View Reviews

Paper ID 5582

Paper Title L^2-GCN: Layer-Wise and Learned Efficient Training of Graph Convolutional Networks

Reviewer #2

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.
The authors proposed LWGCN, disentangling feature aggregation and feature transformation during training to reduce time and memory constraints. The effect of this design is analyzed theoretically. Also, the authors proposed RL-based controller to automatically adjust the training epochs per layer in LWGCN. With this design, the authors claim that the proposed method scales better with improved performance.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.
- A novel idea of disentangling feature aggregation and feature transformation
- Another interesting idea of using RL to control the number of epochs
- Scalable solution with theoretical analysis

3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel).
- Scores in experiment do not match with original papers of baselines, losing trustworthiness of the experiment.

4. [Overall rating] Paper rating (pre-rebuttal)
Weak accept

5. [Justification of rating] Please explain how the strengths and weaknesses aforementioned were weighed in for the rating. Please also mention what you expect to see from the rebuttal that may change your rating.
1. It is indeed a brilliant idea that we can separate feature aggregation and feature transformation, and train layer-wise to scale up. Section 3 describes this clearly.

2. It is great to see that the authors have trained and evaluated the proposed method on large-scale datasets such as Amazon-3M.

3. One concern with the experiment is that we are not sure what is the accuracy metric the authors have used. It simply says 'accuracy', so I guess it is classification accuracy. Now, let's compare the figures for baselines in Table 4 with their original papers. In GraphSAGE [8], PPI was 50.2 (unsup-F1) and 61.2 (sup-F1), Reddit was 90.8 and 95.4, respectively, while the reported numbers here are 68.8 and 93.4. Same thing for FastGCN [2]: 85.5 vs. 85.0 (Cora), 87.4 vs. 88.0 (PubMed), and 92.6 vs. 93.7 (Reddit). Why these scores do not match? I guess it might be because of the mismatch of metrics, F1 in these papers vs. classification accuracy in this submission. Then, why don't you use the same metric to compare against the original paper? Unlike the first two baselines, VRGCN [3] indeed provided classification accuracy in their original paper, but scores do not match again: 85.4 vs. 82.0 (Cora), 86.4 vs. 78.7 (PubMed), and 98.6 vs. 97.8 (PPI). Please include justification for this mismatch, with precise definition of the metrics used.

4. There's another interesting recent work on extending GCN from last year's UAI. We encourage the authors to compare against this as well:http://auai.org/uai2019/proceedings/papers/310.pdf

It will be more interesting to apply the proposed idea (layer-wise training) into this extension as well.

5. (minor) In section 2.1: When we define a graph, it specifies two sets (V and E) only. F is technically not part of the graph, and A is purely defined by E.

11. Final rating
Weak accept

12. Explanation of final rating. Describe the rationale for your final rating, including notes based on the rebuttal, discussion, and other reviews.
Thanks for the rebuttal. I keep the weak accept rating as original.

Reviewer #4

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.
the paper propose to speed up the training of GCN by decouping the training process to two step: feature aggregation and feature transformer. And employ a rnn network to automatically train the GCN layer by layer. Theoretical analysis prove that layer-wise training is able to achieve the same capacity as conventional training.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.
1. The motivation for the paper is useful. Proposed method is an important aspect of deep learning and can be of high practical impact.
2. Rationale analysis provided some insights about how and why would this approach work.
3. Some experiments are reasonable enough to make the approach credible

3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel).
Some mathematical symbols are confusing and should be presented better:
1. what is N(i) in eq8. ?
2. At L77 in supplementary, what's difference between x^(0)_{i,G_1} , x^(0)_{i,G1,Con} and x^(0)_{i,G1,Lay}
3. lacking of page number

4. [Overall rating] Paper rating (pre-rebuttal)
Weak accept

5. [Justification of rating] Please explain how the strengths and weaknesses aforementioned were weighed in for the rating. Please also mention what you expect to see from the rebuttal that may change your rating.
The idea is interesting and the results look good. I wish the author could handle the mathematical problem and make it easy to understand.

11. Final rating
Weak accept

12. Explanation of final rating. Describe the rationale for your final rating, including notes based on the rebuttal, discussion, and other reviews.
The authors have addressed my concern. As for the concerns raised by other reviewers, the author should address them in the final version.

Reviewer #5

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.

This paper proposes an efficient layer-wise training framework for GCN. It reduces time and memory complexities by disentangling feature aggregation and feature transformation during training. Besides, it learns a controller for each layer, which can automatically adjust the training epochs per layer in LWGCN.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.

1. The paper proposed a training framework that disentangles feature aggregation and feature transformation during training. The algorithm is faster than SOTA by at least an order of magnitude.

2. It learns a controller for each layer to make the early stopping decision, which can further cut the training time in half with tiny performance loss.

3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel).

I have many concerns about this paper.

