Uniform Priors for Meta-Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep Neural Networks have shown great promise on a variety of downstream applications; but their ability to adapt and generalize to new data and tasks remains a challenging problem. However, the ability to perform few-shot adaptation to novel tasks is important for the scalability and deployment of machine learn-ing models. It is therefore crucial to understand what makes for good, transferable features in deep networks that best allow for such adaptation. In this paper, we shed light on this by showing that features that are most transferable have high uniformity in the embedding space and propose a uniformity regularization scheme that encourages better transfer and feature reuse for few-shot learning. We evaluate our regularization on few-shot Meta-Learning benchmarks and show that *uniformity regularization* consistently offers benefits over baseline methods while also being able to achieve state-of-the-art on the Meta-Dataset.

1 Introduction

Deep Neural Networks have enabled great success in various machine learning domains such as computer vision [14, 16, 32], natural language processing [56, 9, 4], decision making [44, 45, 12] or in medical applications [39, 17]. This can be attributed to the ability of networks to extract abstract features from data, which, given sufficient data, can effectively generalize to held-out test sets.

However, the degree of generalization scales with the semantic difference between test and training tasks, caused e.g. by domain or distributional shifts between training and test data. Understanding how to achieve generalization under such shifts is an active area of research in few-shot Meta-Learning [48, 11, 6], where a meta-learner is tasked to quickly adapt to novel test data given its training experience and a limited labeled data budget. There exists a large corpus of meta-training methods that propose how to extract features from the training data. However, in this paper, we seek to investigate what fundamental properties learned features and feature spaces should have to facilitate adaptation in Meta-Learning.

Fortunately, recent literature provides pointers towards one such property: the notion of "feature uniformity" for improved generalization: For Unsupervised Representation Learning, [58] highlight a link between the uniform distribution of hyperspherical feature representations and the transfer performance in downstream tasks, which has been implicitly adapted in the design of modern contrastive learning methods [30, 1, 50, 51]. Similarly, [41] show that for Deep Metric Learning, uniformity in hyperspherical embedding space coverage as well as uniform singular value distribution embedding spaces are strongly connected to zero-shot generalization performance. Both [58] and [41] link the uniformity in the feature representation space to the preservation of maximal information and reduced overfitting. Thus, actively imposing a uniformity prior on learned feature representations should encourage better transfer properties by retaining more information and reducing bias towards training tasks, and as such facilitate better adaptation to novel tasks.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Motivated by this, we propose *uniformity regularization* for meta-learning deep neural networks, which places a uniform hypercube prior on the learned features space during training. Unlike e.g. a multivariate Gaussian, the *uniform* prior puts equal likelihood over the feature space, which then enables the network to make fewer assumptions about the data, limiting model overfitting to the training task. This incentivizes the model to learn more task-agnostic and reusable features, which in turn improve generalization [36]. Our *uniformity regularization* follows an adversarial learning framework that allows us to apply our proposed uniformity prior, as a uniform distribution does not have a closed-form divergence minimization scheme. Using this setup, we experimentally demonstrate that *uniformity regularization* aids test-time adaptation to novel tasks in few-shot Meta-Learning. We find that it consistently improves performance of baseline methods on four benchmarks, while also bein able to set a new state-of-the-art in Meta-Learning on the Meta-Dataset [54].

Overall, our contributions can be summarized as:

- We propose to perform *uniformity regularization* in the embedding spaces of a deep neural network, using a GAN-like alternating optimization scheme, to increase the transferability of learned features and the ability for better adaptation to novel tasks and data.
- Using our proposed regularization, we achieve strong improvements over baseline methods in Meta-Learning. Furthermore, *uniformity regularization* allows us to set a new state-of-the-art on the Meta-Dataset [54].

