A Simple Framework for Cross-Domain Few-Shot Recognition with Unlabeled Data

Ashraful Islam[1], Chun-Fu Chen[2], Rameswar Panda[2], Leonid Karlinsky[2], Rogerio Feris[2], Richard Radke[1]
[1]Rensselaer Polytechnic Institute [2]IBM T. J. Watson Research
islama6@rpi.edu, chenrich,rsferis@us.ibm.com, rpanda@ibm.com
leonidka@il.ibm.com, rsferis@us.ibm.com, rjradke@ecse.rpi.edu

Abstract

Most existing works in few-shot learning rely on meta-learning the model on a large base dataset which is typically from the same domain as the target dataset. We tackle the problem of cross-domain few-shot learning where there is a large shift between the base and target domain. We propose a simple solution to utilize unlabeled images from the novel/base dataset. We calculate pseudo soft-label from the weakly-augmented version of the unlabeled image and compare it with the strongly augmented version. We also minimize the supervised cross-entropy loss for the labeled base dataset at the same time. We show that the proposed network learns representation that can be easily adapted to the target domain even though it has not been trained with target-specific classes during the pretraining phase. Our model outperforms the current state-of-the art method by 2.7% for 5-shot and 3.6% for 1-shot classification in the BSCD-FSL benchmark.

1. Introduction

The tremendous success of deep learning in visual recognition tasks is, to a great extent, attributed to the availability of large scale labeled datasets. While humans can recognize an object by looking only at a few examples, modern deep neural networks require hundreds or thousands of images for each category to achieve human-level visual recognition capability. This has led to the research on few-shot learning which aims at learning from a much smaller dataset. In a typical few-shot learning setting, there are two stages: meta-training and meta-testing. In the meta-training stage, a base dataset with labeled images is provided to train the model. In the meta-testing stage, the learned model is quickly adapted to a set of novel classes with only a few examples per class (the support set) and evaluated on a set of test images from the same novel classes (the query set). The base classes and novel classes are typically disjoint, but the images are obtained from the same domain. However, in many real world settings, training the model on a base dataset from the same domain as the target dataset is difficult and infeasible. Guo *et al*. proposed a cross-domain few-shot benchmark, BSCD-FSL, which contains datasets from extremely different domains. The benchmark shows that traditional pretraining and finetuning outperforms more complicated meta-learning based few-shot learning methods by a great margin.

In real-world scenarios, the target domain should have many unlabeled images, and it might be beneficial to use the unlabeled data to learn more target domain specific representations. One potential solution is to use self-supervised representation learning. However, as pointed out by [9], plain self-supervised learning struggles to outperform the naive transfer learning baseline. We propose a pretraining strategy using both the base dataset and unlabeled images from the target domain to learn a representation that can be easily adapted to the few-shot task. We show that labeled images from the base dataset are still important to learn generic image features, and images from the target domain, even if unlabeled, can help developing more target domain specific representations. Our main contribution is a simple approach to train a model for cross-domain few-shot learning using the unlabeled images from the target domain. Our method significantly outperforms the current state of the art in the BSCD-FSL benchmark with unlabeled images by **2.7**% for 5-shot and **3.6**% for 1-shot learning in terms of average top-1 accuracy. Figure 1 illustrates our approach.

2. Related Work

Few-shot learning methods can be divided into three broad categories - generative [18], metric-base [11, 15, 13] and adaptation-based [4, 6]. Guo *et al*. [2] proposed a cross-domain few-shot learning benchmark, and noted that existing state-of-the-art approaches fail to achieve good accuracy on this benchmark. One potential solution could be to use an unlabeled dataset from the target to learn representations that are adaptable to a completely different domain. Many approaches also explored few-shot learning with unlabeled data [5, 7, 10]; however, most of these still assume a smaller gap between the base and target domains. Re-

