General notes

We would like to thank reviewers and area chairs for their time and effort required to handle the review process of our paper. We would like to sincerely thank the reviewers for their thorough analysis of our paper and for their insightful feedback. We have addressed all the comments provided by reviewers and we feel that it helped to further clarify and strengthen the paper.

Reviewer #2

Questions

1. Please provide an "overall score" for this submission.

3: A clear reject. I vote and argue for rejecting this submission.

2. Please provide a "confidence score" for your assessment of this submission.

5: You are absolutely certain about your assessment. You are very familiar with the related work.

3. Please provide detailed comments that explain your "overall score" and "confidence score" for this submission. You should summarize the main ideas of the submission and relate these ideas to previous work at NIPS and in other archival conferences and journals. You should then summarize the strengths and weaknesses of the submission, focusing on each of the following four criteria: quality, clarity, originality, and significance.

Reviewer #2, point 1

This work proposes improvements to the prototypical network approach to few-shot classification introduced in [25]. The improvements are three-fold: (i) the addition of a learned temperature parameter in the computation of the softmax-normalized centroid distance; (ii) a modification to conditional normalization [16, 17] in which the conditioning is done as a function of the task-specific dataset as a whole; and (iii) an auxiliary prediction task implemented as an additional head in the architecture that outputs the logits for a 64-way classification task.

The method gives better performance on the standardized miniImageNet benchmark as well as on a few-shot split of CIFAR100 introduced by the authors.

Authors, response to Reviewer #2, point 1

We thank the reviewer for an accurate summary of the contributions of the paper

QUALITY

Reviewer #2, point 2

The theoretical contribution of this paper is to write out the gradient of the objective function from [25] with the addition of the temperature scaling parameter in the softmax computation. The limiting cases (zero and infinite temperature) are discussed for intuition even though they do not occur in practice as I assume the temperature parameter is transformed to be positive finite (although this is not stated).

Authors, response to Reviewer #2, point 2

The main theoretical contribution is the mathematical analysis of the limiting cases of α. To our knowledge, this is novel and non-trivial. It covers a very different case and provides very different insights than Hinton et al. *Distilling ... 2015 (incorporated as a reference). To address reviewer's concern we have moved the proof of Lemma 1 into the main body of the paper.*

Moreover, we would like to emphasize that we explicitly used Lemma 1 and its discussion to drive algorithm design and empirical studies. Based on Lemma 1 and insights obtained from the analysis of the limiting cases, we posed the following hypothesis: "there is an optimal value of scaling parameter α for a given combination of dataset, metric and task". This hypothesis was validated on two independent datasets in the experimental part of the paper. Figure 3 (especially (a) and (b)) demonstrates the behavior of generalization error as α approaches the extreme values.

Reviewer #2, point 3

There are some flaws in quality throughout the paper:

Several technical claims / hypotheses made in the paper are not evidenced:

"Neural network initialization and batch norm encourages this regime."

"If Df is large, the network may have to work outside of its optimal regime to be able to scale down the feature representation."

"We hypothesize that the importance of TEN is not uniformly distributed over the layers in the feature extractor. The lower layers closer to the similarity metric need to be conditioned more than the shallower layers closer to the image because pixel-level features extracted by shallow level layers are not task specific."

Authors, response to Reviewer #2, point 3

To address the concerns of the reviewer we have removed these statements.

Reviewer #2, point 4

A conceptual point is made in the conclusion unnecessarily: it does not appear that the fact that "the scaled cosine similarity...does not belong to the class of Bregman divergences" is used anywhere in the paper, yet it is mentioned in the conclusion.

Authors, response to Reviewer #2, point 4

In order to address the reviewers comment we removed this point and replaced it with the following: "We showed that the scaled cosine similarity performs at par with Euclidean distance, unlike its unscaled counterpart"

CLARITY

The paper is reasonably clear. Some points of are omitted:

Reviewer #2, point 5

- Why not report results with temperature scaling and matching networks (instead of prototypical networks)?

