Responses to the reviews of ACL rolling review paper #162: “Latent Group Dropout for Multilingual and Multidomain Machine Translation”

We thank the reviewers for sharing their suggestions which could improve the clarity of our paper and strengthen our claim over the efficacy of our proposed method

Reviewer iMf4

Paper Summary:

This paper considers multidomain and multilingual NMT as multi-task learning and proposed a method to automatically determine task-specific and task-agnostic parameters in the neural networks without explicitly defining or adding any additional layers. learning task dependent sub-networks. They formulate such problem as latent dropout mask learning in which the parameters in each transformer layer are divided into multiple groups and learn to select the k active groups for each task so that the multitask loss is minimized. Experiments in multidomain NMT and multilingual NMT show that their proposed methods achieve better average BLEU scores than the vanilla transformer and adapter methods.

Summary of Strengths:

- The problem setting is interesting and well-formulated.

Summary of Weaknesses:

- Missing citation and baseline: Learning task-specific and task-agnostic parameters in multitask learning for NMT has been proposed in Lin et al.2021. In this work, they learn a mask for each neuron instead of group of neurons. You should compare the proposed LaMGD method with this baseline.
- Lack of ablation study on the dropout group selection. What if the groups are selected randomly or using some predefined heuristics, e.g. first 2 groups are shared across tasks, some other groups for language family, or a single language.
- You should have an analysis on the effect of k which potentially controls the level of sharing between tasks.
- In table 5, although the average BLEU scores of LaMGD is higher, most of the results are not statistically significant.

Comments, Suggestions And Typos:

I Table 7 can be improved in presentation by using heatmap.

Overall Assessment: 2

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Best Paper: No

Replicability: 2 = They would be hard pressed to reproduce the results: The contribution depends on data that are simply not available outside the author's institution or consortium and/or not enough details are provided.

Datasets: 1 = No usable datasets submitted.

Software: 2 = Documentary: The new software will be useful to study or replicate the reported research, although for other purposes it may have limited interest or limited usability. (Still a positive rating)

Author Identity Guess: 1 = I do not have even an educated guess about author identity.

Rebuttal

We thank the reviewer for his suggestion on a citation that we accidentally missed. For the sake of the clarification of our approach, we would like to claim several misunderstandings in the summary made by Reviewer iMf4:

- Our approach does not aim to find task-specific and task-agnostic parameters in the neural network. The objective of our algorithm is to optimize the assignment of a sub-group of "nodes", which are the dimensions of the output of each intermediate layer, to each task.
- Our approach does not mask the sub-parts of each weight matrix but masks a fixed number of sub-groups of elements of the output of each intermediate layer.

We would like to answer the concerns of the reviews as follows.

- We admit the weakness in our choice of baseline according to the reviewer. However, our claim is that embedding the similarity between tasks in the construction of the multi-task network is beneficial, and the fine-tuning-based approaches ignore these similarities. We prove our claim by proposing a method that automatically optimizes the structure of a multi-task network by learning from data, i.e., implicitly learning the similarity between tasks and embedding it to the organization of the network. Besides, the Adapters are a robust method for multi-task learning, and a comparison with this baseline strongly supports our claim over the efficiency of our proposed method.
- We are happy to provide more comparisons with heuristic strategies of the dropout group selection in the final version. The flaws of these heuristic strategies motivate us to develop this technique.
- We have finished the analysis on the effect of k on the performance and will provide them in the final version.
- The multilingual experiments use two groups of 8 languages, which are all low-resourced; this might lower the significance of the improvement as observed in the

paper of Lin et al. 21 (only 0.7 for low). Our method significantly improves the minor languages among these languages.

We hope the reviewer could reconsider the solid mathematical foundation of our method, the improvement with high margin in multi-domain, the reduced computational cost in the inference compared to the Adapters, and the novelty in the use of node-masking instead of weight-masking as in previous work such as in Lin et al.

Reviewer 681g

Paper Summary:

This paper proposes a routing strategy for multitask machine translation. For a given language or domain, only a fixed amount of nodes is activated at each layer and the model learns which nodes should be activated. They experiment on both multilingual and multidomain settings, and the experimental results demonstrate that they can achieve comparable and sometimes marginally better performance than adapter-based methods with fewer parameters.

Summary of Strengths:

- The general idea is intuitive and makes sense, and the paper proposes a technically sound way of implementing it.
- The empirical results are good as they can achieve comparable performance with an adapter-based method with fewer parameters.
- The paper is well-organized.

Summary Of Weaknesses:

- There is little analysis of their methods. Sensitivity analysis of their hyper-parameters (n_d, k, τ), some qualitative analysis are required.
- Requiring partitioning nodes into groups and a fixed number of nodes to be activated at each layer makes their method less flexible.
- More baselines should be included (e.g. Li et al., (2020), Gong et al., (2021a, b) as mentioned in the paper).

Comments, Suggestions And Typos:

It'd be good to also include a mixture-of-experts baseline.

Overall Assessment: 3 = Good: This paper is of interest to the *ACL audience and could be published, but might not be appropriate for a top-tier publication venue. It would likely be a strong paper in a suitable workshop.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Best Paper: No

Replicability: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Author Identity Guess: 1 = I do not have even an educated guess about author identity.

