View Meta-Reviews

Paper ID
6521

Paper Title
Are We Overfitting to Experimental Setups in Recognition?

META-REVIEWER #1

META-REVIEW QUESTIONS

3. [Meta-review] Consolidation report explaining the decision for the paper based on reviews, rebuttal and discussion with reviewers and AC-triplet
The paper initially received a weak accept, a borderline and a weak reject overall scores. While the reviewers appreciated the FLUID setting, there were a number of important concerns, such as missing references to work on MDT (R1), clarification of the technical contributions and novelty of the paper (R2, R4). The reviewers considered the author response, and although acknowledging that some of the questions were addressed in the rebuttal, noted that several major concerns still remain about the paper. The initial borderline score was also reduced to a weak reject final score.

The AC also considered the message from the authors. In conclusion, while the paper presents an interesting view of the problem, it does not follow up with strong insights. Based on this, the paper, the reviews, the discussions at the AC panel meeting, and the author response, the conclusion is that the weaknesses of the paper dominate its positive points. The final decision is to not accept the paper in its current form.

META-REVIEWER #2

Are We Overfitting to Experimental Setups in Recognition?

We thank the reviewers for their helpful feedback. We appreciate the comments regarding the value of the new framework (FLUID) and insights (R1), extensive experiments and analysis (R1, R2), and quality of the writing (R1, R2, R4). We clarify our evidence for overfitting (R2, R4) and motivation for a new framework (R2, R4) and hope that our rebuttal addresses your individual concerns.

Reviewer 1

MDT Clarification. We will incorporate the missing references & place MDT in better context. There are some practical differences between the reference methods & MDT, but the principles are similar. We will reclassify MDT as a baseline to make the claims more grounded.

For clarification, MDT, unlike the references, does not explicitly optimize for the out-of-distribution task and can be used with a standard classification network. Also, MDT can be done simultaneously with classification in a single forward pass for negligible additional compute by thresholding the maximum/minimum logit.

Reviewer 2 & 4 - Common Concerns

Continual & few-shot learning methods are not overfitting and only underperform relative to the authors' methods because they were not designed for FLUID. The n-shot k-way for few-shot consists of providing exactly n training examples for k classes then evaluating while standard continual learning consists of presenting a batch of training data from the current task to the learner then evaluating on that task and all previous tasks. FLUID is a generalization of these setups that makes more realistic assumptions about the data and learning conditions; therefore, we would expect performance on the previous tasks to transfer to FLUID. FLUID primarily differs from the few-shot setup in that it does not assume provision of exactly n examples from k classes. As in the real-world, the number of classes may increase and the number of examples may vary. FLUID differs from the standard CL setup in two ways: 1) data comes one at a time rather than in fixed-size batches and 2) model may be trained offline before the online phase.

Simple baselines of fine-tuning (FT) and nearest class mean (NCM) significantly outperform SOTA few-shot and CL methods in this more general setting which is strong evidence of overfitting and important knowledge for the community. In general, *the need to develop an entirely new set of algorithms for each specific setting is one definition of overfitting*. We argue against designing methods for narrowly constructed scenarios as 1) the insights and methods do not transfer well to other settings 2) ML systems should be robust to realistic changes in learning conditions.

Motivation for the proposed task. Motivation for FLUID comes from our experimental evidence that methods and insights developed for narrow, unrealistic evaluations do not transfer well to more real-world use cases. Further we empirically identify the unrealistic assumptions in continual learning and few-shot learning which methods leveraged for performance, but ultimately result in baselines outperforming SOTA when the assumptions do not hold in more general settings. FLUID removes these assumptions that we show to be problematic, in general makes as few assumptions as possible, and is a step towards more general evaluations. We further discuss the value of FLUID below.