2 Background

2.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs, [15]) were proposed as a generative model which utilizes an alternative optimization scheme that solves a minimax two-player game between a generator, G, and a discriminator, D. The generator $G(z)$ is trained to map samples from a prior $z \sim p(z)$ to the target space, while the discriminator is trained to be an arbiter between the target data distribution $p(x)$ and the generator distribution. The generator is trained to *trick* the discriminator into predicting that samples from $G(z)$ actually stem from the target distribution. While many different GAN objectives have been proposed, the standard "Non-Saturating Cost" defines the generator objective

$$\mathcal{L}_G = \min_G \mathbb{E}_{z\sim p(z)}[1 - \log D(G(z))] \tag{2.1}$$

with discriminator objective

$$\mathcal{L}_D = \max_D \mathbb{E}_{z\sim p(z)}[1 - \log D(G(z))] + \mathbb{E}_{x\sim p(x)}[\log D(x)] \tag{2.2}$$

and $p(z)$ the generator prior and $p(x)$ a defined target distribution (e.g. natural images).

2.2 Fast Adaptation and Generalization in Meta-Learning

Throughout this work, we use the notion of "fast adaptation" to novel tasks to measure the transferability of learned features, and as such the generalization and adaptation capacities of a model. This term has recently been popularized by different meta-learning strategies [11, 48]. These methods assume distinct meta-training and meta-testing task distributions, where the goal of a meta-learner is to adapt fast to a novel task given limited samples for learning it.
Specifically, a few-shot meta-learner is evaluated to perform n-way classification given k 'shots', corresponding to k examples taken from n previously unseen classes. Generally, one distinguishes two types of meta-learners: ones requiring m training iterations for finetuning [11, 37], and ones that do not [48, 29]. In the meta-learning phase, the meta-learner is trained to solve entire tasks as (meta-training) datapoints. Its generalization is measured by how well it can quickly adapt to novel test tasks.
Many different strategies have been introduced to maximize the effectiveness of the meta-learning phase such as episodic training, where the model is trained by simulating 'test-like' conditions [57], or finetuning, where the model performs up to m gradient steps on the new task [11].

3 Extending Meta-Training with Uniformity Priors

In this section, we introduce the proposed *uniformity regularization* and detail the employed alternating GAN-like optimization scheme to perform it in a computationally tractable manner.

3.1 Prior Matching

Given a neural network $q(y|x)$ that is parameterized by θ we formally define the training objective as $\mathcal{L}_T(q(y|x), y)$ where $\mathcal{L}_T$ is any task-specific loss such as a cross-entropy loss, (x, y) are samples from the training distribution $\mathcal{D}_{\text{train}}$ and $q(y|x)$ the probability of predicting label y under q. This is a simplified formulation; in practice, there are many different ways to train a neural network, such as ranking-based training with tuples [7]. We define the embedding space z as the output of the final convolutional layer of a deep network. Accordingly, we'll note $q(z|x)$ as the conditional distribution for that embedding space which, due to the convnet being a deterministic mapping, is a dirac delta distribution at the value of the final convolutional layer. Section 4.1 further details how to apply *uniformity regularization* in practice.

As we ultimately seek to impose a uniformity prior over the learned aggregate feature/embedding "posterior" $q(z) = \int_x q(z|x)p(x)dx$, we begin by augmenting the generic task-objective to allow for the placement of a prior $r(z)$. For priors $r(z)$ with closed-form KL-divergences $\mathbf{D}$, one can define a prior-regularized task objective as

$$\mathcal{L} = \min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}} \left[\mathcal{L}_T(q(y|x), y)\right] + \mathbf{D}_{x\sim\mathcal{D}_{\text{train}}} \left(q(z|x) \| r(z)\right) \tag{3.1}$$

similar to the Variational Autoencoder formulation in [24]. However, to improve the generalization of a network by encouraging uniformity in the learned embeddings, we require regularization by matching the learned embedding space to a uniform distribution prior $\mathcal{U}(-\alpha, \beta)$, defined by the lower and upper bounds α and β, respectively. Unfortunately, such a regularization does not have a simple solution in practice, as a bounded uniform distribution has no closed-form KL divergence metric to minimize.