1. The time reported in Table 4 is unfair, which excludes search time. The evidence is that training for 80+80 and 75+75 epochs cost 0.45 seconds and 0.38 seconds, respectively. Supposing during hyper-parameter optimization, you have searched 10000 rounds of hyper-parameter, the time you used maybe 3,800 seconds, which is far time-consuming that the competitors. Therefore, the proposed method is inefficient.

2. The proposed controller learning is exactly an AutoML technique and an algorithm of hyper-parameter optimization (HPO). Therefore, discussion and comparison with existing AutoML and HPO are desirable in the related works.

3. It is claimed as a contribution that the proposed method does not compromise in its graph-representative power in theory. However, this theory is invalid because there does not exist an injective mapping for each layer for DNN. Therefore, the theoretical analysis in Section 3.2 and the appendix may be useless.

4. I doubt the effectiveness of the proposed controller learning. Actually, recent works have proved the ineffectiveness of existing AutoML, suggesting the many AutoML techniques are not better than random selection in the search space. To prove the effectiveness of hyper-parameter search in method, a curve showing the search strategy obtains a better hyper-parameter than the previous hyper-parameter is needed.

5. The value of the horizontal axis of Figure 1 should be re-order from small to large. Similarly, the value of the vertical axis should be re-order from small to large.

6. In general, the target/optimization of end-to-end training is different from that of the greedy layer-wise training. Imposing a loss on each of the low-level layer may bring unpredictable results. A theoretical explanation should be provided to analyze how the proposed method closes the gap.

7. Reducing training time is the main focus of the paper. However, due to the model size and the data volume, the training time of all the competing methods are not consuming. Actually, it appears to us that training a GCN does not suffers from the high computation. For example, training on Cora only costs 18 seconds for the competing method, which is very efficient compared to the time of training on ImageNet (1 week) in computer vision and the time of training a BERT (1000+ GPU days) in NLP. Therefore, it seems the proposed method is useless. To show the advance

of the proposed method, training with a large model or on a large scale dataset is desirable. Recently, training a very deep GCN has been enabled. I suggest the authors examine the effectiveness of the proposed method with a very deep model (e.g., 100 layers).

8. The proposed method seems to be unscalable. Saving the internal representation of the graph into the hard disk has many problems. First, it would cost IO time, especially when the dataset is large-scale. Second, when the dataset is very large, and the internal features is highly dimensional, it may cost awesomely huge storage of the hard disk. Third, it cannot be applied to other tasks such as image recognition. Theoretically, an image is a special graph with each pixel being a graph node. Consider training a ResNet-152 on ImageNet. It is infeasible to store the internal representations of all the images on ImageNet to the hard disk.

9. In the definition of the MDP, could the authors explain why the loss is considered as a part of the state? Since the definition of the state is continuous, the space of the state is thus very large, which may lead to the curse of dimensionality. Solving such an MDP problem is very hard. This is one of the reasons why I doubt the effectiveness of the proposed controller learning.

10. Existing HPO methods should be compared to in the experiment section.

11. I am confused by the number of epochs in Table 5. Why are the searched epochs divisible by 5 or 10.

12. In Table 4, when compared with VRGCN and LWGCN, the advantage of the proposed method over the competitor is insignificant. Sometimes, the proposed method is even inferior to the competitor. According to your analysis, the memory complexity of LWGCN and VRGCN are O(BD) and O(LND), respectively. Why is the memory usage of VRGCN lower than LWGCN on dataset Cora and PubMed? (LN >> B?) Do you use a different D for VRGCN and LWGCN?

13. There can be more descriptions for Figure 4, which I think is not clear enough.

If the authors address my above concerns, I would consider raising my rating to acceptance. Otherwise, I may decrease the rating to strong rejection.

4. [Overall rating] Paper rating (pre-rebuttal)
Weak reject

5. [Justification of rating] Please explain how the strengths and weaknesses aforementioned were weighed in for the rating. Please also mention what you expect to see from the rebuttal that may change your rating.

Please refer to the above [Weaknesses].

6. [Detailed comments] Additional comments regarding the paper (e.g. typos, any suggestions to make the submission stronger).

Please refer to the above [Weaknesses].

11. Final rating
Borderline

12. Explanation of final rating. Describe the rationale for your final rating, including notes based on the rebuttal, discussion, and other reviews.

The authors have addressed many of my concerns. However, according to the rebuttal, the efficiency of the proposed method is far lower than that was presented in the paper because the search time is excluded in Table 4.

With regard to other reviewers, I appreciate their carefully reading, especially R2 finding the inconsistent numbers between the official results and the results reported in this paper. This inconsistency should be clarified by the authors in the revision.

I would like to raise my rating to a borderline (inclined to a weak acceptance). However, even if this paper were accepted, the concerns raised by other reviewers and me should be addressed in the revision, especially the experiments the author promises to include.