Filling, Low Resource, Pre-training

1. INTRODUCTION

Medical Slot Filling (MSF), which intends to automatically convert medical queries into structured information by detecting medical terms, has recently received increased attention [1–3]. It plays a vital role in diagnosis dialogue systems [4,5] . Different from conventional slot filling tasks in NLP that label the explicit words in a given utterance and extract the structured information (a.k.a slot-value pairs) based on the labeled words [6–9], MSF exists the non-alignment issue between a patient query and corresponding provided medical term slots [2, 3, 10–12]. To be specific, colloquial expressions of terms in patient queries vary from formal expressions. As shown in Tabel 1, the slot-value `Symptom:Bellyache` does not explicitly appear in any specific spans but is mentioned implicitly in the query. Therefore, MSF requires a deeper understanding of term semantics with medical knowledge. Besides, medical data is more dependent on expert annotation, and the annotation is expensive to obtain in practice, which makes annotation data particularly insufficient.

Patient Query
My stomach feels bad these days, pain in the area above the navel, poop twice a day, belly bulge, shapeless, take cefixime currently, what happens? 我这几天肚子感觉难受，肚脐眼上面的位置疼痛，一天大便两次，肚子胀，不成型，目前在吃头孢克肟，这是怎么回事呢?
Slot-values Pairs Label
Symptom:Bellyache (腹痛) Symptom:Diarrhea (稀便) Symptom:Abdominal Distension (腹胀) Medicine:Cefixime (头孢克肟)

Table 1. An example of a patient query and the label that consists of slot-value pairs (e.g, Symptom:Diarrhea).

Recent works [2, 3, 13] have been proposed to address the above problems, which can be generally grouped into two categories: multi-label classification and generative methods. The first category of methods [2, 3] regards pre-defined slot-value pairs as different classes. They can utilize unlabeled patient queries with doctor responses to produce pseudo labels as weak supervision. However, it requires that unlabeled data map to the limited pre-defined terms, making it challenging to exploit a larger unlabeled medical conversation corpus. The second category of methods [13] commonly models MSF as a response generation task through a dialog prompt. In this way, MSF benefits from dialogue-style pre-training utilizing the large unlabeled medical dialogue corpus. However, the divergence between MSF and the response generation task inevitably undermines the performance. Besides, as the model generates terms in sequential order, the errors accumulated from previous steps will be propagated to the later steps [14].

Unlike these approaches, we propose a Term Semantics Pre-trained Matching Network (TSPMN) that takes both

This work is supported by the Key Programs of Chinese Academy of Sciences (No.ZDBS-SSW-JSC006-2), and the National Natural Science Foundation of China (No.62206294).

[1]Our codes can be found at https://github.com/FlyingCat-fa/TSPMN.