Reviewer 2

Do merits of new evaluation warrant increased complexity? FLUID is not a complex evaluation workflow but on the contrary, encompasses multiple siloed setups while simultaneously removing unrealistic assumptions. The design choices make FLUID a general, pragmatic test-bed. Yes, the community often simplifies setting to solve the tasks at hand but the assumptions made result in future non-generalizable setups as demonstrated by FLUID.

Novelty of the Insights. We strongly believe that most of the insights presented are novel and unbeknownst to the community at large. Insights like pretraining mitigates catastrophic forgetting and higher capacity networks generalize better to new classes are highly non-trivial and are contrary to the existing understanding (line 639). Systematic analysis to support our intuition helps the community with the finer details while being a baseline for future work.

Evaluating only MoCo is too specific. Most SOTA self-supervised learning methods like MoCo, SimCLR, PiRL etc., rely on contrastive learning and often produce similar representations. Owing to this, we believe that our insights generalize to other SOTA self-supervised learning methods. We also observe the same insight with self-supervised representations learnt from videos via VINCE. FT on ResNet18 via VINCE for same metrics as Table 2 in paper: {18.00; 14.61; 1.89; 1.56; 7.25; 26.27 0.16 / 5.73}.

Exemplar Tuning (ET) is an overfitting algorithm for FLUID. ET is a combination of simple baselines: FT & NCM. Both FT & NCM (strong baselines in standard, few-shot & CL) individually outperform all the few-shot & CL approaches benchmarked in FLUID (lines 733-740). ET adopts the best of FT & NCM and is as simple & generalizable. In supervised learning setup on ImageNet-LT with ResNet18, FT's accuracy is 42.01% while ET is slightly better with 42.5% showing its generalization beyond FLUID.

Reviewer 4

Code & Data. Code is in the supplementary material. As stated in the abstract, code & dataset will be open-sourced.

FLUID appears to just be nothing more than multi-task online learning therefore novelty is limited. The framework should be associated with the online and multi-task community rather than continual learning. CL specifically studies the scenario of learning multiple tasks in an online setting. We found no references to multi-task online in the literature, and online multi-task differs from continual learning in that the tasks change or increment over time in the latter. We clearly enumerate the important ways FLUID differs from current continual learning setups in the paper (lines 206 & 397) and reiterate them in the previous sections of the rebuttal. Also, we empirically show that these subtle differences matter as they significantly affect method performance.

View Reviews

Paper ID
6521

Paper Title
Are We Overfitting to Experimental Setups in Recognition?

Reviewer #1

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.
This work proposes “FLUID” a novel sequential and open-world classification setting based on ImageNet-22K. The setting also considers realistic data availability, sampling head and tail classes. Methods for classification and OOD classification are proposed alongside extensive experimentation of pre-training, meta-learning and supervised learning methods. Both the setting and results are significant, together they demonstrate the inability of current few-shot methods to scale to larger, more realistic settings.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.
The proposed FLUID setting and benchmark provides a significantly improved test-bed for the development of models for deployment in realistic settings. The setting models agents observing head and tail classes in a sequential and open-world setting.

The experimentation is extensive and demonstrates a significant deficiency in current few-shot methods to perform in a more realistic setting. A wide range of prominent methods are thoroughly evaluated using a set of metrics which capture important aspects of performance in the FLUID setting. The authors provide an in-depth discussion of the experimental results and the relevance of each metric in the setting.

The paper is well written and provides substantial supporting material containing supporting experimentation.

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).
Citation should be provided for the significant amount of literature surrounding the Minimum Distance Thresholding (MDT) method. It seems that this method is very similar to methods that have been proposed previously in the OOD literature [1, 2] and should not be claimed as a contribution.

[1] Lee, Kimin, et al. "A simple unified framework for detecting out-of-distribution samples and adversarial attacks." Advances in Neural Information Processing Systems. 2018.
[2] M. Masana, I. Ruiz, J. Serrat, J. van de Weijer, and A. M. Lopez. Metric learning for novelty and anomaly detection. In Proceedings of the British Machine Vision Conference, 2018.

