https://doi.org/10.2352/ISSN.2470-1173.2021.6.IRIACV-313

EVALUATING DEEP SEMI-SUPERVISED LEARNING METHODS FOR COMPUTER VISION APPLICATIONS

Habib Ullah[1] *Mohib Ullah*[2] *Sultan Daud Khan*[3] *Faouzi Alaya Cheikh*[2]

[1] College of Computer Science and Engineering, University of Ha'il, Saudi Arabia.
[2] Department of Computer Science, Norwegian University of Science and Technology, Norway.
[3] National University of Technology (NUTech), Islamabad, Pakistan.

ABSTRACT

Deep semi-supervised learning (SSL) have been significantly investigated in the past few years due to its broad spectrum of theory, algorithms, and applications. The extensive use of the SSL methods is dominant in the field of computer vision, for example, image classification, human activity recognition, object detection, scene segmentation, and image generation. In spite of the significant success achieved in these domains, critically analyzing SSL methods on benchmark datasets still presents important challenges. In the literature, very limited reviews and surveys are available. In this paper, we present short but focused review about the most significant SSL methods. We analyze the basic theory of SSL and the differences among various SSL methods. Then, we present experimental analysis to compare these SSL methods using standard datasets. We also provide an insight into the challenges of the SSL methods.

Index Terms— Semi-supervised learning, unlabeled data, data augmentation, Mixmatch, Fixmatch.

1. INTRODUCTION

Deep neural networks have shown tremendous success in areas such as image [1–4] and speech recognition [5–7] by using a large collection of labeled data. In order to learn useful abstractions, deep neural networks set up millions of parameters, thus making them prone to over-fitting. Therefore, these achievements depend on collecting large datasets which typically require extensive human effort to manually label the datasets. The labeling processing may also requires pain and/or risk considering medical datasets [8, 9] driven by invasive tests. Moreover, labeling huge amount of data is expensive to label data requiring expert knowledge. It is worth noticing that for many practical applications, we do not have sufficient resources to collect a large labeled dataset [10, 11], which restricts the wide-spread use of deep learning techniques. An alternative and appealing way to cope with the lack of data is semi-supervised learning (SSL) approach. SSL methods are introduced to study the impact of using the labeled and unlabeled data together to improve performance. In comparison with supervised learning methods, which require labels for all samples, SSL methods can improve their performance by also taking into account the unlabeled samples as shown in Fig. 1. SSL methods also learn about the structure of the data from huge amount of unlabeled samples, therefore, alleviating the need for labeling all the available data for training. Currently, some state-of-the-art SSL methods reached the performance of solely supervised learning considering limited labeled data.

In the literature many SSL methods are proposed. For example, temporal ensembling [12] maintained an exponential moving average of label predictions on each training example, and penalized predictions that are inconsistent with this target. However, because the targets change only once per epoch, temporal ensembling becomes unwieldy when learning large datasets. Zhai et al. [13] unified semi-supervised learning and self-supervised learning to derive two novel semi-supervised image classification methods. Jeong et al. [14] proposed a consistency-based semi-supervised learning method for object Detection, which is a way of using consistency constraints as a tool for improving detection performance by making full use of available unlabeled data. Zhu et al. [15] proposed a semi-supervised deep learning method, using temporal ensembling of deep long short-term memory, to recognize human activities with smartphone inertial sensors. With the deep neural network processing, features are extracted for local dependencies in the recurrent framework. Chen et al. [16] introduced a novel memory-assisted deep neural network capable of using the memory of model learning to enable semi-supervised learning. Specifically, they introduced a memory mechanism into the network training process as an assimilation-accommodation interaction between the network and an external memory module.