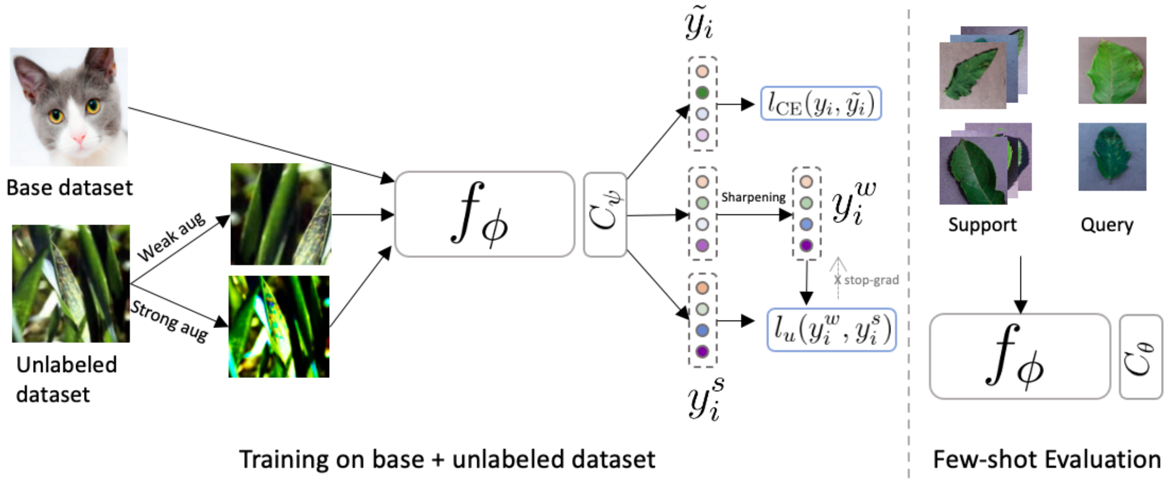


Figure 1: Diagram of our approach. In the pretraining stage, we use both the labeled base dataset and the unlabeled target dataset to learn the feature extractor f_ϕ. During the few-shot evaluation stage, we use the frozen feature extractor f_ϕ to learn a linear header C_θ on the support set and test on the query set.

cently, Phoo *et al*. [9] proposed STARTUP, which learns by self-training a source domain specific representation on the unlabeled target data. Our method is inspired from their insight that a classifier pretrained on the base domain can induce a grouping on the target domain, even though the unlabeled target images might be from a completely different domain. One major difference between STARTUP and our approach is that STARTUP uses a pretrained fixed teacher and learns a student model by self-distillation. We argue that the fixed teacher might not be adaptive enough to learn a representation for the target dataset. We adopt a dynamic approach similar to FixMatch [12] by imposing consistency regularization. However, while FixMatch is a semi-supervised technique where the unlabeled data is assumed to be from the same domain, our approach is applicable to the cross-domain few-shot learning problem.

3. Method

Problem Formulation. A few-shot learning task consists of a support set S, which containing K data points from N classes for N-way K-shot task, and a query $Q = \{x_i\}_{i=1}^m$ consisting of data points only from the N classes of the support set. The goal is to classify the query points with the help of the labeled support set. In the typical few-shot learning setting, (1) an embedding is learned from the base/source dataset $\mathcal{D}_S$, (2) a linear classifier is learned on top of the fixed embedding on the support set, and (3) the classifications of the query data points are determined. The difference between the typical few-shot learning setup and cross-domain few-shot learning is that the base/source dataset is drawn from a very different domain than the target domain. In our setting, we are additionally provided unlabeled data points $\mathcal{D}_U = \{x_i\}_{i=1}^{N_U}$ from the target domain. Given the base dataset $\mathcal{D}_S$, and an unlabeled set $\mathcal{D}_U$, we need to learn an embedding that can extract a representation that can be used for few-shot learning evaluation in the target-domain.

Approach. Denote the embedding network as f_ϕ that embeds an input image x to a d-dimensional vector $f_\phi(x)$. We add a classifier header C_ψ on top of f_ϕ, which predicts n_c logits from the embeddings, where n_c is the total number of classes in the base dataset. Since the labels of the data points of the base dataset are provided, we can learn the embedding f_ϕ and header C_ψ by optimizing the supervised cross-entropy loss: $l_{\text{CE}}(y_i, \tilde{y}_i) = H(y_i, \tilde{y}_i)$, where $\tilde{y}_i = \texttt{Softmax}(C_\psi(f_\phi(x_i)))$.

