MixMatch: A Holistic Approach to Semi-Supervised Learning

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, Colin Raffel

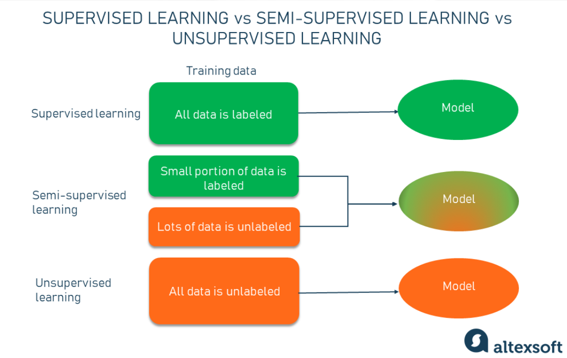
Coby Penso

What we'll see

1. Introduction
2. MixMatch
3. Experiments

Semi-Supervised Learning (SSL)

Semi-supervised learning is an approach that combines a small amount of labeled data with a large amount of unlabeled data during training.



Consistency Regularization

- data augmentations - applies input transformations assumed to leave class semantics unaffected
- **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 example x should be classified the same as Augment(x), an augmentation of itself.
- For unlabeled points x, Consistency Regularization loss term is

$$||p_{model}(y|Augment(x);\theta) - p_{model}(y|Augment(x);\theta)||_2^2$$

- Augment(x) - stochastic transformation
- **Drawback:** use domain specific data augmentations strategies.

Entropy Minimization

- SSL assumes that the classifier’s decision boundary should not pass through high-density regions of the marginal data distribution.
- One way to enforce - require that the classifier output low-entropy predictions on unlabeled data.
- Loss term is

$$p_{model}(y|x_{unlabeled}; \theta)$$

- "Pseudo-Label" does it implicitly.

Traditional Regularization

Imposing a constraint on a model to make it harder to memorize the training data and therefore hopefully make it generalize better to unseen data.

- Weight decay
- L2 Regularization
- MixUp -
  to encourage convex behaviour "between" examples.

MixMatch - Overview

- Batch *mathcalX* - labeled data.
- Batch $\mathcal{U}$ - equally-sized unlabeled data.
- Generate $\mathcal{X}'$ and $\mathcal{U}'$

$$\mathcal{X}', \mathcal{U}' = MixMatch(\mathcal{X}, \mathcal{U}, T, K, \alpha)$$

$$\mathcal{L}_{\mathcal{X}} = \frac{1}{|\mathcal{X}'|} \sum_{x,p \in \mathcal{X}'} H(p, p_{model}(y|x;\theta))$$

$$\mathcal{L}_{\mathcal{U}} = \frac{1}{L|\mathcal{U}'|} \sum_{u,q \in \mathcal{U}'} H(q, p_{model}(y|u;\theta))$$

$$\mathcal{L} = \mathcal{L}_{\mathcal{X}} + \lambda_{\mathcal{U}} \mathcal{L}_{\mathcal{U}}$$

- H(p,q) - cross-entropy.
- T,K,*alpha*, and $\lambda_{\mathcal{U}}$ are hyperparameters.

MixMatch - Data Augmentation

- Both on labeled and unlabeled data.
- For each x_b in $\mathcal{X}$: $\hat{x}_b = Augment(x_b)$.
- For each u_b in $\mathcal{U}$: $\hat{u}_{b,k} = Augment(u_b)$, $k \in (1, .., K)$

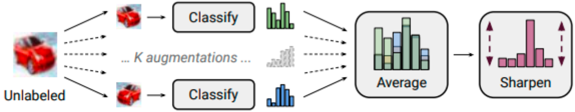


Figure 1: Diagram of the label guessing process used in MixMatch. Stochastic data augmentation is applied to an unlabeled image K times, and each augmented image is fed through the classifier. Then, the average of these K predictions is "sharpened" by adjusting the distribution's temperature. See algorithm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

MixMatch - Label Guessing

- For each unlabeled example in $\mathcal{U}$, MixMatch produces a "guess" label.
- Later on, used in the unsupervised loss term.
- $$\hat{q}_b = \frac{1}{K}\sum_{k=1}^{K} p_{model}(y|\hat{u}_{b,k};\theta)$$
- Apply a sharpening function to reduce the entropy of the label