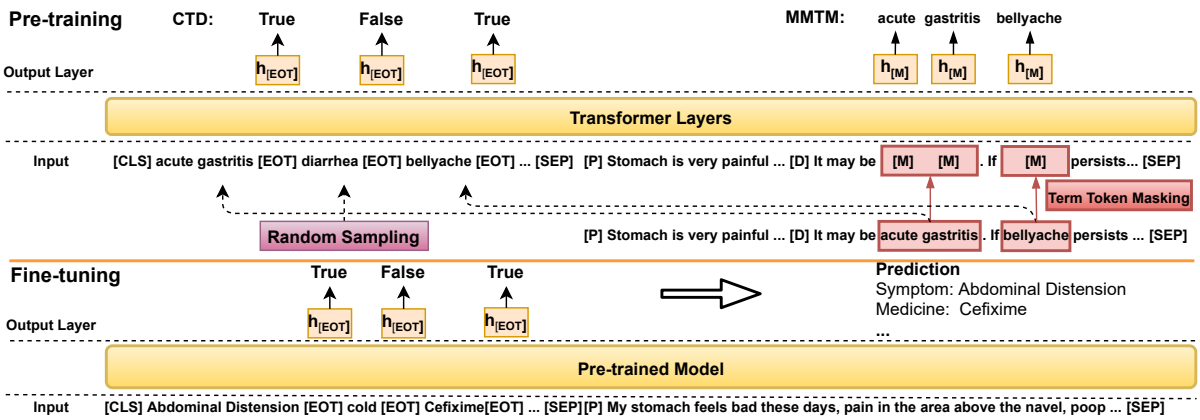

Fig. 1. Illustration of Term Semantics Pre-trained Matching Network (TSPMN). The input consists of a term sequence and a medical dialogue/patient query. `[EOT]` is the separator for the previous term. `[P]` and `[D]` represents patient and doctor, respectively. TSPMN first learns term semantics through our self-supervised tasks. Then the pre-trained TSPMN is fine-tuned to match candidate terms and the patient query for MSF. Note that all examples are translated from Chinese, and the example of fine-tuning is from Table 1.

terms and queries as input. Therefore, the model only needs to learn how to match between the queries and given terms rather than map the queries into limited pre-defined labels, which reduces the data restrictions. Moreover, two self-supervised tasks are proposed for TSPMN to learn term semantics better, including Contrastive Term Discrimination (CTD) and Matching-based Mask Term Modeling (MMTM). CTD is a matching task close to MSF, while MMTM is an adaptive Mask language Modeling (MLM) task to predict masked term tokens better by matching with golden tokens. In this way, TSPMN can not only use large-scale medical dialogue corpora for pre-training, but also reduce the divergence between the pre-training and fine-tuning phases. Experimental results on two Chinese benchmarks show that TSPMN outperforms strong baselines, especially in few-shot settings .

2. PROBLEM STATEMENT

Given a patient query q containing colloquial expressions, Medical Slot Filling (MSF) task aims at transforming the query q into the grounded formal representation with discrete logical forms `(slot: value)`. The candidates of `slot` and `value` are pre-defined according to Medical Knowledge Graphs, where the value is a medical term, and the slot represents the category of the term (e.g., `(Symptom: Bellyache)`). We formulate MSF as a matching problem, in which we match each term candidate to the patient query q to determine whether the candidate appears in q.

3. APPROACH

This section presents how Term Semantics Pre-trained Matching Network (TSPMN) models Medical Slot Filling (MSF) by matching terms and patient queries. Then two term semantics pre-training tasks for TSPMN will be introduced.

3.1. Matching for MSF

For efficient matching while considering the length limitations of the model input, we first construct multiple term sequences by concatenating the terms in the term set T. Each term sequence is in the following form:

$$S_t = (T_1, \texttt{[EOT]}, \ldots, T_i, \texttt{[EOT]}, \ldots, T_n, \texttt{[EOT]}), \quad (1)$$

where n and `[EOT]` are the term number and the separator following each term, and T_i represents the tokens of the i-th term. Given a patient query, we concatenate each term sequence with the query as a whole sequence x, and encode x with a pre-trained language model $\mathcal{H}$ such as BERT [15] to capture semantic information adequately:

$$(\mathbf{h}_{\texttt{[CLS]}}, ..., \mathbf{h}_{\texttt{[EOT]}}, ..., \mathbf{h}_{\texttt{[SEP]}}) = \mathcal{H}(x). \quad (2)$$

We take each hidden state $\mathbf{h}_{\texttt{[EOT]}}$ as the hidden state of the term before `[EOT]`. Then the probability about whether T_i appears in the query is predicted as follows:

$$p_i(x;\theta) = \text{Softmax}(\text{FFN}(\mathbf{h}_{T_i})), \quad (3)$$

where $\mathbf{h}_{T_i}$ means the hidden state of term i and θ is the model parameter. We map $\mathbf{h}_{T_i}$ to the scores of `True` and `False` independently through FFN, which means that term i is mentioned in the query (`True`) or not (`False`), respectively. $p_i(x;\theta) \in \mathbb{R}^2$ represents the scores normalized by softmax function. If the normalized score of `True` is bigger than the `False`, we choose the term i to fill the corresponding slot.