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 strengths of this paper greatly outweigh the weaknesses. Other recent work, such as Meta-Dataset [3], also argues that the problem formulation used by few-shot benchmarks fails to adequately model realistic settings. FLUID provides evidence of this and a realistic alternative. Citation of existing OOD literature or clarification of the MDT novelty may improve the rating.

[3] Triantafillou, Eleni, et al. "Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples." International Conference on Learning Representations. 2019.

Reviewer #2

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.
This paper argued that existing experimental setups in recognition do not fit the practical setting and created a new setting that considers (1) sequential steaming data, (2) flexible training phases, (3) compute aware, (4) open-world.
In this new setting, supervised learning methods, fine-tuning methods, NCM, few-shot learning methods, continue learning methods, OOD methods, etc are evaluated and worse than the proposed method.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.
- Codes are submitted with a comprehensive README.
- An interesting title and attractive abstract.
- Much empirical analysis, systemical experimental results.
- As a research problem, the proposed evaluation framework is interesting and shows some value.

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).
- The authors in the paper argued that previous works overfit the previous experimental setups. However, the proposed Exemplar Tuning is yet another overfiting algorithm for the proposed setup. Together with the contribution #4, it is not surprising that other methods perform bad on the proposed framework. For example, the evaluated few-shot learning methods are not designed to handle the problem of large training samples and continue learning scenario.

- For the contribution #3, the conclusion for MoCo is too specific and less general to other self-supervised learning methods. The claimed other new insights are the same as our intuition. Even if no previous works have similar observations, it is still a relatively minor contribution.

- As a paper focus on the new setting, I would suggest the authors use more words to argue this setting is valuable. From my point of view, the claimed 5 features are important. However, is it necessary to design such a complex workflow for evaluation. In the industrial community, we usually simplify the setting to solve the real world application. I'm worried about this work is creating new setting that are not that valuable to real world.

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

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.
From my personal pespective, it is obvious that old methods work not perfect on a new settings, because they are not designed to handle the new objective. The claimed new findings seem restricted to specific method (e.g., MoCo).

10. [Final rating] Paper rating (post-rebuttal)
Weak reject

11. [Justification of final rating] Explanation of final rating. Describe the rationale for your final rating, including notes based on the rebuttal, discussion, and other reviews.
I thin the evidence for overfitting and movitation in the rebuttal is not strong.

Reviewer #4

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.
The authors introduced a problem setting that involved sequentially evolving different tasks and evaluated several state-of-the-art continal learning methods. They proposed two new methods that perform better than these baselines in their proposed problem setting, and claims the existing methods to overfit easily.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.
The problem of more realistic continal learning is an important one.

The introduction and related work is well-written and offers a good summary of the work.

The plots are clear to illustrate the main intuition.

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).
The novelty is very limited. The entire, so called "Flexible Sequential Data" appears nothing more than a traditional multitask online learning setting, that has been widely studied in a wide range of work in online learning and multi-task learning communities. I think instead of renaming it to another term to serve the continal learning community, it would be more reasonable to place it in somewhere it really belongs.

I am not convinced of its motivations and evaluations. The different state-of-the-art methods included in the comparison is not built for the very specific problem setting that the authors is proposing, and thus, expectedly, underperforms against the authors' methods. Similarly, the argument that they "overfit" is more likely connected to this discrepancy of inductive bia and evaluation environments, than to the inferiority of these methods.

The codes and data are unavailable, which makes reproducibility and evaluation challenging.

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.
as above.

10. [Final rating] Paper rating (post-rebuttal)
Weak reject

11. [Justification of final rating] Explanation of final rating. Describe the rationale for your final rating, including notes based on the rebuttal, discussion, and other reviews.
Hi all, I have read other reviewers' comment and the author's rebuttal, and agree with some points and concerned being made, especially the observations made by R3. I don't think the rebuttal fully address my concerns. Thus, I am keeping my original score.