We aim to provide the reader with an overview of the current state of the research area of semi-supervised learning and provide explanations of key algorithms and approaches. We present a perspective on semi-supervised learning that allows for a more thorough understanding of recent methods and the connections between them. In fact, we cannot possibly cover every method in existence, however, we present an overview

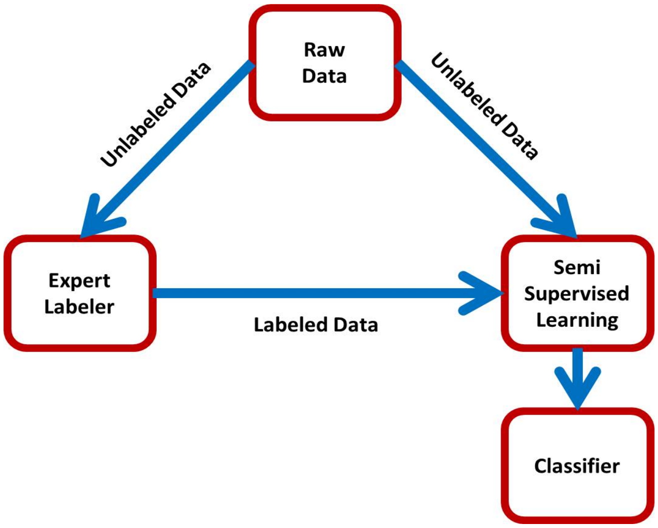

Fig. 1. In an SSL method, huge amount of unlabeled data is combined with limited amount of labeled data to improve the performance of the learning process.

of a set of SSL methods. Due to the sheer size of the literature considering SSL methods, this would not only be beyond the scope of this work, but also distract from the key insights which we wish to provide to the readers. Instead, we focus on the most recent advances in the area over the past few years. For this purpose, we identify five SSL methods gaining popularity in the research community. These methods are Π-model [17], the mean teacher model [18], the mixmatch model [19], the remixmatch model [20], and the fixmatch model [21].

The rest of the paper is structured as follows. The detail explanation of five semi-supervised learning methods are covered in Section 2. We present the experimental analysis and discussion in Section 3. Finally, in Section 4, we provide conclusion of the paper.

2. SEMI-SUPERVISED LEARNING METHODS

In this section, we provide technical details of each SSL method. In fact, each method is presented by taking into account different augmentation techniques and training strategies they used.

2.1. The Π-model

Rasmus et al. [17] proposed the Π-Model that combined the strengths of supervised learning with unsupervised learning in deep neural networks. They train their model to simultaneously reduce the sum of supervised and unsupervised costs by backpropagation, avoiding the need for layer-wise pre-training. The Π-Model is inspired by the Ladder network [22], where the supplementary objective is to denoise representations at every level of the model. The architecture of the model is an autoencoder with skip connections from the encoder to decoder. In fact, the learning task is similar to that in denoising autoencoders but applied to every layer, not just the inputs. The skip connections alleviate the obligation to represent details in the higher layers of the model because, through the skip connections, the decoder can retrieve any details thrown away by the encoder. Previously, the Ladder network was only considered in unsupervised learning. The key characteristics of the Π-Model are compatibility with supervised methods, scalability resulting from local learning, and computational efficiency.

For the decoder of fully connected networks, they exploited vertical mappings whose shape is a transpose of the encoder mapping. The same approach works for the convolution operations. The decoder part of the network used in the paper has convolution operations. The parametrization of these operations mirrors the encoder and effectively just reverses the flow of information. Convolutional networks use pooling operations which downsample the spatial feature maps. Therefore, the decoder entails to compensate for this with an equivalent upsampling procedure. To achieve this, firstly, on the encoder side, pooling operations are considered as separate layers with their own batch normalization and linear activation function. Secondly, the downsampling of the pooling on the encoder side is compensated for by upsampling with copying on the decoder side. They showed that a simultaneous unsupervised learning task improves both deep and shallow feedforward networks. The Π-Model is simple and easy to implement considering different feedforward architectures, as the training is based on backpropagation from a simple cost function.

2.2. The mean teacher model

Tarvainen and Valpola proposed the Mean Teacher Model [18] based on the fact that when a percept is changed slightly, a human typically still considers it to be the same object. Correspondingly, a classification model should support functions that give consistent output for similar data samples. One way to do this is to add noise to the input of the model. To consolidate the model to learn more abstract invariances, the noise may be added to intermediate representations, an insight that has motivated many regularization techniques. Rather than minimizing the classification cost at the zero-dimensional data samples of the input space, the regularized model minimizes the cost on a manifold around each data sample, thus pushing decision boundaries away from the labeled data samples.