The MSF loss function is defined as:

$$\mathcal{L}_{MSF} = -\sum_{i=1}^{n}\sum_{k} y_{i,k} \log p_{i,k}, \quad k \in \{0,1\}, \quad (4)$$

where n and k denote the number of terms and the index of `True` or `False`, and $y_i \in \{[1,0],[0,1]\}$ is the label indicating whether T_i appears in the patient query.

3.2. Term Sementics Pre-training

Pre-trained language models (PrLMs) show excellent performance in many tasks [15, 16]. There are also numerous PrLMs for dialogue representation [17–21] or medical domain adaptation [22–26]. Inspired by these works, we focus on spoken understanding in medical dialogues and propose two self-supervised tasks to better model term semantics, which can not only use large-scale medical dialogue corpora, but also narrow the gap between those tasks and MSF as much as possible. The details are described as follows.

3.2.1. Matching for Pre-training

We use three public unlabeled medical dialogue datasets MedDialog [27], KaMed [28], and ReMeDi-large [29], as pre-training corpora, which contain over 3.5M dialogues in more than 100 medical departments. The public sougoupinyin medical dictionary[2] and the medical dictionary THUOCL [30] are merged as a large medical terminology T_{large}. The terms in the knowledge base of the pre-training corpora are also added to T_{large}. Based on T_{large}, we retrieve the terms in each medical dialogue by string matching and construct dialogue-terms pairs for pre-training. Similar to the matching for MSF in section 3.1, we construct multiple term sequences by concatenating the terms and the dialogue as the input. Specifically, each term sequence consists of the sampled positive terms from the current dialogue and negative terms that are not in the current dialogue. The difference from MSF is that we only mask the sampled positive terms. In this way, the model can learn term semantics from dialogue contexts rather than just focus on string matching. We denote the input as x_{mask} and encode it in the same way as equation 2:

$$(\mathbf{h}_{\texttt{[CLS]}}, ..., \mathbf{h}_{\texttt{[EOT]}}, ..., \mathbf{h}_{\texttt{[M]}}, ..., \mathbf{h}_{\texttt{[SEP]}}) = \mathcal{H}(x_{mask}), \quad (5)$$

where M means a masked token of terms. The hidden states are used for our two self-supervised tasks. The two tasks share the same input and encourage the pre-trained model to capture different aspects of semantics through multi-task learning. We define the total pre-training loss as the summation of two aforementioned losses:

$$\mathcal{L}_{pretrain} = \lambda \mathcal{L}_{CTD} + (1-\lambda)\mathcal{L}_{MMTM}, \quad (6)$$

where λ is a tunable weight used to adjust the contribution of different losses, $\mathcal{L}_{CTD}$ and $\mathcal{L}_{MMTM}$ are the losses of CTD and MMTM, respectively. The details of those two tasks are introduced in the following subsection.

[2] `https://pinyin.sogou.com/dict/detail/index/15125`, updated to October 13, 2017.

3.2.2. Self-supervised Tasks

Contrastive Term Discrimination. For each term in the term sequence, CTD aims to determine whether it belongs to the current dialogue. Similar to MSF in section 3.1, we use hidden states of `[EOT]` ($\mathbf{h}_{T_i} = \mathbf{h}_{\texttt{[EOT]}_i}$) to represent the front term after matching with the patient query. The operation is the same as equations 3 and 4.

Matching-based Mask Term Modeling. This task is motivated by masked language modeling (MLM) [15] with two improvements to match MSF: 1) MMTM only masks medical terms, 2) The masked always appears in the T_{large}. Therefore, the model can not only learn semantics from the dialogue context but also more fully model the semantic interactions of the term and the dialogue. $\mathbf{h}_{\texttt{[M]}}$ is used to predict the mask. We compute the same cross-entropy loss as MLM.

