← Go to **ICLR 2021 Conference** homepage (/group?id=ICLR.cc/2021/Conference)

In the Wild: From ML Models to Pragmatic ML Systems

(/pdf?id=fpew0Il7wl)

Matthew Wallingford (/profile?email=mcw244%40cs.washington.edu), Aditya Kusupati (/profile?id=~Aditya_Kusupati1), Keivan Alizadeh-Vahid (/profile?id=~Keivan_Alizadeh-Vahid1), Aaron Walsman (/profile?id=~Aaron_Walsman1), Aniruddha Kembhavi (/profile?id=~Aniruddha_Kembhavi1), Ali Farhadi (/profile?id=~Ali_Farhadi3)

28 Sept 2020 (modified: 05 Mar 2021) ICLR 2021 Conference Withdrawn Submission Readers: Everyone Show Revisions (/revisions?id=fpew0Il7wl)

Post Decision Revision

Keywords: Benchmark, Real-world, Framework, Few-shot Learning, Sequential Learning, Continual Learning, Long tail, Open-world, Deep Learning

Abstract: Enabling robust intelligence in the wild entails learning systems that offer uninterrupted inference while affording sustained learning from varying amounts of data and supervision. Such ML systems must be able to cope with the openness and variability inherent to the real world. The machine learning community has organically broken down this challenging task into manageable sub tasks such as supervised, few-shot, continual, and self-supervised learning; each affording distinct challenges and a unique set of methods. Notwithstanding this remarkable progress, the simplified and isolated nature of these experimental setups has resulted in methods that excel in their specific settings, but struggle to generalize beyond them. To foster research towards more general ML systems, we present a new learning and evaluation framework - I{N} TH{E} WIL{D} (NED). NED naturally integrates the objectives of previous frameworks while removing many of the overly strong assumptions such as predefined training and test phases, sufficient labeled data for every class, and the closed-world assumption. In NED, a learner faces a stream of data and must make sequential predictions while choosing how to update itself, adapt quickly to novel classes, and deal with changing data distributions; while optimizing for the total amount of compute. We present novel insights from NED that contradict the findings of less realistic or smaller-scale experiments which emphasizes the need to move towards more pragmatic setups. For example, we show that meta-training causes larger networks to overfit in a way that supervised training does not, few-shot methods break down outside of their narrow experimental setting, and self-supervised method MoCo performs significantly worse when the downstream task contains new and old classes. Additionally, we present two new pragmatic methods (Exemplar Tuning and Minimum Distance Thresholding) that significantly outperform all other methods evaluated in NED.

One-sentence Summary: We introduce a new framework, NED, that integrates past frameworks while more closely modeling the real world, present findings to validate the need for pragmatic frameworks like NED & propose two new methods that outperform current methods in NED.

Supplementary Material: zip (/attachment?id=fpew0Il7wl&name=supplementary_material)

Code Of Ethics: I acknowledge that I and all co-authors of this work have read and commit to adhering to the ICLR Code of Ethics

Reviewed Version (pdf): /references/pdf?id=nP8cd7axbA (/references/pdf?id=nP8cd7axbA)

5 Replies

Add **Comment**

Show all from everybody

[−] Submission Withdrawn by the Authors

ICLR 2021 Conference Paper85 Authors Aditya Kusupati (/profile?id=~Aditya_Kusupati1) (privately revealed to you)

10 Nov 2020 ICLR 2021 Conference Paper85 Withdraw Readers: Everyone

Withdrawal Confirmation: I have read and agree with the venue's withdrawal policy on behalf of myself and my co-authors.

Add **Comment**

[−]

Evaluations for Paper "In the Wild: From ML Models to Pragmatic ML Systems"

ICLR 2021 Conference Paper85 AnonReviewer1

08 Nov 2020 (modified: 10 Nov 2020) ICLR 2021 Conference Paper85 Official Review Readers: Everyone

Review:

The authors present a new machine learning and evaluation framework - IN THE WILD (NED). Given a data stream, NED automatically decides how to update itself, adapt to novel classes, deal with out-of-distribution (OOD) samples, and at the same time consider computational cost. NED integrates solutions from a range of sub fields, such as supervised classification, few-shot learning, continual learning, and efficient ML. In addition, the authors introduce two new approaches: (1) Exemplar Tuning for class representations initialization; (2) Minimum Distance Thresholding for OOD detection. They evaluate the proposed approaches and methods from different sub-fields on an image classification problem.

Strengths:

(1) It is important for ML practitioners to make sound decisions during the entire life cycle of model training. This paper makes good efforts towards building more pragmatic ML systems in the wild. The authors present a broad set of practical observations under the framework of NED.