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

Lower T -> lower-entropy predictions. distr

MixMatch - Mixup

- Modified version of MixUp

$$(x_1, p_1), (x_2, p_2) \rightarrow (x', p')$$

$$\lambda \sim Beta(\alpha, \alpha)$$

$$\lambda' = max(\lambda, 1 - \lambda)$$

$$x' = \lambda' x_1 + (1 - \lambda') x_2$$

$$p' = \lambda' p_1 + (1 - \lambda') p_2$$

  Vanilla MixUp uses $\lambda' = \lambda$
- $\lambda' = max(\lambda, 1 - \lambda)$ ensures that x' is closer to x_1.
- MixMatch transforms $\mathcal{X}$ into $\mathcal{X}'$, a collection of labeled examples which have had data augmentation and MixUp (potentially mixed with an unlabeled example) applied.
- Similarly, $\mathcal{U}$ is transformed into $\mathcal{U}'$, a collection of multiple augmentations of each unlabeled example with corresponding label guesses.

MixMatch - Algorithm

Algorithm 1 MixMatch takes a batch of labeled data $\mathcal{X}$ and a batch of unlabeled data $\mathcal{U}$ and produces a collection $\mathcal{X}'$ (resp. $\mathcal{U}'$) of processed labeled examples (resp. unlabeled 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**

  $\hat{x}_b = \mathrm{Augment}(x_b)$ *// Apply data augmentation to x_b*

  for $k = 1$ **to** K **do**

    $\hat{u}_{b,k} = \mathrm{Augment}(u_b)$ *// Apply k^{th} round of data augmentation to u_b*

  end for

  $\bar{q}_b = \frac{1}{K} \sum_k \mathrm{p}_{\mathrm{model}}(y \mid \hat{u}_{b,k}; \theta)$ *// Compute average predictions across all augmentations of u_b*

  $q_b = \mathrm{Sharpen}(\bar{q}_b, T)$ *// 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*

$\mathcal{W} = \mathrm{Shuffle}\big(\mathrm{Concat}(\hat{\mathcal{X}}, \hat{\mathcal{U}})\big)$ *// Combine and shuffle labeled and unlabeled data*

$\mathcal{X}' = \big(\mathrm{MixUp}(\hat{\mathcal{X}}_i, \mathcal{W}_i); i \in (1, \dots, |\hat{\mathcal{X}}|)\big)$ *// Apply MixUp to labeled data and entries from $\mathcal{W}$*

$\mathcal{U}' = \big(\mathrm{MixUp}(\hat{\mathcal{U}}_i, \mathcal{W}_{i+|\hat{\mathcal{X}}|}); i \in (1, \dots, |\hat{\mathcal{U}}|)\big)$ *// Apply MixUp to unlabeled data and the rest of $\mathcal{W}$*

return $\mathcal{X}', \mathcal{U}'$

Implementation details

- Model = Wide ResNet-28.
- Checkpoint every $2^1$6 training samples and report the median error rate of the last 20 checkpoints.
- 5000 examples to select hyperparameters.
- Weight decay of 0.0004.
- Datasets - CIFAR-10, CIFAR-100, STL-10, and SVHN.

Hyperparameters

- Sharpening temperature T.
- Number of unlabled augmentations K.
- α for the Beta distribution used in Mixup.
- Unsupervised loss weight $\lambda_{\mathcal{U}}$.
- They find in practice that most of MixMatch's hyperparameters can be fixed. Specifically, for all experiments set T = 0.5 and K = 2.
- $\alpha = 0.75$, $\lambda_{\mathcal{U}} = 100$ good starting point.
- Linearly ramp up $\lambda_{\mathcal{U}}$ to its maximum value over the first 16,000 steps of training.

SSL - CIFAR10/100

Method	CIFAR-10	CIFAR-100
Mean Teacher [44]	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 (with 4,000 and 10,000 labels respectively) with larger models (26 million parameters).

Method	1000 labels	5000 labels
CutOut [12]	-	12.74
IIC [20]	-	11.20
SWWAE [48]	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.