4. EXPERIMENTS

4.1. Datasets and Evaluation Metrics

Dataset	Train	Dev	Test	Slot	Value(Term)
MSL	1152	500	1000	1	29
MedDG	50965	6956	3645	4	155

Table 2. Data statistics of MSL and MedDG datasets.

We evaluate our method on two Chinese medical datasets: MSL [2] and MedDG [1]. MedDG was initially constructed for the medical dialogue system and labeled with the medical slots, which can be used for Medical Slot Filling (MSF). The statistics of the datasets are shown in Table 2. For evaluation, we follow the MSL guidance [2] for all individual metrics: **Precision**, **Recall**, **Micro F1**, **Macro F1** and **Accuracy**.

4.2. Implementation Details

We initialize our model with Chinese BERT-base [15]. During the pre-training phase, the batch size is 48, and 1-bit Adam [31] is used as the optimizer. We set the learning rate and pre-training epoch as 3×10^{-5} and 5, respectively. And λ is set to 0.9. During the fine-tuning phase, AdamW [32] is used as our optimizer with an initial learning rate of 1×10^{-5}. The batch size is 8 for MedDG and 32 for MSL. We set the term number n of each term sequence to 20, 15 and 20 for pre-training, fine-tuning on MSL and MedDG, respectively.

4.3. Main Results

Full Training Evaluation. From Table 3 we can see that our model achieves new state-of-the-art results. The improvements of all metrics over baselines are statistically significant where $p < 0.05$ from significance testing. Compared with the classification method BERT+TST and the generative-based method PromptGen, which are enhanced or pre-trained

Model	MSL					MedDG				
	P	R	mi-F1	ma-F1	Acc	P	R	mi-F1	ma-F1	Acc
DRNN [2]†	83.43	67.85	74.83	65.17	52.5	96.55	97.34	96.95	82.8	73.39
DRNN+A [2]†	82.11	70.86	76.07	67.42	51.9	98.53	96.69	97.6	83.62	75.25
DRNN+A+WS [2]‡	82.94	79.44	81.15	76.95	58.3	-	-	-	-	-
TextCNN-Raw [3]†	90.37	64.31	75.14	64.28	51.6	97.5	95.65	96.57	80.64	72.57
BERT-Raw [3]†	89.78	88.63	89.2	87.03	70.9	99.3	99.38	99.34	84.82	74.15
BERT+TST [3]‡	90.95	90.81	90.88	89.28	72.9	-	-	-	-	-
PromptGen [13]§	89.11	87.57	88.34	87.75	79.6	-	-	-	-	-
TSPMN-MedBERT	90.91	**91.11**	91.01	90.45	81.4	99.38	**99.5**	99.44	86.51	98.44
TSPMN	**92.33**	90.66	**91.49**	**90.62**	**83.4**	**99.61**	99.43	**99.52**	**86.55**	**98.77**
w/o MMTM	91.14	90.66	90.90	89.74	82.80	99.45	99.43	99.44	86.34	98.44
w/o Pre-train	85.76	88.40	87.06	86.36	74.50	99.21	99.28	99.25	86.00	97.86

Table 3. Full training evaluation on MSL and MedDG datasets. †: we cite the results of these models on MSL from the original papers [2,3], and obtain the results on MedDG based on their released codes. ‡: the models require homologous unlabeled data. §: as the authors did not release their code, we cite the results of PromptGen on MSL from the original paper [13].