Rebuttal

We thank the reviewer for appreciating our mathematical work to support our implementation and the suggestions for a better understanding of our proposed approach. We would like to answer the reviewer's concerns as follows.

- We did admit our lack of analysis on the hyperparameters in the conclusion of our paper. We have collected more data for better analyzing the effect of hyperparameters on the performance of our method since the submission. We will provide these results in our final version.
- We agree with the reviewer. However, we could reduce the group size to 1 node only and use hierarchical modeling to optimize the number of nodes to be activated. These changes would improve the method's flexibility in exchange for computational cost in training.
- We are aware of these works, but their implementations are not explicit and require time to be well-calibrated.

We agree that a mixture of experts like Gshard is an excellent candidate for the baseline. We have been working on their implementations but lack the computational resource for training.

Reviewer 3koi

Paper Summary:

This paper proposes a latent-variable model based on a variational probabilistic modeling framework for conditional compute in multi-domain and multilingual machine translation. They introduce latent variables which are learned during model training. These latent variables decide which sub-network of the model is selected for each task / language pair. The results on multi-domain machine translation show promise compared to vanilla Transformer and Adapter baselines. The results on multi-lingual machine translation are somewhat mixed with modest improvements in 3 settings and worse accuracy in 1 setting. The paper includes some interesting analysis of the models as well. On the other hand, some important ablations/hyper-parameter tuning experiments are missing and it is unclear how statistically significant the results on multilingual translation are.

Summary of Strengths:

- This work proposes a technique to learn which subnetworks to use for which task / language pair based on learning from the data.
- The proposed probabilistic framework is mathematically sound and explained clearly and completely.
- There are experiments compared with decent baselines and multiple benchmarks and there has been a good attempt to evaluate the new technique well.

- The analysis showing that accuracy of Adapter baselines declines on a domain when that domain is divided into two pseudo-domains, where as the accuracy on this proposed framework remains the same motivates the reason of using the proposed technique very well
- A nice advantage of this approach is that it doesn't need additional parameters to the added to the baseline Transformer model. There are some additional hyper-parameters to tune correctly per setup though.
- There is also analysis showing whether the learned subnetworks for related domains/languages overlap significantly.

Summary of Weaknesses:

- It is unclear how the latent variables that decide the subnetworks to be used per task/domain are initialized at the start of model training.
- While the authors mention leaving this to future work, the study of how many groups to divide each layer into and how many groups to dropout per layer is missing.
- The ablation about the effect of tuning temperature parameter τ is missing.
- It is unclear whether the baselines have been allowed to train all the way till convergence. The multi-domain baseline model is trained for 20K updates, while the proposed model is trained for 300K updates. This looks weird.
- Similarly, but to a lesser extent, the multilingual MT baselines are trained for 40K/(40K+5K) updates. The proposed model is trained for 50K updates. So it is unclear if the baselines have been given enough chance to convergence well.
- It is unclear how statistically significant the improvements or decline in accuracy in the 4 setups for multilingual MT are.

Comments, Suggestions and Typos:

[nice to have] comparison to Adapter baseline with the same number of parameters and approximate FLOPs per update as the Transformer baseline and the proposed approach would be useful. While the proposed approach and results are quite interesting, there is some need to qualify some of the claims in the Introduction and Conclusion such as "very little extra computational cost" (it looks like training cost increases by 33%); "significant gains of this method with respect to baselines" (the gains are great in the multi-domain benchmark, but modest in the multilingual translation benchmark).

Overall Assessment: 3 = Good: This paper is of interest to the *ACL audience and could be published, but might not be appropriate for a top-tier publication venue. It would likely be a strong paper in a suitable workshop.

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Best Paper: No

Replicability: 2 = They would be hard pressed to reproduce the results: The contribution depends on data that are simply not available outside the author's institution or consortium and/or not enough details are provided.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Author Identity Guess: 1 = I do not have even an educated guess about author identity.

Rebuttal

We thank the reviewer for appreciating our mathematical modeling and the suggestions for a better understanding of our proposed technique. We would like to answer the concerns of the reviewer as follows.

- We initialize the probabilistic distribution of the latent variables as uniform distribution.
- We will provide more data on how k affect the performance of our approach in the final version.
- As reported from many works using the Gumbel reparameterization trick, the effect of τ is relatively insignificant. Therefore, we focused more on selecting k.
- Because we would like to have the same total amount of training iterations between Adapter-based and LaMGD, we extend the training of LaMGD to 300K. The convergence of the standard Transformer is before 200K as its validation curve became flat near the 200K-th iteration.
- The previous reason is applied for the reviewer's concern over multilingual experiments.
- As we pointed out, the significant gains are observed in several very low-resourced languages, not globally. This might be explained by the fact that low-resourced languages benefit from the sharing nodes with their related well-resourced languages while less suffering from the inference of unrelated languages.

Our claim on the advantage in computational cost repose on the inference stage while there are no extra parameters added to the standard Transformer compared to Adapter-based methods. However, the training time will cost more than usual because of the joint optimization of the network's organization and the parameters.