Authors, response to Reviewer #2, point 5

Table 2 shows that the prototypical approach with Euclidean distance and scaled cosine are close and both are superior to [30] (results are based on the same feature extractor as [25,30]). Therefore, we base our results on [25]. We state this clearly in experimental section now.

Reviewer #2, point 6

- What is the architecture of the TEN? What motivated this architecture (specifically, why use the centroids, which enable parameter sharing amongst the feature extractors, but still requires the TEN with additional parameters)?

Authors, response to Reviewer #2, point 6

TEN is defined in detail in the supplementary material and we reference it in the revised manuscript as we discuss the architecture in the main body. The task representation defined as the mean of task class centroids (i) reduces the dimensionality of TEN input and (ii) replaces expensive RNN/CNN/attention modeling. On the other hand, it is an effective way to cluster tasks. Tasks having larger number of similar classes in common will tend to cluster closer in the task representation space. Alternative task encoding schemes could be considered. This is outside of the scope of the current paper. The extra fully connected layers in TEN are necessary to (i) sufficiently decouple the few-shot classification from the encoding of the task and (ii) decouple the generation of shift/scale parameters across convolutional layers in the feature extractor. It might not be necessary with higher capacity feature extractors.

Reviewer #2, point 7

- What is the reason that the TEN "underperformed"? Was it overfitting?

Authors, response to Reviewer #2, point 7

*We observed two difficulties with the TEN training. We saw convergence problems initially and overfit (or convergence to a local minimum) later on in the optimization process. Accordingly, in Table 3 we see that when task conditioning (TC) is added without auxiliary training (AT) the overall performance drops significantly (*cf. *rows 2, 4).*

Reviewer #2, point 8

The experimental comparison in and the discussion of Table 1 does not identify the differences in architectural complexity between the reported methods (i.e., [1, 14, 15] and the proposed work employ a ResNet architecture while the other methods employ a shallow convolutional network).

Authors, response to Reviewer #2, point 8

To address the reviewer's concern we have included a discussion of the complexity of the entries in Table 1

ORIGINALITY

Reviewer #2, point 9

The architecture and algorithm are a combination of previously proposed methods (see "SIGNIFICANCE" below).

Authors, response to Reviewer #2, point 9

*In our view, the proposed conditioning architecture and its training scheme constitute a significant novel contribution advancing state of the art in FSL. Conditioning in [3,16,17] was not used for FSL. We had to introduce two significant modifications to the original conditioning architecture to achieve performance gains. First, we introduced the post-multipliers γ_0 and β_0 encoding prior importance of each TEN layer. Second, we introduced auxiliary co-training with a normal multi-way classifier. Without either of these components, the task conditioning network has problems with convergence initially and overfit later on in the optimization process. This is illustrated in Table 3, where we see that when task conditioning (TC) is added without auxiliary training (AT) the overall performance drops significantly (*cf. *rows 2, 4).*

Reviewer #2, point 10

The problem setup is not novel, although the authors apply the few-shot episode generation procedure of [30] to the CIFAR100 dataset.

Authors, response to Reviewer #2, point 10

The novelty of the problem setup has never been part of the contributions that we claim. We are convinced that the contributions that we claim are novel and significant.

SIGNIFICANCE

The work is incremental as a combination of previously proposed methods applied to prototypical networks for the task of few-shot classification. In particular:

Reviewer #2, point 11

- "Metric scaling" is the addition of a learnable temperature parameter to the normalized distance computation in the regime of Matching Networks [30] or Prototypical Networks [25].

Authors, response to Reviewer #2, point 11

To our knowledge, this is the first study to (i) propose metric scaling to improve few-shot algorithms, (ii) mathematically analyze its effects on objective function updates and (iii) empirically demonstrate its positive effects on few-shot performance. In our view, this makes it a novel and significant contribution. To address the reviewer's concern we have further clarified this point in the section "Summary of contributions".

