MixMatch: A Holistic Approach toSemi-Supervised Learning (2019)

David Berthelot et al.

Notes

1 Introduction

In this paper the authors propose a new algorithm for semi supervised learning (SSL) called MixMatch that unifies the current dominant approaches in SSL with a single loss, this method gives state of the art results with very limited number of data and very useful for differentiall private learning.

The currect state of the art SSL technics are:

- Consistency Regularization: applies data augmentation to semi-supervised learning by leveraging the idea that a classifier should output the same class distribution for an unlabeled example even after it has been augmented. More formally, consistency regularization enforces that an unlabeled examplex should be classified the same as $Augment(x)$, where $Augment(x)$ is a stochastic data augmentation functionlike a random spatial translation or adding noise (e.g. Pi model where the loss is the difference between the model's predictions with two inputs, each one with a different data augmentation applied to it, and for Mean teacher, the second prediction is made by an exponontially moving average of the models weights).
- Entropy Minimization: an other way to enforce low density decision boundaries is to force the model to make confident predictions even if they're incorrect, this is done by adding a loss term that minimizes the entropy of the predictions $P(y|x)$.
- Traditional Regularization: when we add some contraint on the model to make it harder to memorize the trianing data, like L2-regularization, or mixup regulizers.

2 MixMatch

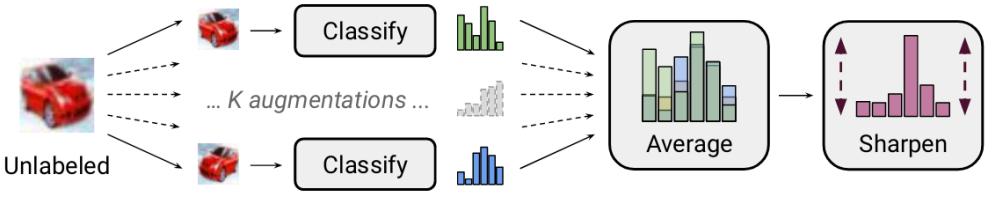

Figure 1: Diagram of the label guessing process used in MixMatch. Stochastic data augmentation is applied to an unlabeled image K times, and each augmented image is fed through the classifier. Then, the average of these K predictions is "sharpened" by adjusting the distribution's temperature. See algorithm 1 for a full description.

In a nutshell, MixMatch creates a batch of augmented examples $\mathcal{U}'$ and $\mathcal{X}'$, from a batch of labeled examples and $\mathcal{X}$ and unlabeled examples $\mathcal{U}$, these are then used to calculate a combined loss $\mathcal{U}$ for semi-supervised learning, with a supervised $\mathcal{L}_{\mathcal{X}}$ and unspervised loss $\mathcal{L}_{\mathcal{U}}$ and a weighting factor

$\lambda_{\mathcal{U}}$, the supervised loss is a standart cross entropy, and the unsupervised loss is an MSE loss, that corresponds to a multiclass brier loss.

$$
\begin{aligned}
\mathcal{X}', \mathcal{U}' &= \text{MixMatch}(\mathcal{X}, \mathcal{U}, T, K, \alpha) \\
\mathcal{L}_{\mathcal{X}} &= \frac{1}{|\mathcal{X}'|} \sum_{x,p \in \mathcal{X}'} \mathrm{H}\left(p, \mathrm{p}_{\text{ model}}\left(y|x;\theta\right)\right) \\
\mathcal{L}_{\mathcal{U}} &= \frac{1}{L\,|\mathcal{U}'|} \sum_{u,q \in \mathcal{X}'} \left\| q - \mathrm{p}_{\text{ model}}\left(y|u;\theta\right) \right\|_2^2 \\
\mathcal{L} &= \mathcal{L}_{\mathcal{X}} + \lambda_{\mathcal{U}} \mathcal{L}_{\mathcal{U}}
\end{aligned}
$$

Algorithm 1 MixMatch ingests a batch of labeled data $\mathcal{X}$ and a batch of unlabeled data $\mathcal{U}$ and produces a collection $\mathcal{X}'$ of processed labeled examples and a collection $\mathcal{U}'$ of processed unlabeled examples with "guessed" labels.