3.2 Uniformity Regularization

To address the practical limitation of solving Eqn. 3.1, we draw upon the GAN literature, in which alternate adversarial optimization has been successfully used to match a generated distribution to a defined target distribution using implicit divergence minimization. Latent variable models such as the Adversarial Autoencoder [33] have successfully used such a GAN-style adversarial loss, instead of a KL divergence, in the latent space of the autoencoder to learn a rich posterior. Such implicit divergence minimization allows us to match any well-defined distribution as a prior, but more specifically, ensures that we can successfully match learned embedding spaces to $\mathcal{U}(-\alpha, \alpha)$, which we set to the unit hypercube $\mathcal{U}(-1, 1)$ by default.

To this end, we adapt the GAN objective in Eqn. 2.2 and 2.1 for uniformity regularization optimization and train a discriminator, D, to be an arbiter between which samples are from the learned distribution $q(z|x)$ and from the uniform prior $r(z)$.
As such, the task model q (parameterized by θ) aims to *fool* the discriminator D into thinking that learned features, $q(z|x)$, come from the chosen uniform target distribution, $r(z)$, while the discriminator D learns to distinguish between learned features and samples taken from the prior, $\tilde{z} \sim r(z)$. Note that while the task-model defines a deterministic mapping for $q(z|x)$ instead of a stochastic one, the aggregate feature "posterior" $\int_x q(z|x)p(x)dx$, on which we apply our uniformity prior, is indeed a stochastic distribution [33].

Concretely for our *uniformity regularization*, we rewrite the discriminator objective from Eqn. 2.2 to account for the uniform prior matching, giving

$$\mathcal{L}_D = \max_D \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}[\log(1 - D(q(z|x)))] + \mathbb{E}_{\tilde{z}\sim\mathcal{U}(-1,1)}[\log D(\tilde{z})] \tag{3.2}$$

Consequently, we reformulate the generator objective from Eqn. 2.1 to reflect the task-model q,

Table 1: **Uniform Priors for Meta-Learning baselines**. Comparison of several meta-learning algorithms on four few-shot learning benchmarks: Omniglot [27], Double MNIST [28], CIFAR-FS [26] and Mini-Imagenet [57]. The models are evaluated with and without *uniformity regularization* ($\mathcal{UR}$) and we report the mean **error rate** over 5 seeds. No hyperparameter tuning is performed on the meta-learner and we use the exact hyperparameters as proposed in the original paper.

Baseline Study	Omniglot		Double MNIST		CIFAR-FS		MiniImageNet	
Methods ↓	(5, 1)	(5,5)	(5, 1)	(5,5)	(5, 1)	(5,5)	(5, 1)	(5,5)
Matching Networks [57]	2.1 ± 0.2	**1.0** ± 0.2	4.2 ± 0.2	**2.7** ± 0.2	46.7 ± 1.1	**62.9** ± 1.0	43.2 ± 0.3	50.3 ± 0.9
Matching Networks + $\mathcal{UR}$	**1.7**± 0.1	**0.9**± 0.1	**3.2**± 0.1	**2.3**± 0.3	**49.3** ± 0.4	**63.1** ± 0.7	**47.1** ± 0.8	**53.1** ± 0.7
MAML [11]	**4.8**± 0.4	**1.5** ± 0.4	**7.9**± 0.7	**1.9**± 0.3	**52.1** ± 0.8	**67.1** ± 0.9	47.2 ± 0.7	**62.1** ± 1.0
MAML + $\mathcal{UR}$	**4.1**± 0.5	**1.3**± 0.2	**7.3** ± 0.2	**1.5** ± 0.5	**52.9** ± 0.4	**67.1** ± 0.9	**48.9** ± 0.8	**64.1** ± 1.0
Prototypical Network [48]	**1.6** ± 0.2	**0.4** ± 0.1	**1.3** ± 0.2	**0.2** ± 0.2	**52.4** ± 0.7	**67.1** ± 0.5	45.4 ± 0.6	61.3 ± 0.7
Prototypical Network + $\mathcal{UR}$	**1.2** ± 0.3	**0.4** ± 0.1	**1.0** ± 0.2	**0.2** ± 0.2	**52.6** ± 0.8	**66.8** ± 0.5	**46.8** ± 0.5	**64.4** ± 0.9