Model	1-shot			2-shot			5-shot		
	mi-F1	ma-F1	Acc	mi-F1	ma-F1	Acc	mi-F1	ma-F1	Acc
DRNN+A	20.66±3.32	11.63±3.15	7.16±1.88	33.69±3.38	27.96±4.89	11.74±1.15	52.83±4.84	50.24±5.3	23.74±3.56
DRNN+A+WS	70.55±1.62	64.18±4.06	42.08±1.51	71.14±1.09	64.84±0.67	42.96±2.06	75.11±0.87	70.56±1.36	46.74±0.87
BERT-Raw	14.74±4.44	6.93±3.78	4.56±1.99	41.76±13.71	31.25±17.21	17.84±6.29	73.62±1.8	68.3±2.93	46.8±2.14
BERT+TST	71.68±3.19	62.5±5.68	49.76±2.56	72.76±1.06	62.55±3.51	51.46±1.9	77.55±0.68	71.08±1.16	55.66±1.63
TSPMN-MedBERT	78.19±0.86	77.81±0.25	53.32±1.6	79.34±0.93	79.1±1.43	55.76±2.14	83.34±1.04	83.27±1.13	64.08±1.96
TSPMN	**78.52±2.65**	**79.32±1.19**	**55.78±4.61**	**81.87±1.21**	**81.93±1.56**	60.5±2.18	**84.74±0.83**	83.96±0.59	**67.28±1.79**
w/o MMTM	77.82±2.32	78.24±1.47	55.68±3.63	81.37±1.2	81.49±0.59	**61.2±1.86**	84.22±0.83	**84±0.91**	66.08±1.84
w/o Pre-train	75.12±1.3	74.69±0.77	48.78±2.45	77.59±0.45	76.87±0.48	53.86±0.98	81.61±1.64	81.64±1.46	60.78±3.13

Table 4. Few-shot evaluation on MSL. The means and standard deviations over five runs are reported.

on medical corpora, our TSPMN shows consistent improvements, which we attribute to TSPMN utilizing both large-scale unlabeled medical dialogue corpora and narrowing the discrepancy between the pre-training and fine-tuning phases. For further analysis, we remove the MMTM objective and then remove both MMTM and CTD, denoted as TSPMN w/o MMTM and TSPMN w/o Pre-train. We further replace our self-supervised objectives with the original BERT objectives, denote it as TSPMN-MedBERT. From the perspective of pre-training corpora, TSPMN and TSPMN-MedBERT perform better than TSPMN w/o Pre-train, illustrating the importance of continuous pre-training on large-scale medical dialogue corpora. From the perspective of pre-training tasks, the comparison of TSPMN and TSPMN-MedBERT indicates that the closer pre-training tasks get to the target task, the more performance gain can be achieved. The ablation experiments also verify the effectiveness of CTD and MMTM.

Few-shot Evaluation. We further evaluate our method in more challenging few-shot settings. In the k-shot setting, we select k training examples from the original training set for each term to form the training dataset. The initial validation and test data are still used for the few-shot evaluation. As shown in Table 4, TSPMN achieves more performance gains than baselines in few-shot settings. Further, we find that TSPMN outperforms baselines on most measures even without pre-training, which validates the excellence of the novel matching paradigm of TSPMN in low resource scenarios. As shown in Table 3 and Table 4, compared with TSPMN-MedBERT, TSPMN achieves more relative improvement in few-shot settings than full training settings, which indicates that the smaller discrepancy between pre-training and fine-tuning phases is more significant in low resource scenarios.

5. CONCLUSION

The variation of terminology complexity between patients and formal providers requires a deeper and richer semantics understanding, which has been a headache in Medical Slot Filling (MSF) task. To learn term semantics thoroughly, this paper proposes Term Semantics Pre-trained Matching Network (TSPMN) with two self-supervised objectives, including Contrastive Term Discrimination (CTD) and Matching-based Mask Term Modeling (MMTM). We reveal the excellence of TSPMN and the proposed training objectives through detailed experiments. The limitation of this paper is that the value of medical slots only consider as terms, which ignores the possible status corresponding to terms in more complicated scenarios, and we leave it to our future work.

6. REFERENCES

[1] Wenge Liu et al., "Meddg: A large-scale medical consultation dataset for building medical dialogue system," *arXiv preprint*, 2020.

[2] Xiaoming Shi et al., "Understanding medical conversations with scattered keyword attention and weak supervision from responses," in *Proc. of AAAI*, 2020.