Input: Batch of labeled examples and their one-hot labels $\mathcal{X} = \big((x_b, p_b); b \in (1, \dots, B)\big)$, batch of unlabeled examples $\mathcal{U} = \big(u_b; b \in (1, \dots, B)\big)$, sharpening temperature T, number of augmentations K, Beta distribution parameter α for MixUp.
for $b = 1$ **to** B **do**
  $\hat{x}_b = \text{Augment}(x_b)$ *// Apply data augmentation to x_b*
  for $k = 1$ **to** K **do**
    $\hat{u}_{b,k} = \text{Augment}(u_b)$ *// Apply k^{th} round of data augmentation to u_b*
  end for
  $\bar{q}_b = \frac{1}{K} \sum_k \mathrm{p}_{\text{model}}(y \mid \hat{u}_{b,k}; \theta)$ *// Compute average predictions across all augmentations of u_b*
  $q_b = \text{Sharpen}(\bar{q}_b, T)$ *// Apply temperature sharpening to the average prediction (see eq. (7))*
end for
$\hat{\mathcal{X}} = \big((\hat{x}_b, p_b); b \in (1, \dots, B)\big)$ *// Augmented labeled examples and their labels*
$\hat{\mathcal{U}} = \big((\hat{u}_{b,k}, q_b); b \in (1, \dots, B), k \in (1, \dots, K)\big)$ *// Augmented unlabeled examples, guessed labels*
$\mathcal{W} = \text{Shuffle}\big(\text{Concat}(\hat{\mathcal{X}}, \hat{\mathcal{U}})\big)$ *// Combine and shuffle labeled and unlabeled data*
$\mathcal{X}' = \big(\text{MixUp}(\hat{\mathcal{X}}_i, \mathcal{W}_i); i \in (1, \dots, |\hat{\mathcal{X}}|)\big)$ *// Apply MixUp to labeled data and entries from $\mathcal{W}$*
$\mathcal{U}' = \big(\text{MixUp}(\hat{\mathcal{U}}_i, \mathcal{W}_{i+|\hat{\mathcal{X}}|}); i \in (1, \dots, |\hat{\mathcal{U}}|)\big)$ *// Apply MixUp to unlabeled data and the rest of $\mathcal{W}$*
return $\mathcal{X}', \mathcal{U}'$

The algorithm First with a given number of examples in a single batch, we apply some form of data augmentation to the labled example x_b obtaining $\hat{x}_b$, and for the unlabeled examples we apply k augmentation resulting in k unlabeled examples $\hat{u}_{b,k}$, we then create a label for unlabled example as the average of the model prediction given the k unlabeled augmented inputs:

$$
\bar{q}_b = \frac{1}{K} \sum_{k=1}^{K} \mathrm{p}_{\text{model}}\left(y|\hat{u}_{b,k};\theta\right)
$$

And to add some form of entropy minimization we apply a sharpening to our predictions, the closer T is to zero the sharper the predictions, for T = 0 we have a dirac / one-hot distribution:

$$
\text{Sharpen}(p, T)_i := p_i^{\frac{1}{T}} \Big/ \sum_{j=1}^{L} p_j^{\frac{1}{T}}
$$

Now that we have a label for the unlabeled examples (k examples), we can apply the mixup:

$$
\begin{aligned}
\lambda &\sim \text{Beta}(\alpha, \alpha) \\
\lambda' &= \max(\lambda, 1 - \lambda) \\
x' &= \lambda' x_1 + (1 - \lambda')\, x_2 \\
p' &= \lambda' p_1 + (1 - \lambda')\, p_2
\end{aligned}
$$

One difference from the traditionnal mixup is that we always choose λ so that the mixed examples are closed to the labeled examples $\mathcal{X}$, so first we have all our augmented labeled examples $\hat{\mathcal{X}} =$