Since the classification cost is not defined for unlabeled samples, the noise regularization by itself does not help in semi-supervised learning. To cope with this, the model [17] assesses each data sample with and without noise, and then applies a consistency cost between the two predictions. In this case, the model assumes a dual role as a teacher and a student. As a student, it learns as before; as a teacher, it produces targets, which are then utilized by itself as a student for learning. Since the model itself produces targets, they may not be very well correct. If too much weight is given to the produced targets, the cost of inconsistency outweighs that of misclassification, hindering the learning of new information. In effect, the model faces a confirmation bias, a problem that can be alleviated by improving the quality of targets. There are at least two options to enhance the target quality. One option is to identify the perturbation of the representations carefully instead of barely applying additive or multiplicative noise. Another option is to identify the teacher model carefully instead of barely replicating the student model. The Mean Teacher Model followed the second option and showed that it too provides significant benefits. In fact, they investigated a better teacher model from the student model without additional training.

2.3. The mixmatch model

Many recent approaches for semi-supervised learning add a loss term which is computed on unlabeled data and encourages the model to generalize better to unseen data. This loss term falls into one of three techniques namely entropy minimization (EM), consistency regularization (CR), and generic regularization (GR). Entropy minimization fosters the model to output confident predictions on unlabeled data, consistency regularization fosters the model to produce the same output distribution when its inputs are perturbed, and generic regularization fosters the model to generalize well and avoid overfitting the training data.

Berthelot et al. [19] introduced the mixmatch method, an SSL algorithm which introduces a single loss that effectively combines the three techniques to semi-supervised learning. Therefore, the mixmatch method introduces a unified loss term for unlabeled data that seamlessly reduces entropy while maintaining consistency and remaining compatible with traditional regularization techniques. In the mixmatch method, consistency regularization applies data augmentation (DA) to semi-supervised learning by leveraging the idea that a classifier should output the same class distribution for an unlabeled sample even after it has been augmented. The DA technique applies input transformations without affecting the semantics of the input data. For instance, in image analysis, the DA technique is applied by deforming or adding noise to an input image. This process changes the pixel content of an image without changing its label [23] [24] [25]. Therefore, the process extends the size of the training data by producing modified data. Entropy minimization encourages the fact that the classifier's decision boundary should not pass through high-density regions of the marginal data distribution. To achieve this the classifier should output low-entropy predictions on unlabeled data. The MixMatch method achieves entropy minimization through the use of a "sharpening" function on the target distribution for unlabeled data. Traditional regularization imposes a constraint on a model to make it harder to memorize the training data and therefore hopefully make it generalize better to unseen data [26]. In the mixmatch method, weight decay is used which penalizes the L2 norm of the model parameters [27]. The MixUp [28] technique is also used to encourage convex behavior "between" samples. MixUp is used both as a regularizer (applied to labeled samples) and a semi-supervised learning method (applied to unlabeled samples).

2.4. The remixmatch model

Berthelot et al. [20] improved the recently introduced the mixmatch method, a semi-supervised learning method by proposing two new techniques namely distribution alignment and augmentation anchoring. Distribution alignment nurtures the marginal distribution of predictions on unlabeled data to be close to the marginal distribution of ground truth labels. Augmentation anchoring feeds multiple strongly augmented versions of an input into the model and nurtures each output to be close to the prediction for a weakly-augmented version of the same input. To generate strong augmentations, the authors [20] introduced a variant of AutoAugment [24] which learns the augmentation policy while the model is being trained.

The distribution alignment technique encourages the distribution of a model's aggregated class predictions to match the marginal distribution of ground-truth class labels. Bridle et al. [29] introduced the distribution alignment technique a fair objective, where a related loss term was shown to arise from the maximization of mutual information between model inputs and outputs. Berthelot et al. [20] show how distribution alignment can be straightforwardly added to the mixmatch method by modifying the "guessed labels" using a running average of model predictions. In fact, the main objective of an SSL algorithm is to incorporate unlabeled data in a way which improves a model's performance. The authors of the remixmatch method formalize this intuition by maximizing the mutual information between the model's input and output for unlabeled data. A good classifier's prediction should depend as much as possible on the input. Taking into account this, the mixmatch method already considers a form of entropy minimization via the "sharpening" functions which makes the guessed labels (synthetic targets) for unlabeled data have lower entropy. Therefore, the remixmatch method also incorporates a form of "fairness" in the remixmatch method. For this purpose, over the course of training, the remixmatch method maintains a running average of the model's predictions on unlabeled data. Given the model's prediction on an unlabeled example, the remixmatch method scales the prediction by a ratio term and then renormalizes the result to form a valid probability distribution.