[3] Xiaoming Shi et al., "Understanding patient query with weak supervision from doctor response," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[4] Zhongyu Wei et al., "Task-oriented dialogue system for automatic diagnosis," in *Proc. of ACL*, 2018.

[5] Mina Valizadeh and Natalie Parde, "The ai doctor is in: A survey of task-oriented dialogue systems for healthcare applications," in *Proc. of ACL*, 2022.

[6] Grégoire Mesnil et al., "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE ACM Trans. Audio Speech Lang. Process.*, 2015.

[7] Ngoc Thang Vu et al., "Bi-directional recurrent neural network with ranking loss for spoken language understanding," in *Proc. of ICASSP*, 2016.

[8] Qian Chen et al., "BERT for joint intent classification and slot filling," *CoRR*, 2019.

[9] Libo Qin et al., "A co-interactive transformer for joint slot filling and intent detection," in *Proc. of ICASSP*, 2021.

[10] Deyu Zhou and Yulan He, "Learning conditional random fields from unaligned data for natural language understanding," in *Proc. of ECIR*, 2011.

[11] Lina Maria Rojas-Barahona et al., "Exploiting sentence and context representations in deep neural models for spoken language understanding," in *Proc. of COLING*, 2016.

[12] Lin Zhao and Zhe Feng, "Improving slot filling in spoken language understanding with joint pointer and attention," in *Proc. of ACL*, 2018.

[13] Jun Liu et al., "Prompt-based generative approach towards multi-hierarchical medical dialogue state tracking," *CoRR*, 2022.

[14] Xiujun Li et al., "Investigation of language understanding impact for reinforcement learning based dialogue systems," *CoRR*, 2017.

[15] Jacob Devlin et al., "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc. of NAACL*, 2019.

[16] Linghui Meng et al., "Offline pre-trained multi-agent decision transformer: One big sequence model conquers all starcraftii tasks," *arXiv preprint*, 2021.

[17] Shikib Mehri et al., "Pretraining methods for dialog context representation learning," in *Proc. of ACL*, 2019.

[18] Zhuosheng Zhang et al., "Structural pre-training for dialogue comprehension," in *Proc. of ACL*, 2021.

[19] Yi Xu and Hai Zhao, "Dialogue-oriented pre-training," in *Proc. of ACL Findings*, 2021.

[20] Zekang Li et al., "Conversations are not flat: Modeling the dynamic information flow across dialogue utterances," in *Proc. of ACL*, 2021.

[21] Xueliang Zhao et al., "Towards efficient dialogue pretraining with transferable and interpretable latent structure," in *Proc. of EMNLP*, 2022.

[22] Ningyu Zhang et al., "Conceptualized representation learning for chinese biomedical text mining," *arXiv preprint*, 2020.

[23] Benyou Wang et al., "Pre-trained language models in biomedical domain: A systematic survey," *arXiv preprint*, 2021.

[24] Taolin Zhang et al., "Smedbert: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining," in *Proc. of ACL*, 2021.

[25] Renqian Luo et al., "Biogpt: generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, 2022.

[26] Hongyi Yuan et al., "Biobart: Pretraining and evaluation of a biomedical generative language model," *BioNLP @ ACL*, 2022.

[27] Guangtao Zeng et al., "Meddialog: Large-scale medical dialogue datasets," in *Proc. of EMNLP*, 2020.

[28] Dongdong Li et al., "Semi-supervised variational reasoning for medical dialogue generation," in *Proc. of SIGIR*, 2021.

[29] Guojun Yan et al., "Remedi: Resources for multi-domain, multi-service, medical dialogues," in *Proc. of SIGIR*, 2022.

[30] Shiyi Han et al., "Thuocl: Tsinghua open chinese lexicon," *Tsinghua University*, 2016.

[31] Hanlin Tang et al., "1-bit adam: Communication efficient large-scale training with adam's convergence speed," in *Proc. of ICML*, 2021.

[32] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *Proc. of ICLR*, 2019.