Reviewer #2, point 12

- "Task conditioning" makes use of a task-dataset-conditioning network to predict the scale and offset parameters of batch normalization layers as in [3, 16, 17].

Authors, response to Reviewer #2, point 12

To address the point raised by the reviewer, we have extended the related work section discussion of the modifications we introduced to make this approach work in the few-shot learning setting.

Reviewer #2, point 13

- Auxiliary tasks are known to be beneficial to few-shot learning [*,**] and learning in general.

Authors, response to Reviewer #2, point 13

References [,**] do not seem to address FSL scenario. Even though auxiliary co-training is beneficial for learning in general, according to [*] "little is known on* when *multitask learning works and whether there are data characteristics that help to determine its success". Indeed, Table 3 demonstrates that the addition of AT alone only provides marginal gain in most cases (*cf. *rows 2, 3). AT provides significant gain* only *in combination with TC (*cf. *rows 2, 3, 5). This cannot be inferred from existing literature and significantly advances state of the art. Therefore, in our opinion this is a significant and novel contribution. To address the reviewer's concern, we discussed [*] as a reference in related work and extended the discussion of Table 3.*

Reviewer #2, point 14

Moreover, I disagree with the argument (in the conclusion section) that the task sampling technique should be modified for improved performance on a few-shot task, as we should be hesitant about tuning the dataset (which is a feature of the problem) to an algorithm.

Authors, response to Reviewer #2, point 14

We thank the reviewer for this insightful comment. To address it, we removed the part of discussion related to task sampling tuning from the conclusion section.

SPECIFIC COMMENTS

Reviewer #2, point 15

pg. 1: "...one aims to...learn a model that extracts information from a set of labeled examples (sample set)..." Introduce the terminology "support set" alongside "sample set". I believe the restriction to the labelled support/unlabelled query setting is not representative of recent works in few-shot learning; e.g., consider [***, ****], which deal with learning with unlabelled data in the support set.

Authors, response to Reviewer #2, point 15

We addressed the comment by introducing the notion of support set together with sample set. We chose not to expand the discussion on the possibility of using the unlabelled samples in the support set, because it is clearly outside of the scope of the paper.

Reviewer #2, point 16

pg. 1: It is strange to introduce few-shot learning with Ravi & Larochelle as the first citation, then to claim that the problem has subsequently been reframed by Ravi & Larochelle as meta-learning – they are the same paper! This needs to be rewritten to correctly capture the nuanced difference between few-shot and meta-learning.

Authors, response to Reviewer #2, point 16

We addressed the comment by removing Ravi & Larochelle from the first citation.

Reviewer #2, point 17

pg. 1: The claim that certain approaches to few-shot learning and meta-learning are "influential" and "central to the field" is subjective and

Authors, response to Reviewer #2, point 17

We replaced "influencial" and removed "central to the field" with "recent" to make the statement more balanced.

Reviewer #2, point 18

pg. 1: "a feature extractor (or more generally a learner)" A feature extractor is more general than a neural network with learned parameters, so this relationship needs to be reversed. Since you consider models with learned parameter, it would be sufficient to rewrite this as "a feature extractor with learned parameters."

Authors, response to Reviewer #2, point 18

We would like to thank the reviewer for the insightful comment. We addressed it by rewriting text fragement simply as "a feature extractor".

Reviewer #2, point 19

line 57: "parameterized by ϕ, mapping x to z, a representation space of dimension D_z" z is an element of the representation space, not the representation space itself.

Authors, response to Reviewer #2, point 19

We thank the reviewer for pointing out this inconsistency. We addressed it by shortening the sentence to "parameterized by ϕ, mapping x to z."

Reviewer #2, point 20

line 59: "can directly be used to solve the few-shot learning classification problem by association" This needs a more thorough explanation and a more thorough description of the differences between [30] and [25]. In particular, the training criterion for [25] is discussed in lines 62-63 but [30]'s is not.