In the remixmatch method, augmentation anchoring replaces the consistency regularization component of the mixmatch method. For each given unlabeled input, augmentation anchoring first produces a weakly augmented version (e.g. using only a flip and a crop) and then produces multiple strongly augmented versions. The model's prediction for the weakly-augmented input is considered as the basis of the guessed label for all of the strongly augmented versions. To produce strong augmentations, the remixmatch method introduces a variant of AutoAugment [24] based on control theory which is called CTAugment. Unlike AutoAugment, CTAugment learns an augmentation policy alongside model training, making it particularly convenient in SSL settings.

2.5. The fixmatch model

Sohn et al. [21] followed the trend of recent state-of-the-art methods that unify different techniques for generating artificial labels [19] [30] [20] [31]. They proposed the fixmatch method, which generates artificial labels using both consistency regularization and pseudo-labeling. The artificial label is generated based on a weakly-augmented unlabeled image (e.g., using only flip-and-shift data augmentation). The weakly-augmented image is used as a target when the model is fed a strongly-augmented version of the same image. Inspired by UDA [30] and the remixmatch method [20], the fixmatch method leverages CutOut [32], CTAugment [20], and RandAugment [33] for strong augmentation, which all produce heavily distorted versions of a given image.

Overall, the fixmatch method is a simple combination of two common techniques to SSL namely consistency regularization and pseudo-labeling. The main contribution of the FixMatch method comes from the combination of these two ingredients as well as the use of a separate weak and strong augmentation when performing consistency regularization. The loss function for the fixmatch method consists exclusively of two cross-entropy loss terms. The first term is a supervised loss applied to labeled data and the second term is an unsupervised loss which is the standard cross-entropy loss on weakly augmented labeled samples. For unlabeled data, the fixmatch method estimates an artificial label for each sample which is then used in a standard cross-entropy loss. To obtain an artificial label, the fixmatch method first calculates the model's predicted class distribution given a weakly augmented version of a given unlabeled image. Considering it as a pseudo-label, except the fixmatch method enforces the cross-entropy loss against the model's output for a strongly-augmented version of unlabeled sample.

The fixmatch method leverages weak and strong kinds of augmentations. Weak augmentation is a standard flip-and-shift augmentation strategy. Images are randomly flipped horizontally and randomly translated vertically and horizontally. For strong augmentation, two approaches are considered based on AutoAugment [24] technique. The AutoAugment technique learns an augmentation strategy based on transformations using reinforcement learning. This requires labeled data to learn the augmentation pipeline, making it problematic to use in SSL settings where limited labeled data is available. Therefore, two variants of AutoAugment are considered in the fixmatch method namely RandAugment and CTAugment. In fact, the authors use Cutout followed by either of these augmentation strategies. The RandAugment technique randomly selects transformations for each sample in a minibatch from a collection of transformations including but not limited to color inversion, translation, and contrast adjustment. For this purpose, the RandAugment technique exploits a single fixed global magnitude that controls the severity of all distortions and the magnitude is a hyperparameter that must be optimized on a validation set.

3. EXPERIMENTAL ANALYSIS AND DISCUSSION

We perform experimental analysis on Π-Model [17] proposed the l, the mean teacher model [18], the mixmatch model [19], the remixmatch model [20], and the fixmatch model [21]. For this purpose, we consider standard SSL datasets namely CIFAR10/100 [34] and SVHN [35]. In particular, we carry out experiments with different amounts of labeled data and augmentation strategies on CIFAR10/100 [34] and SVHN [35]. To this end, we perform experiments with fewer labels than previously considered to investigate an SSL method showing promising results in extremely label-scarce settings. We compare the SSL methods using the same network architecture and training protocol, including the optimizer, learning rate schedule, and data preprocessing. Considering the work of Berthelot et al. [19], we use a Wide ResNet-28-2 [36] with 1.5M parameters for CIFAR-10 and SVHN, and WRN-28-8 for CIFAR-100. Higher performance with reduced magnitude of supervision is the main objective of SSL methods because it reduces the dependency on the labeled data. We also carry out experiments taking into account only four labeled samples for each class on each dataset.