$((\hat{x}_b, p_b)\,; b \in (1, \ldots, B))$, and all our unlabeled examples (each example has k augmented copies) with the labels we've created $\hat{\mathcal{U}} = ((\hat{u}_{b,k}, q_b)\,; b \in (1, \ldots, B), k \in (1, \ldots, K))$, we combine the two after shuffling to form $\mathcal{W}$, and for each ith labeled example we compute the MixUp $\left(\hat{\mathcal{X}}_i, \mathcal{W}_i\right)$ to get the augmented enteries $\mathcal{X}'$, now for the unlabeled examples we compute $\mathcal{U}'_i = \text{MixUp}\left(\hat{\mathcal{U}}_i, \mathcal{W}_{i+|\hat{\mathcal{X}}|}\right)$ using the elements of $\mathcal{X}$ we did not use in the first mixup, we the new sets $\mathcal{X}'$ and $\mathcal{U}'$ we can calculate the supervised and unsupervised losses.

3 Experiments

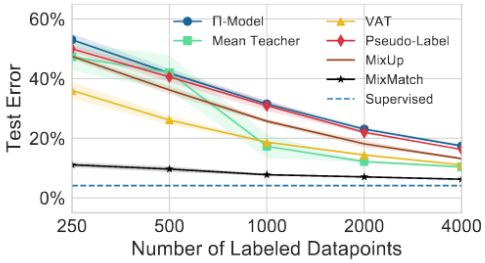

Figure 2: Error rate comparison of MixMatch to baseline methods on CIFAR-10 for a varying number of labels. Exact numbers are provided in table 5 (appendix). "Supervised" refers to training with all 50000 training examples and no unlabeled data. With 250 labels MixMatch reaches an error rate comparable to next-best method's performance with 4000 labels.

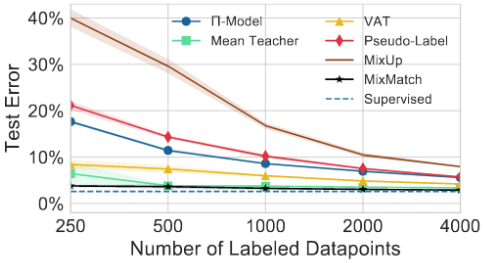

Figure 3: Error rate comparison of MixMatch to baseline methods on SVHN for a varying number of labels. Exact numbers are provided in table 6 (appendix). "Supervised" refers to training with all 73257 training examples and no unlabeled data. With 250 examples MixMatch nearly reaches the accuracy of supervised training for this model.

Method	CIFAR-10	CIFAR-100
Mean Teacher [42]	6.28	-
SWA [2]	5.00	28.80
MixMatch	4.95 ± 0.08	25.88 ± 0.30

Table 1: CIFAR-10 and CIFAR-100 error rate comparison with larger (26 million parameter) models.

Method	1000 labels	5000 labels
CutOut [12]	-	12.74
IIC [20]	-	11.20
SWWAE [47]	25.70	-
CC-GAN^2 [11]	22.20	-
MixMatch	10.18 ± 1.46	5.59

Table 2: STL-10 error rate using 1000-label splits or the entire 5000-label training set.

3.1 Ablation Study

Ablation	250 labels	4000 labels
MixMatch	11.80	6.00
MixMatch without distribution averaging ($K = 1$)	17.09	8.06
MixMatch without temperature sharpening ($T = 1$)	27.83	10.59
MixMatch with parameter EMA	11.86	6.47
MixMatch without MixUp	39.11	10.97
MixMatch with MixUp on labeled only	32.16	9.22
MixMatch with MixUp on unlabeled only	12.35	6.83
MixMatch with MixUp on separate labeled and unlabeled	12.26	6.50
Interpolation Consistency Training [44]	38.60	6.81

Table 4: Ablation study results. All values are error rates on CIFAR-10 with 250 or 4000 labels. ICT uses EMA parameters and unlabeled mixup and no sharpening.