Authors, response to Reviewer #2, point 20

We addressed the comment by including a more detailed explanation of the Matching networks by inserting the following sentence "For example, Matching networks[30] use sample-wise attention mechanism to perform kernel label regression)."

Reviewer #2, point 21

line 60: "proposed to introduce inductive bias" Framing the computation of the class centroid as an "inductive bias" is not useful in this context unless it is identified why it is a useful inductive bias.

Authors, response to Reviewer #2, point 21

We addressed the comment by removing the "inductive bias"

Reviewer #2, point 22

line 61: "a unique feature representation" The mean of embeddings is not necessarily unique.

Authors, response to Reviewer #2, point 22

we replaced this by "defined a feature representation"

Reviewer #2, point 23

line 61: "for each class k" It is confusing to use k to index classes when above K have been used to count examples in each class.

Authors, response to Reviewer #2, point 23

We addressed the comment by using K for the number of classes and M for the number of shots consistently throughout the paper.

Reviewer #2, point 24

line 77-8: "This is the case of Matching Networks [30], which use a Recurrent Neural Network (RNN) to accumulate information about a given task." The vanilla version of MNs does NOT use an RNN; only the "full-context embedding" version requires it.

Authors, response to Reviewer #2, point 24

We addressed this comment by rephrazing the relevant part of the sentence as follows: "...which optionally use a Recurrent Neural Network (RNN)..."

Reviewer #2, point 25

line 108-109: "We observed that this improvement could be directly attributed to the interference of the different scaling of the metrics with the softmax." This needs an experimental result, and so likely does not belong in the "Model Description" section.

Authors, response to Reviewer #2, point 25

We would like to thank the reviewer for pointing this out. We have addressed the concern by replacing "observed" with "hypothesize". The experimental result confirming the hypothesis is provided in Table 2.

Reviewer #2, point 26

lines 224: "We re-implemented prototypical networks..." Do the experimental results remain the same when employing the authors' original code (https://github.com/jakesnell/prototypical-networks) with the addition of the temperature scaling parameter?

Authors, response to Reviewer #2, point 26

We did not experiment with the original code, because we were able to reproduce the original results using our re-implementation.

[*] Alonso, Héctor Martínez, and Barbara Plank. "When is multitask learning effective? Semantic sequence prediction under varying data conditions." arXiv preprint arXiv:1612.02251 (2016).

[**] Rei, Marek. "Semi-supervised multitask learning for sequence labeling." arXiv preprint arXiv:1704.07156 (2017).

[***] Finn, Chelsea, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. "One-shot visual imitation learning via meta-learning." In CoRL, 2017.

[****] Metz, Luke, Niru Maheswaranathan, Brian Cheung, and Jascha Sohl-Dickstein. "Learning Unsupervised Learning Rules." arXiv preprint arXiv:1804.00222 (2018).

4. How confident are you that this submission could be reproduced by others, assuming equal access to data and resources?

3: Very confident

Reviewer #3

Questions

1. Please provide an "overall score" for this submission.

8: Top 50% of accepted NIPS papers. A very good submission; a clear accept. I vote and argue for accepting this submission.

2. Please provide a "confidence score" for your assessment of this submission.

4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

3. Please provide detailed comments that explain your "overall score" and "confidence score" for this submission. You should summarize the main ideas of the submission and relate these ideas to previous work at NIPS and in other archival conferences and journals. You should then summarize the strengths and weaknesses of the submission, focusing on each of the following four criteria: quality, clarity, originality, and significance.

This paper introduces a novel prototypical network-inspired architecture for few-shot learning that incorporates several useful modifications that lead to wins in classification accuracy. It analyzes the effect of temperature scaling the output softmax and discusses the nature of gradients in both the low temperature and high temperature regimes. It also proposes a method to condition the embedding network on the few-shot task. Finally, it proposes to train jointly on an auxiliary prediction task. The combination of these modifications leads to significant performance wins on an established few-shot classification benchmark. Overall a strong paper.