$$\mathcal{L}_{\text{max}} = \min_{\theta} \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}[\log(1 - D(q(z|x)))] \tag{3.3}$$

where we used the notation $\mathcal{L}_{\text{max}}$ to reflect that optimization maximizes the feature uniformity by learning to fool D. Our final, *uniformity regularized* objective for θ is then given as

$$\mathcal{L} = \min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}}[\mathcal{L}_T(q(y|x), y)] + \gamma\mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}[\log(1 - D(q(z|x)))] \tag{3.4}$$

with task-objective $\mathcal{L}_T$ and training data distribution $\mathcal{D}_{\text{train}}$. Using this objective, the learned feature space is implicitly encouraged to become more uniform. The amount of regularization is controlled by the hyperparameter γ, balancing generalization of the model to new tasks and performance on the training task at hand. Large γ values hinder effective feature learning from training data, while values of γ too small result in weak regularization, leading to a non-uniform learned feature distribution with reduced generalization capabilities.

4 Experiments

We now study how *uniformity regularization* can facilitate generalizability of learned features and the ability of a model to perform fast adaptation to novel tasks and data in Meta-Learning. We divide this study into two experiments. First, we measure the improvements over distinct baseline methods on four Meta-Learning benchmarks in §4.2: Omniglot [27], Double MNIST [28], CIFAR-FS [26] and MiniImageNet [57]. To study more realistic applications, we then evaluate the benefits of *uniformity regularization* on the diverse, large-scale Meta-Dataset [54] in §4.3.

For all experiments, *we do not perform hyperparameter tuning on the base algorithms*, and use the same hyperparameters that the respective original papers proposed; we simply add the *uniformity regularization*, along with the task loss as in Eqn. 3.4.

4.1 Experimental Details

Uniformity regularization was added to the output of the CNNs for all networks. Specifically, the regularization is applied directly on the learned metric space for the metric-space based meta-learners [57, 48, 31], and applied to the output of the penultimate layer for MAML [11]. The discriminator is parameterized using a three-layer MLP with 100 hidden units in each layer and trained using the Adam optimizer [23] with a learning rate of 10^{-5}. The value of γ is chosen to be 0.1 for all experiments.

4.2 Uniform Priors improve Baseline Methods

We first examine the impact of *uniformity regularization* on three distinct meta-learning baselines: Matching Networks [57], Prototypical Networks [48] and MAML [11]. Performance is evaluated on

Table 2: **Uniform Priors achieve State-of-the-art on Meta-Dataset**. Application of *uniformity regularization* with Universal Representation Transformer Layers [31] on Meta-Dataset improves further upon the state-of-the-art performance of URT. Numbers listed in **blue** represent the state-of-the-art on the MetaDataset tasks.

Meta-Dataset (1/2)	ILSVRC	Omniglot	Aircrafts	Birds	Textures	QuickDraw
TaskNorm	50.6 ± 1.1	90.7 ± 0.6	83.8 ± 0.6	74.6 ± 0.8	62.1 ± 0.7	74.8 ± 0.7
SUR	56.3 ± 1.1	93.1 ± 0.5	85.4 ± 0.7	71.4 ± 1.0	71.5 ± 0.8	81.3 ± 0.8
SimpleCNAPS	**58.6** ± 1.1	91.7 ± 0.6	82.4 ± 0.7	74.9 ± 0.8	67.8 ± 0.8	77.7 ± 0.7
URT	55.7 ± 1.0	94.4 ± 0.4	85.8 ± 0.6	**76.3** ± 0.8	71.8 ± 0.7	82.5 ± 0.6
URT + $\mathcal{UR}$	**58.3** ± 0.9	**95.2** ± 0.2	**88.0** ± 0.9	**76.7** ± 0.8	**74.9** ± 0.9	**84.0** ± 0.3