Given an image x_i from the unlabeled set $\mathcal{D}_U$, we compute the model's prediction from a weakly-augmented version (denoted as x_i^w) and from a strongly-augmented version (denoted as x_i^s) of the unlabeled image :

$$y_i^w = \texttt{stopgrad}(\texttt{Softmax}(C_\psi(f_\phi(x_i^w))/\tau)) \tag{1}$$

$$y_i^s = \texttt{Softmax}(C_\psi(f_\phi(x_i^s))) \tag{2}$$

where τ is a sharpening parameter. Note that we do not let gradient pass through y_i^w which is denoted as $\texttt{stopgrad}$ operation. We minimize the cross-entropy loss function $l_U(y_i^w, y_i^s) = H(y_i^w, y_i^s)$ that works like a consistency regularizer so that the network predicts similar scores for different augmented versions of the image. It is important that we apply $\texttt{stopgrad}$ to the weakly-augmented version and make output of the strongly-augmented image similar to the output of the weak-augmentation. See Sec. 4.1 for details. The total loss function is:

$$\mathcal{L} = \frac{1}{N_S} \sum_{(x_i, y_i) \in \mathcal{D}_S} l_{\text{CE}}(y_i, \tilde{y}_i) + \frac{1}{N_U} \sum_{x_i \in \mathcal{D}_U} l_U(y_i^w, y_i^s) \tag{3}$$

4. Experiments

Dataset. We use the BSCD-FSL benchmark [2], which contains novel data from CropDisease [8], EuroSAT [3],

	EuroSAT		CropDisease		ChestX		ISIC	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MAML*	-	71.70±.72	-	78.05±.70	-	23.48±.48	-	40.13±.58
ProtoNet*	-	73.29±.71	-	79.72±.79	-	24.05±1.01	-	39.57±.57
MetaOpt*	-	64.44±.73	-	68.41±.73	-	22.53±.91	-	36.28±.50
Transfer†	60.73±.86	80.30±.64	69.97±.85	90.16±.49	22.71±.40	26.71±.46	30.71±.59	43.08±.57
SimCLR †	43.52±.88	59.05±.70	78.23±.83	92.57±.48	22.10±.41	25.02±.42	26.25±.53	36.09±.57
STARTUP (no SS)†	62.90±.83	81.81±.61	73.30±.82	91.69±.47	22.87±.41	26.68±.45	32.24±.62	46.48±.61
STARTUP†	63.88±.84	82.29±.60	75.93±.80	93.02±.45	**23.09**±.43	26.94±.45	32.66±.60	47.22±.61
Transfer	58.95±.86	80.39±.60	69.95±.88	89.85±.51	21.62±.39	25.19±.44	31.62±.58	44.89±.58
Ours	**69.08**±.82	**88.26±.46**	**83.22**±.77	**95.78**±.34	22.85±.42	**27.82**±.44	**34.84**±.59	**48.26**±.56

Table 1: 5-way 1-shot and 5-shot scores on the BSCD-FSL benchmark datasets. The mean and 95% confidence interval of 600 runs are reported. The * indicates that the numbers are reported from [2] where no unlabeled data is used. The † are the numbers reported from [9], which uses 20% of the original set as the unlabeled dataset. We also use similar numbers of unlabeled images as [9]; however, the splits might be different for random sampling.

ISIC [1], and ChestX [17]. The base dataset is mini-ImageNet [16]. The novel datasets are chosen based on increasing dissimilarity from the mini-ImageNet dataset. Following [9], we randomly sample 20% of the data from each novel dataset to construct the unlabeled set $\mathcal{D}_U$, and the remaining images are used for evaluation, where we perform 5-way 1-shot and 5-way 5-shot classification. For evaluation metric, we report top-1 accuracy and 95% confidence interval over 600 runs.

Implementation details. We use ResNet-10 as the backbone network [2, 9]. Our pretraining has two steps. In the first step, we train our network only on the mini-ImageNet dataset for 200 epochs. We use SGD with momentum 0.9, weight decay 1e-4, learning rate 0.01, batch size 32, and the cosine learning rate scheduler. In the next step, we use the mini-ImageNet-pretrained network, and use both the base dataset and the unlabeled dataset to optimize the loss function in Eq. 3 for 60 epochs. The sharpening temperature is set to 0.1. For the base images and weakly-augmented unlabeled images, we use the random-resize-crop, horizontal flip and normalization augmentations. For strong augmentation, we additionally use the color jitter, Gaussian blur, and random gray scale transformations. The other hyperparameters are kept the same. For few-shot evaluation, we learn a logistic regression classifier on the support set, and evaluate on the query set, similar to [2].