We present the experimental results of the SSL methods in Table 1. We calculate the mean and variance of accuracy when training on 5 different "folds" of labeled data. The mixmatch method and the remixmatch method perform reasonably well with 40 and 250 labels, but we discover that the fixmatch model outperforms each of these methods significantly. For instance, the fixmatch model achieves an average error rate of 11.39% on CIFAR-10 with 4 labels per class. Moreover, the lowest error rate achieved on CIFAR-10 with 400 labels per class was 13.13%. Our experimental analysis also shows that the remixmatch method presents favorably good results. For example, considering the CIFAR-100 dataset, the remixmatch method performs better. The experimental results on a few variants of the fixmatch model are also presented. For this purpose, these variants copy different elements of the remixmatch method into the fixmatch method.

Table 1. Performance of the SSL methods [21]. We present error rates of the Π-model [17], the mean teacher model [18], the mixmatch model [19], the remixmatch model [20], and the fixmatch model [21] considering the CIFAR10/100 and SVHN datasets. These experimental results are presented using different number of labeled samples per class of each dataset.

Methods	CIFAR-10			CIFAR-100			SVHN		
	40	250	4000	400	2500	10000	40	250	1000
Π -Model	-	54.26±3.97	14.01±0.38	-	57.25±0.48	37.88±0.11	-	18.96±1.92	7.54±0.36
Mean teacher	-	32.32±2.30	9.19±0.19	-	53.91±0.57	35.83±0.24	-	3.57±0.11	3.42±0.07
Mixmatch	47.54 ±11.50	11.05±0.86	6.42±0.10	67.61±1.32	39.94±0.37	28.31±0.33	42.55±14.53	3.98±0.23	3.50±0.28
Remixmatch	**19.10**±9.64	**5.44**±0.05	4.72±0.13	**44.28**±2.08	**27.43**±0.31	**23.03**±0.56	**3.34**±0.20	**2.92**±0.48	2.65±0.08
Fixmatch (RA)	**13.81**±3.37	**5.07**±0.65	**4.26**±0.05	48.85±1.75	28.29±0.11	**22.60**±0.12	**3.96**±2.17	**2.48**±0.38	**2.28**±0.11
Fixmatch (CTA)	**11.39**±3.35	**5.07**±0.33	**4.31**±0.15	49.95±3.01	28.64±0.24	23.18±0.11	7.65±7.65	**2.64**±0.64	**2.36**±0.19

We investigate that the most important element is the distribution alignment (DA) term. In fact, the DA term encourages the model predictions to have the same class distribution as the labeled set. Putting together the fixmatch model with the DA term obtains a 40.14% error rate with 400 labeled samples. This is the significant improvement considering 44.28% achieved by the remixmatch method. As can be seen in Table 1, the fixmatch model presents similar performance in general using the augmentation techniques: CTAugment and RandAugment. However, this is not the case considering the setup of four labels per class. The reason may be that these scores are specifically high-variance. For instance, the variance over 5 various folds for CIFAR-10 with 4 labels per class is 3.35%, which is substantially higher than that with 25 labels per class (0.33%).

In general, these results show that the performances of the most recent SSL methods namely the mixmatch method, the remixmatch method, and the fixmatch model are very good in most cases. However, there are still some challenges when developing and using an SSL method in comparison with a purely supervised method. For example, one of the most significant problems to be addressed in SSL is the potential performance reduction due to the unlabeled data during the training process [37]. Although researchers have paid little attention to this in the literature, several SSL methods only present good results than their supervised counterparts or base methods in specific scenarios. In other cases, the supervised methods considered for experimentally assessing the performance of the SSL methods are relatively weak, presenting a skewed perspective on the advantages of using unlabelled data. Additionally, the potential performance reduction is generally much more substantial than the potential enhancement, especially in machine learning problems where strong performance is obtained with solely supervised learning.

4. CONCLUSION

In this paper we identify five potential methods in semi-supervised learning. We describe these methods in details that effectively leverage unlabeled data for training, and point out motivating advantages that arise if huge amount of unlabeled can be combined with limited amount of labeled data. We make a prospect into potential elements, application scenarios considering different datasets, and challenges that come along semi-supervised learning. We hope that this paper can lead to more attempts in more effective utilization of unlabeled data, and better learning methods.