(2) Extensive experimental results are reported to compare the methods from a wide range of sub-fields.

Weaknesses:

(1) The proposed two approaches (Exemplar Tuning and Minimum Distance Thresholding) are relatively simple and straightforward.

(2) The proposed framework is only evaluated on a single image classification task. Due to the nature of the research problem that this work aims to address (a more generic ML framework), it is important to perform evaluations on a broader set of machine learning problems. This is also acknowledged by the authors in their conclusion section (...incorporating other mainstream tasks into NED is an immediate next step).

(3) NED provides compute cost analysis (i.e., the total numbers of multiply-accumulate operations) in the evaluations. However, there are no further insights on how such evaluations enable "compute aware" beyond simple reporting (e.g., how to optimize the total amount of compute).

Rating: 5: Marginally below acceptance threshold

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Add **Comment**

[−]

This paper offers a useful idea, but needs more concrete motivation and evaluation

ICLR 2021 Conference Paper85 AnonReviewer2

28 Oct 2020 (modified: 10 Nov 2020) ICLR 2021 Conference Paper85 Official Review Readers: Everyone

Review:

In this work the authors primarily introduce a new framework for evaluating machine learning models that accounts for a variety challenges posed in real-world systems. The authors argue that most machine learning methods are evaluated in ways that oversimplify real world tasks, specifically they either do not account for streaming data, they do not account for the introduction of new classes or they do not account for practical limitations on computing resources. The “NED” framework that the authors introduce is designed to replicate scenarios where these issues must be accounted for.

In the NED framework, data is provided in a streaming manner and the target classifier is tasked with evaluating each new batch or instance. It is then provided the true class and allowed to update. Crucially, new instances may belong to previously unseen classes or classes that were never seen in pretraining. The number of examples in these classes is drawn from a heavy tailed distribution to integrate a challenge similar to few-shot learning for rare classes. The authors track metrics such as classification accuracy and total compute used. They distinguish their framework from continual learning in that data in their approach may be presented as a stream of individual instances rather than large batches of data and the fact that evaluation is performed continuously.

The authors setup an example of their framework using data derived from the large ImageNet-22K dataset. Models are pre-trained on the Imagenet 1k dataset and then evaluated with NED streaming on new images from Imagenet-22k. The streaming images are limited to 1000 classes of which 250 overlap with pretraining. The authors evaluated a number of existing approaches to continual and few-shot learning, as well as a novel method called exemplar tuning, which is similar nearest class mean approach, but introduces a learnable offset to each class centroid.

Strengths

The work is clearly presented and the authors' goal of pushing a more rigorous form of evaluation for machine learning methods is potentially useful and relevant. The authors do a good job of collecting and contrasting recent relevant methods in the literature as well as performing a useful, independent comparison of these methods on a non-trivial dataset. The methods the authors introduce, such as exemplar tuning, although simple, do seem to offer some benefit over the baseline methods shown and could be promising directions for future research.

I think other researchers could benefit from the comparisons and insights given by these authors.

Weaknesses

While I think the evaluation procedure the authors introduce is potentially useful, I’m not sure about its novelty. It appears to me to be a form of online or continual learning, but with particular goals in terms of its selected class distribution and evaluation metrics. Perhaps the argument that the various aspects of NED need to be considered together is novel.

I think the introduction of the NED framework would greatly benefit from some discussion motivating real-world examples where perhaps other methods of evaluation fall short. The authors introduce it with the motivation of a general recognition system, which is an ambitious, but somewhat abstract goal. Some aspects of the framework, such as “flexible training phases”, are presented as making fewer strong assumptions about the real-world setting, but in this case the framework is assuming the availability of labels and that evaluation should be continuous in real world settings.

It would also be beneficial if empirical results showed some contrast to other methods of evaluation. If these experiments were replicated in a “more traditional” few-shot or continual learning setup, would the results look similar?

This work would also benefit from a greater range of datasets for experiments. The authors present this framework is being very general, but only discuss a single synthetic example for empirical evaluations. Similarly, it’s difficult to asses whether the new methods they present are more generally useful or if their performance on this particular benchmark is anecdotal.

Conclusion

The authors correctly identify a gap between current evaluation of machine learning algorithms and the evaluation needed for general recognition. The work shown has some merits, but I think the authors need to show a more substantial justification for why this approach is novel or why the specifics of the approach are necessary to improving the current state of ML research.

Other questions:

How sensitive is the ImageNet experimental setup to small changes? Can changing the class distribution, the number of overlapping classes or the streaming order lead to different results on these methods? Could reproducibility be seen as a benefit of “more traditional” evaluation methods?

With the goals of this approach, it seems like it might be relevant to think about the effect of methods that can incorporate unlabeled data. Is this something you have considered testing as well?