Meta-Dataset (2/2)	Fungi	VGGFlower	TrafficSigns	MSCOCO	Average Rank
TaskNorm	48.7 ± 1.0	89.6 ± 0.6	67.0 ± 0.7	43.4 ± 1.0	4.5
SUR	**63.1** ± 1.0	82.8 ± 0.7	70.4 ± 0.8	52.4 ± 1.1	3.2
SimpleCNAPS	46.9 ± 1.0	**90.7** ± 0.5	**73.5** ± 0.7	46.2 ± 1.1	3.2
URT	**63.5** ± 1.0	88.2 ± 0.6	69.4 ± 0.8	52.2 ± 1.1	2.6
URT + $\mathcal{UR}$	**62.8** ± 1.1	**90.3** ± 0.4	**72.9** ± 0.8	**54.6** ± 1.1	**1.5**

four few-shot learning benchmarks: Double MNIST [28], Omniglot [27], CIFAR-FS [26] and Mini-Imagenet [57]. For our implementation, we utilize TorchMeta [8]. Results for each meta-learning method with and without regularization are summarized in Table 1a)[1].
As can be seen, the addition of *uniformity regularization* benefits generalization across method and benchmark, in some cases notably.
We find that this holds regardless of the number of shots used at meta-test-time, though we find the largest performance gains in the 1-shot scenario.
Overall, the results highlight the benefit of reduced training-task bias introduced by *uniformity regularization* for fast adaptation to novel test tasks.

4.3 Uniform Priors achieve State-of-the-art on Meta-Dataset

To measure the benefits for large-scale few-shot learning problems, we further examine *uniformity regularization* on the Meta-Dataset [54], which contains data from diverse domains such as natural images, objects and drawn characters. We follow the setup suggested by [54], used in [31], in which eight out of the ten available datasets are used for training, while evaluation is done over all. Results are averaged across varying numbers of ways and shots.
We apply *uniformity regularization* on the state-of-the-art Universal Representation Transformer (URT) [31], following their implementation and setup without hyperparameter tuning. As shown in Table 1b) *uniformity regularization* provides consistent improvements upon URT, matching or even outperforming the state-of-the-art on all sub-datasets.

5 Related Work

Adversarial Representation Learning. Latent variable models (e.g. Adversarial Autoencoders), have used GAN-style training [15] in the latent space [33, 53] to learn a rich posterior. Recent efforts have made such training effective in different contexts like active learning [47, 22] or domain adaptation [55, 18]. It has also found usage in Unsupervised representation learning (URL) [3, 2], ensemble-based representation learning [46, 34, 40] and continual learning [10]. In this work, we utilize adversarial training to introduce efficient *uniformity regularization* to improve fast adaptation and generalization of networks.

Meta-Learning. Many types of meta-learning algorithms for few-shot learning have recently been proposed such as memory-augmented methods [38, 35, 43], metric-based approaches [57, 48, 49] or optimization-based techniques [29, 11, 37, 59, 36]. More recently, finetuning using ImageNet [42] pretraining [6, 13] and episode-free few-shot approaches [52] have shed new light on alternative approaches. Different unsupervised approaches have also been used to learn such initializations

[1] For Double MNIST and Omniglot, error rates are listed instead of accuracies.

[5, 21]. Conversely, Meta-Learning has also been utilizes as a process of refinement for unsupervised representation [19]. Meta-learning has also been explored for fast adaptation of novel tasks in reinforcement learning [25, 60, 20].

Conclusion

In this paper, we propose a regularization technique for the challenging task of fast adaptation to novel tasks and data in neural networks. We present a simple and general solution, *uniformity regularization*, to reduce training bias and encourage networks to learn more reusable features. Over Meta-Learning baselines and benchmarks as well as the large-scale Meta-Dataset, we find improvements and even achieve state-of-the-art performance, highlighting the role of uniformity of the prior over learned features for generalization and adaptation.

References

[1] P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views, 2019.

[2] C. Beckham, S. Honari, V. Verma, A. M. Lamb, F. Ghadiri, R. D. Hjelm, Y. Bengio, and C. Pal. On adversarial mixup resynthesis. In *Advances in Neural Information Processing Systems*, pages 4348–4359, 2019.

[3] D. Berthelot, C. Raffel, A. Roy, and I. Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. *arXiv preprint arXiv:1807.07543*, 2018.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[5] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[6] Y. Chen, X. Wang, Z. Liu, H. Xu, and T. Darrell. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390*, 2020.