The contributions of this work are significant. The analysis of scaling presented in section 2.1 is helpful to further understand pathologies that may arise when training prototypical networks and justifies the use of the scaling parameter. Conditioning the embedding network on the task is an intuitive yet unexplored avenue to improved few-shot classification performance. The ablation study in Table 3 shows the contributions of these components as well as the effect of auxiliary training on classification. Moreover, the introduction of the few-shot CIFAR-100 variant will likely prove useful for the community.

The paper is very well-written. Experimental procedures are described in sufficient detail and the related work section sufficiently covered relevant material in my opinion.

I verified equations 1-4 and they are correct to my knowledge.

Were task representations other than the mean prototype experimented with? It seems that higher order statistics of the class representations might be helpful.

4. How confident are you that this submission could be reproduced by others, assuming equal access to data and resources?

3: Very confident

Authors, response to Reviewer #3

We would like to thank Reviewer #3 for the review. We agree that the higher order statistics of class representations might be helpful. Designing more powerful task representations definitely looks like a very promising venue for future work. We included that in the conclusions section.

Reviewer #4

Questions

1. Please provide an "overall score" for this submission.

6: Marginally above the acceptance threshold. I tend to vote for accepting this submission, but rejecting it would not be that bad.

2. Please provide a "confidence score" for your assessment of this submission.

5: You are absolutely certain about your assessment. You are very familiar with the related work.

3. Please provide detailed comments that explain your "overall score" and "confidence score" for this submission. You should summarize the main ideas of the submission and relate these ideas to previous work at NIPS and in other archival conferences and journals. You should then summarize the strengths and weaknesses of the submission, focusing on each of the following four criteria: quality, clarity, originality, and significance.

Reviewer #4, point 1

Summary

This paper investigates how to learn metrics for few-shot learning. The authors theoretically and empirically show the importance of metric scaling in the learning objective. They further show that learning task conditioning metrics could improve the results (by roughly 1 %). The paper provides an insightful introduction and related work to categorize existing few-shot learning algorithms. However, the idea of learning conditioning metrics (or features) seems to be proposed already but the discussion is missing. The experiments also need more explanation to clarify the contributions.

Strengths

S1. The idea of metric scaling is theoretically and empirically justified. Section 2.1 should benefit not only few-shot learning, but the whole metric learning community.

S2. The introduction and related work are well organized and inspiring.

Authors, response to Reviewer #4, point 1

We thank the reviewer for outlining the strengths of the paper.

Main weakness (comments)

Reviewer #4, point W1

W1. The idea of learning task conditioning metrics (or feature extractors) has been proposed (or applied) in [30], where they learn f(x|S) and g(x|S) where S is the data of the current task. The authors should discuss the relatedness and difference. More informatively, why the proposed method significantly outperforms [30] in Table 1.

Authors, response to Reviewer #4, point W1

Task conditioning in [30] is implicit: there is no notion of task embedding. In our work, we explicitly introduce a task representation (see Fig. 1) computed as the mean of the task class centroids. This is much simpler than individual sample level LSTM/attention models in [30]. Conditioning in [30] is applied as a postprocessing of a fixed feature extractor. We condition the feature extractor by predicting its own batch normalization parameters with the TEN thus making feature extractor dynamic without cumbersome fine-tuning on S. We included the clarifying remarks outlining relatedness and difference in the Related work section.

Reviewer #4, point W2

W2. The comparison in Table 1 might not be fair, since different models seem to use different feature extractors. The authors should emphasize / indicate this. The authors should either re-implement some representative methods ([25], [30]) using the same architecture (e.g., Res-12) or implement the proposed method in the architecture of [25], [30] for more informative comparison. Is the first row of Table 3 exactly [25] but with the Res-12 architecture? If not, since the proposed model is based on [25], the authors should include it (with the same architecture of the proposed method) in Table 3.