Rating: 4: Ok but not good enough - rejection

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Add **Comment**

[−] My main concern is experiments.

ICLR 2021 Conference Paper85 AnonReviewer4

28 Oct 2020 (modified: 10 Nov 2020) ICLR 2021 Conference Paper85 Official Review Readers: 🌏 Everyone

Review:

This paper proposed experimental methods and algorithms that perform well in those experiments to adequately evaluate machine learning systems in a more realistic world. This paper is well written, and the visualization and explanation of the problem setting is easy to understand as shown in Fig. 1. This research is considered to be important in that it aims to point out and solve very important problems in machine learning which is mainly studied in the closed world. However, my main concern is experiments.

The experiments are mainly closed in ImageNet. This experimental setting seems to contradicts the purpose of this paper. In this experiments, supervised pre-training totally performs well. I'm not sure if they have the same properties for different datasets. Moreover, the proposed algorithm, exampler tuning, is based on supervised pre-training. Since Eq. (1) is very similar to ProtoNet, you need to show the results based on meta-training not using supervised pre-training as in ProtoNet. The reason for this is that I'm not sure what the proposed method contributes to the performance in the current experimental settings. In terms of modeling, Eq. (1) doesn't seem to be much better than ProtoNet.

Summary: You need to use different datasets. You need to show the results the proposed method based on meta-training.

Rating: 4: Ok but not good enough - rejection

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Add **Comment**

[−] To play devil's advocate, this paper reinvents the area of data stream classification

ICLR 2021 Conference Paper85 AnonReviewer3

27 Oct 2020 (modified: 10 Nov 2020) ICLR 2021 Conference Paper85 Official Review Readers: 🌏 Everyone

Review:

The paper identifies shortcomings of standard evaluation strategies used in deep learning when considering classification of evolving data streams. It creates a benchmarking scenario based on ImageNet that can be used to evaluate image classification methods in a streaming setting in which new classes appear over time. In addition to metrics based on classification accuracy, the paper also considers the number of compute operations required, and the AUROC observed when detecting novel classes. The paper uses the benchmark to evaluate standard variants of transfer learning, simple nearest centroid classification based on extracted features, several few-shot methods, two methods for detecting samples that are out-of-distribution, and the learning without forgetting method for continual learning. The authors additionally propose their own threshold-based method for out-of-distribution detection, and a method called "exemplar tuning", which is closely related to a centroid classifier that is fine-tuned using discriminative training. In addition to standard supervised pretraining, the papers also evaluates momentum contrast pretraining. Based on the results, it is fair to say that the proposed exemplar tuning generally outperforms all other methods when performing classification, in spite of its simplicity, and the proposed threshold-based out-of-distribution method outperforms the other two out-of-distribution methods.

There are four main points of concern:

a) The paper investigates a particular form of data stream classification (where deep learning is used for image classification and the concept drift is limited to the occurrence of new classes) and appears to imply that this is a new setting that nobody has considered previously. It completely ignores existing work on data stream classification and evaluation methodologies developed in this area.

b) The proposed exemplar tuning strategy seems closely related to existing "centroid-based" methods, particularly the scaled cosine similarity classifier used in few-shot learning (cf. von Mises-Fisher mixture), which is generally initialized to the set of centroids ("generative model") and then fine-tuned ("discriminative model"). It is unclear whether there is an advantage to the existing approach.

c) The proposed out-of-distribution approach is very simple and computationally appealing. However, it seems difficult to believe that it can compete with the state-of-the-art in density estimation or one-class classification. It is unclear to me whether the two baseline out-of-distribution detectors considered in the paper are state of the art.

d) The paper considers the number of operations performed but not the amount of memory consumed (which is a standard evaluation criterion in data stream classification). A case in point is that D never gets pruned in Algorithm 1 so it will invariably run out of memory in a true data stream context. This reinforces the point that the paper does not consider any existing work in data stream classification.

Rating: 3: Clear rejection

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Add **Comment**

About OpenReview (/about)
Hosting a Venue (/group?id=OpenReview.net/Support)
All Venues (/venues)

Frequently Asked Questions (/faq)
Contact (/contact)
Feedback
Terms of Service (/legal/terms)
Privacy Policy (/legal/privacy)

OpenReview is created by the Information Extraction and Synthesis Laboratory (http://www.iesl.cs.umass.edu/), College of Information and Computer Science, University of Massachusetts Amherst. We gratefully acknowledge the support of the OpenReview sponsors: Google, Facebook, NSF, the University of Massachusetts Amherst Center for Data Science, and Center for Intelligent Information Retrieval, as well as the Google Cloud Platform for donating the computing and networking services on which OpenReview.net runs.