In our future work, we aim to extend our experimental analysis taking into account more SSL methods and more challenging datasets.

5. REFERENCES

[1] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[2] Mohib Ullah, Mohammed Ahmed Kedir, and Faouzi Alaya Cheikh, "Hand-crafted vs deep features: A quantitative study of pedestrian appearance model," in *2018 Colour and Visual Computing Symposium (CVCS)*. IEEE, 2018, pp. 1–6.

[3] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[4] Habib Ullah, Ahmed B Altamimi, Muhammad Uzair, and Mohib Ullah, "Anomalous entities detection and localization in pedestrian flows," *Neurocomputing*, vol. 290, pp. 74–86, 2018.

[5] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang, "Disan: Directional self-attention network for rnn/cnn-free language understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.

[6] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6706–6713.

[7] Basemah Alshemali and Jugal Kalita, "Improving the reliability of deep neural networks in nlp: A review," *Knowledge-Based Systems*, vol. 191, pp. 105210, 2020.

[8] Claudia Steiner, Anne Elixhauser, and Jenny Schnaier, "The healthcare cost and utilization project: an overview.," *Effective Clinical Practice*, vol. 5, no. 3, 2002.

[9] Ahmed Kedir, Mohib Ullah, and Jacob Renzo Bauer, "Spectranet: A deep model for skin oxygenation measurement from multi-spectral data," *Electronic Imaging*, vol. 2020, no. 15, pp. 83–1, 2020.

[10] Bernardino Romera-Paredes and Philip Torr, "An embarrassingly simple approach to zero-shot learning," in *International conference on machine learning*. PMLR, 2015, pp. 2152–2161.

[11] Ahmed Mohammed, Congcong Wang, Meng Zhao, Mohib Ullah, Rabia Naseem, Hao Wang, Marius Pedersen, and Faouzi Alaya Cheikh, "Weakly-supervised network for detection of covid-19 in chest ct scans," *IEEE Access*, vol. 8, pp. 155987–156000, 2020.

[12] Samuli Matias Laine and Timo Oskari Aila, "Temporal ensembling for semi-supervised learning," 2018, US Patent App. 15/721,433.

[13] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer, "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 1476–1485.

[14] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak, "Consistency-based semi-supervised learning for object detection," in *Advances in neural information processing systems*, 2019, pp. 10759–10768.

[15] Qingchang Zhu, Zhenghua Chen, and Yeng Chai Soh, "A novel semisupervised deep learning method for human activity recognition," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 3821–3830, 2018.

[16] Yanbei Chen, Xiatian Zhu, and Shaogang Gong, "Semi-supervised deep learning with memory," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 268–283.

[17] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems*, 2015, pp. 3546–3554.

[18] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[19] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 5049–5059.

[20] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," *arXiv preprint arXiv:1911.09785*, 2019.

[21] 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.

[22] Harri Valpola, "From neural pca to deep unsupervised learning," in *Advances in independent component analysis and learning machines*, pp. 143–171. Elsevier, 2015.

[23] Dan Claudiu Cireşan, Ueli Meier, Luca Maria Gambardella, and Jürgen Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition," *Neural computation*, vol. 22, no. 12, pp. 3207–3220, 2010.

[24] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[25] Patrice Y Simard, David Steinkraus, John C Platt, et al., "Best practices for convolutional neural networks applied to visual document analysis.," in *Icdar*, 2003, vol. 3.

[26] GE Hinton and Drew van Camp, "Keeping neural networks simple by minimising the description length of weights. 1993," in *Proceedings of COLT-93*, pp. 5–13.

[27] I Loshchilov and F Hutter, "Fixing weight decay regularization in adam. arxiv 2017," *arXiv preprint arXiv:1711.05101*.

[28] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[29] John S Bridle, Anthony JR Heading, and David JC MacKay, "Unsupervised classifiers, mutual information and'phantom targets," in *Advances in neural information processing systems*, 1992, pp. 1096–1101.

[30] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[31] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[32] Terrance DeVries and Graham W Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[33] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703.

[34] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[35] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[36] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[37] Jesper E Van Engelen and Holger H Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.