[7] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. pages 539–546. IEEE, 2005.

[8] T. Deleu, T. Würfl, M. Samiei, J. P. Cohen, and Y. Bengio. Torchmeta: A meta-learning library for pytorch. *arXiv preprint arXiv:1909.06576*, 2019.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] S. Ebrahimi, F. Meier, R. Calandra, T. Darrell, and M. Rohrbach. Adversarial continual learning. *arXiv preprint arXiv:2003.09553*, 2020.

[11] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[12] S. Fujimoto, H. Van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*, 2018.

[13] S. Gidaris and N. Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.

[14] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] M. H. Hesamian, W. Jia, X. He, and P. Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019.

[18] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[19] K. Hsu, S. Levine, and C. Finn. Unsupervised learning via meta-learning. *arXiv preprint arXiv:1810.02334*, 2018.

[20] A. Jabri, K. Hsu, A. Gupta, B. Eysenbach, S. Levine, and C. Finn. Unsupervised curricula for visual meta-reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 10519–10530, 2019.

[21] S. Khodadadeh, L. Boloni, and M. Shah. Unsupervised meta-learning for few-shot image classification. In *Advances in Neural Information Processing Systems*, pages 10132–10142, 2019.

[22] K. Kim, D. Park, K. I. Kim, and S. Y. Chun. Task-aware variational adversarial active learning. *arXiv preprint arXiv:2002.04709*, 2020.

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[25] L. Kirsch, S. van Steenkiste, and J. Schmidhuber. Improving generalization in meta reinforcement learning using learned objectives. *arXiv preprint arXiv:1910.04098*, 2019.

[26] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[27] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. The omniglot challenge: a 3-year progress report. *Current Opinion in Behavioral Sciences*, 29:97–104, 2019.

[28] Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[29] K. Lee, S. Maji, A. Ravichandran, and S. Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019.

[30] R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[31] L. Liu, W. Hamilton, G. Long, J. Jiang, and H. Larochelle. A universal representation transformer layer for few-shot image classification, 2020.

[32] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[33] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[34] T. Milbich, K. Roth, H. Bharadhwaj, S. Sinha, Y. Bengio, B. Ommer, and J. P. Cohen. Diva: Diverse visual feature aggregation fordeep metric learning. *arXiv preprint arXiv:2004.13458*, 2020.

[35] T. Munkhdalai, X. Yuan, S. Mehri, and A. Trischler. Rapid adaptation with conditionally shifted neurons. *arXiv preprint arXiv:1712.09926*, 2017.

[36] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

[37] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, pages 113–124, 2019.

[38] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. 2016.

[39] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[40] K. Roth, B. Brattoli, and B. Ommer. Mic: Mining interclass characteristics for improved metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8000–8009, 2019.

[41] K. Roth, T. Milbich, S. Sinha, P. Gupta, B. Ommer, and J. P. Cohen. Revisiting training strategies and generalization performance in deep metric learning. *arXiv preprint arXiv:2002.08473*, 2020.

[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[43] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016.

[44] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

[45] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[46] S. Sinha, H. Bharadhwaj, A. Goyal, H. Larochelle, A. Garg, and F. Shkurti. Dibs: Diversity inducing information bottleneck in model ensembles. *arXiv preprint arXiv:2003.04514*, 2020.

[47] S. Sinha, S. Ebrahimi, and T. Darrell. Variational adversarial active learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5972–5981, 2019.

[48] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[49] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. 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.

[50] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding, 2020.

[51] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning, 2020.

[52] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola. Rethinking few-shot image classification: a good embedding is all you need?, 2020.

[53] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[54] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019.

[55] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[57] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra. Matching networks for one shot learning. In *Advances in neural information processing systems*, 2016.

[58] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020.

[59] M. Yin, G. Tucker, M. Zhou, S. Levine, and C. Finn. Meta-learning without memorization. *arXiv preprint arXiv:1912.03820*, 2019.

[60] L. Zintgraf, K. Shiarlis, M. Igl, S. Schulze, Y. Gal, K. Hofmann, and S. Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.