Results. Table 1 shows the performance comparison of our approach with other methods on the BSCD-FSL benchmark. All models are trained on the mini-ImageNet dataset. Performances for the meta-learning based methods are reported from [2]. "Transfer" denotes the baseline trained by cross-entropy loss on the base dataset. The scores for "SimCLR" are reported from [9], which is trained only on the unlabeled images. Only "STARTUP" [9] and our approach use the additional unlabeled dataset during the representation learning phase. "STARTUP (no SS)" denotes STARTUP without the contrastive loss, which is more comparable to our method. See [9] for details on STARTUP.

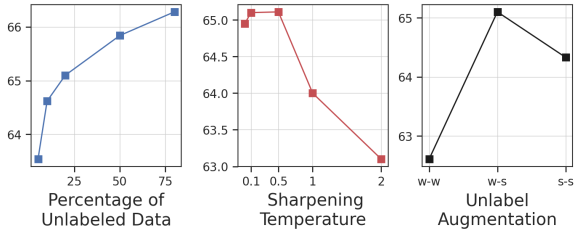
Our method outperforms all meta-learning-based approaches by a significant margin at all settings. Moreover, compared to Transfer, we achieve a 4.95% improvement for 5-shot classification on average. The performance improvement on 1-shot is more significant; we achieve 6.96% improvement on average.

We outperform STARTUP by 5.97% on EuroSAT, 2.76% on CropDisease, 0.88% on ChestX, and 1.04% on ISIC for 5-way 5-shot classification. The performance improvement is also quite significant for 1-shot classification; specifically, our method achieves 7.3% more accuracy than STARTUP on CropDisease. We only perform 0.24% worse on the ChestX dataset for the 1-shot. Considering that our method does not use any self-supervised training or distillation, the performance improvement is impressive. Note that STARTUP uses a fixed teacher to extract pseudo-labels for the unlabeled images, whereas we extract pseudo labels from the weakly-augmented images from the same network that is being trained. In that sense, our model works like a dynamic teacher, where the pseudo labels get more refined as training progress. We hypothesize that the superior performance might be attributed to the dynamic approach of our model over STARTUP.

4.1. Ablation Studies

We perform several ablation studies of different components of our approach. All scores are reported for 5-way 5-shot evaluation.

Percentage of unlabeled data. Fig. 2a shows the average 5-shot accuracy for different amounts of the unlabeled dataset during the pretraining phase. As expected, *more information from the unlabeled dataset helps to learn better representations on the target domain*. For example, the av-



(a) (b) (c)

Figure 2: Ablation studies for (a) amount of unlabeled data, (b) effect of sharpening temperature, and (c) data augmentation on unlabeled images. The Y-axis represents average top-1 accuracy (%) on the four benchmark datasets for 5-shot classification.

	EuroSAT	CropDisease	ChestX	ISIC
Ours	88.26±.46	95.78±.34	27.82±.44	48.26±.56
Ours(w/o base)	62.82±.76	49.87±.77	24.59±.42	35.14±.51
Ours(w/o stopgrad)	78.84±.65	85.76±.58	24.37±.41	44.35±.58
Ours(1-step)	86.16±.52	87.28±.49	25.11±.48	46.10±.59

Table 2: Ablation studies on different settings. Mean and 95% confidence interval over 600 runs.

erage accuracy increases by more than 1% if we use 80% of the dataset as the unlabeled set.

Is sharpening necessary? We perform ablation on the sharpening temperature in Fig. 2b. Note that $\tau = 1$ denotes no sharpening. The results suggest that sharpening increases the average accuracy by at least 1%.

Effect of data augmentation. On the unlabeled images, we apply two types of augmentation: weak augmentation to extract pseudo labels and strong augmentation to impose consistency regularization. This setting is denoted as "weak-strong" (w-s). We also show results with "weak-weak" (w-w) and "strong-strong" (s-s) augmentation settings in Figure 2c. The worst performing is the "w-w" setting. Imposing strong augmentation improves the accuracy.