Authors, response to Reviewer #4, point W2

Yes, the first row of Table 3 is exactly [25] but with the Res-12 architecture. We have provided more details on the architectural complexity of the algorithms compared in Table 1 to provide for for more informative comparison.

Reviewer #4, point W3

W3. In Table 3, it seems that having scaling doesn't significantly improve the performance. Can the authors discuss more on this? Is it because the Euclidean metric is used? If so, the footnote 6 might be misleading because scaling should help a lot for the cosine similarity.

Authors, response to Reviewer #4, point W3

Yes, it is because of the Euclidean metric. Scaled cosine with AT and TC is similar to Table 3, row 5. We amended the footnote accordingly to address the concern of the reviewer.

Reviewer #4, point W4

W4. Section 2.4 on the auxiliary task is not clear. Is the auxiliary task a normal multi-way classification task? It would be great to have the detailed algorithm of the training procedure in the supplementary material.

Authors, response to Reviewer #4, point W4

We have clarified that the auxiliary task is the normal multi-way classification task in the description of the training procedure. We will also release the code to with the exact algorithm used to train the network.

Minor weaknesses (comments)

Reviewer #4, point W5

W5. The idea of scaling (or temperature) is mentioned in some other papers like Geoffrey Hinton et al., Distilling the Knowledge in a Neural Network, 2015 It would be great to discuss the relatedness.

Authors, response to Reviewer #4, point W5

The main theoretical contribution of our paper is the mathematical analysis of the effect of the limiting cases of α on the few-shot update rule. To our knowledge, this is novel and non-trivial. It covers a very different case and provides very different insights than Geoffrey Hinton et al., Distilling the Knowledge in a Neural Network, 2015. To address reviewer's concern, we have incorporated the paper pointed out as a reference and outlined the novelty of the current work with respect to it.

Reviewer #4, point W6

W6. The way FC 100 is created seems to favor learning task conditioning metrics, since now each task is within a superclass, and the difference between tasks (e.g., from different superclasses) will be larger than that if we sample the tasks uniformly from 100 classes. It would be inspiring to compare the improvement of the proposed methods on FC 100 and a version where a task is sampled uniformly.

Authors, response to Reviewer #4, point W6

We thank the reviewer for the insightful comment. The idea behind splitting the CIFAR 100 by superclass was to create a more challenging dataset that will reduce overfitting on classes seen during training phase. For example, if sub-classes of "people" superclass are part of train, validation and test subsets, that may create a potential for overfit based on ability to identify samples from "people" superclass. We now clarify it in the main text. Moreover, we would like to stress that we still sample all the tasks uniformly at random within train, validation and test subsets. Therefore, each task with very high probability contains samples belonging to classes from several superclasses. We further clarified this point in the text to avoid confusion. In other words, each task is not within a superclass. We have stressed and clarified this in the text to avoid confusion.

Reviewer #4, point W7

W7. Why the experiment on the popular Omniglot dataset is not conducted?

Authors, response to Reviewer #4, point W7

We focused on miniImagenet and CIFAR as they are more challenging, and the error rate is more sensitive to model improvements. We included a note pointing this out at the beginning of Section 3.1.

Reviewer #4, point W8

W8. Please add the dataset reference to Table 1. Please include background (i.e., algorithm or equation) on [25] in Section 2.1 since it is the baseline model used in the paper.

Authors, response to Reviewer #4, point W8

We included the dataset reference in the caption of Table 1. The equations describing [25] are presented in Section 1.1. We now reference this section to explicitly link the mention of prototypical networks in Section 2.1 to the equations presented in Section 1.1. by inserting "Snell et al.[25]using approach described in detail in Section 1.1 found that ..."

4. How confident are you that this submission could be reproduced by others, assuming equal access to data and resources?

3: Very confident