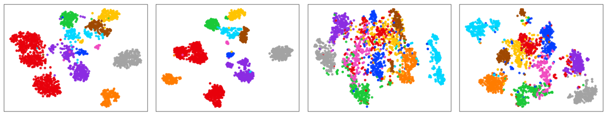
Effect of base dataset. We perform experiments without the first term in Eq. 3, i.e., we train the network on the base dataset first and then train only on the unlabeled images (without joint training on the base dataset), denoted as "Ours (w/o base)". Table 2 shows that the performance of "Ours (w/o base)" is poor, suggesting that the *representation related to the labeled base dataset is still helpful to the target domain.*

Importance of stop-gradient. In Table 2, "Ours (w/o stop-grad)" denotes the results without stop-gradient applied to the class score from the weakly-augmented unlabeled images. Without stop-gradient, the accuracy decreases.

Training without pretrained model on base dataset. We perform 2-step training during the representation learning phase - we first train the model on mini-IN only, and then jointly train on mini-IN and unlabeled images. In Table 2, "Ours(1-step)" denotes training on the unlabeled images and mini-IN from scratch, which performs worse than 2-step training. Our assumption is that the proposed model is also like self-training where a well-trained teacher is needed. Moreover, the mini-IN pretraining provides a good initialization for constructing pseudo-labels.

	EuroSAT	CropDisease	ChestX	ISIC
Ours-EuroSAT	88.26±.46	89.20±.56	25.11±.42	47.11±.61
Ours-CropDisease	82.04±.61	95.78±.34	25.63±.44	47.44±.61
Ours-ChestX	79.81±.62	89.51±.54	27.82±.44	44.17±.58
Ours-ISIC	80.71±.64	87.87±.60	26.58±.43	48.26±.56
Ours-all	83.87±.54	87.40±.56	28.71±.46	46.63±.59

Table 3: Effect of unlabeled datasets from a different domain than the target dataset. "Ours-X" denotes that we use base and "X" dataset during pretraining. "Ours-all" denotes that we use unlabeled images from all four target datasets.



(a) Transfer (b) Ours (c) Transfer (d) Ours

Figure 3: t-SNE plot of 10 classes from CropDisease (a & b) and EuroSAT (c & d) test sets with features obtained from Transfer and our method.

Different unlabeled data. Table 3 reports the few-shot accuracy when our model is trained on different unlabeled datasets. *The best accuracy is achieved when the unlabeled data and target data are from the same domain.* Even if the unlabeled data consists of images from multiple domains including the target domain (denoted as "Ours-all"), it still significantly under-performs the base model.

Quantitative analysis. Fig. 3 shows t-SNE plots [14] from 10 representative classes from the CropDisease and EuroSAT datasets. We compare the embeddings extracted from "Transfer" and our approach. We see that our method creates better grouping on the embeddings of the target datasets, even though we do not use any labels for the target dataset during pretraining.

5. Conclusion

We introduced a novel approach to utilize unlabeled data from the target domain for cross-domain few-shot learning. Experiments show that our method achieves state-of-the-art results in the BSCD-FSL benchmark for both 1-shot and 5-shot classification. Future work can be focused on applying our approach in each task during meta-testing so that the model can learn more category-specific representations.

6. Acknowledgments

This material is based upon work supported by the U.S. Department of Homeland Security under Award Number 2013-ST-061-ED0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

[1] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 3

[2] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. ECCV, 2020. 1, 2, 3

[3] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 2

[4] Taewon Jeong and Heeyoung Kim. Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[5] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2019. 1

[6] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019. 1

[7] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002*, 2018. 1

[8] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016. 2

[9] Cheng Perng Phoo and Bharath Hariharan. Self-training for few-shot transfer across extreme task differences. *arXiv preprint arXiv:2010.07734*, 2020. 1, 2, 3

[10] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018. 1

[11] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4077–4087. Curran Associates, Inc., 2017. 1

[12] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 2

[13] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 1

[14] LJP van der Maaten and GE Hinton. Visualizing high-dimensional data using t-sne.(2008). *Reference Source [Google Scholar]*, 2008. 4

[15] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3630–3638. Curran Associates, Inc., 2016. 1

[16] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 3

[17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2097–2106, 2017. 3

[18] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. *NeurIPS*, 2:8, 2018. 1