Labels	250	500	1000	2000	4000	All
SVHN	3.78 ± 0.26	3.64 ± 0.46	3.27 ± 0.31	3.04 ± 0.13	2.89 ± 0.06	2.59
SVHN+Extra	2.22 ± 0.08	2.17 ± 0.07	2.18 ± 0.06	2.12 ± 0.03	2.07 ± 0.05	1.71

SSL - SVHN

Methods/Labels	250	500	1000	2000	4000
PiModel	17.65 ± 0.27	11.44 ± 0.39	8.60 ± 0.18	6.94 ± 0.27	5.57 ± 0.14
PseudoLabel	21.16 ± 0.88	14.35 ± 0.37	10.19 ± 0.41	7.54 ± 0.27	5.71 ± 0.07
Mixup	39.97 ± 1.89	29.62 ± 1.54	16.79 ± 0.63	10.47 ± 0.48	7.96 ± 0.14
VAT	8.41 ± 1.01	7.44 ± 0.79	5.98 ± 0.21	4.85 ± 0.23	4.20 ± 0.15
MeanTeacher	6.45 ± 2.43	3.82 ± 0.17	3.75 ± 0.10	3.51 ± 0.09	3.39 ± 0.11
MixMatch	3.78 ± 0.26	3.64 ± 0.46	3.27 ± 0.31	3.04 ± 0.13	2.89 ± 0.06

Table 6: Error rate (%) for SVHN.

Methods/Labels	250	500	1000	2000	4000
PiModel	13.71 ± 0.32	10.78 ± 0.59	8.81 ± 0.33	7.07 ± 0.19	5.70 ± 0.13
PseudoLabel	17.71 ± 0.78	12.58 ± 0.59	9.28 ± 0.38	7.20 ± 0.18	5.56 ± 0.27
Mixup	33.03 ± 1.29	24.52 ± 0.59	14.05 ± 0.79	9.06 ± 0.55	7.27 ± 0.12
VAT	7.44 ± 1.38	7.37 ± 0.82	6.15 ± 0.53	4.99 ± 0.30	4.27 ± 0.30
MeanTeacher	2.77 ± 0.10	2.75 ± 0.07	2.69 ± 0.08	2.60 ± 0.04	2.54 ± 0.03
MixMatch	2.22 ± 0.08	2.17 ± 0.07	2.18 ± 0.06	2.12 ± 0.03	2.07 ± 0.05

Table 7: Error rate (%) for SVHN+Extra.

Unlabeled dataset size

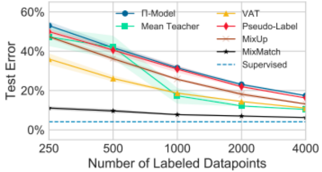


Figure 2: Error rate comparison of MixMatch to baseline methods on CIFAR-10 for a varying number of labels. Exact numbers are provided in table 5 (appendix). "Supervised" refers to training with all 50000 training examples and no unlabeled data. With 250 labels MixMatch reaches an error rate comparable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

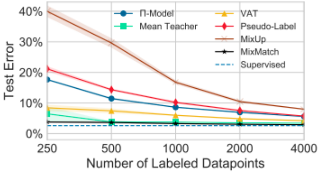


Figure 3: Error rate comparison of MixMatch to baseline methods on SVHN for a varying number of labels. Exact numbers are provided in table 6 (appendix). "Supervised" refers to training with all 73257 training examples and no unlabeled data. With 250 examples MixMatch nearly reaches the accuracy of supervised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

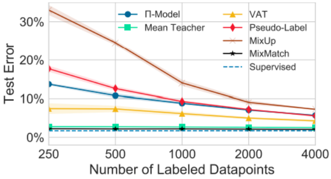


Figure 4: Error rate comparison of MixMatch to baseline methods on SVHN+Extra for a varying number of labels. With 250 examples we reach nearly the state of the art compared to supervised training for this model.

Ablation

Ablation	250 labels	4000 labels
MixMatch	11.80	6.00
MixMatch without distribution averaging ($K = 1$)	17.09	8.06
MixMatch with $K = 3$	11.55	6.23
MixMatch with $K = 4$	12.45	5.88
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 [45]	38.60	6.81

Table 4: Ablation study results. All values are error rates on CIFAR-10 with 250 or 4000 labels.

Questions?

Bibliography

- MixMatch: A Holistic Approach to Semi-Supervised Learning
  https://arxiv.org/pdf/1905.02249.pdf
- mixup: Beyond Empirical Risk Minimization
  https://arxiv.org/abs/1710.09412

- Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks - https://www.researchgate.net/publication/280581078$_{P}seudo - Label_{T}he_{S}imple_{a}nd_{E}fficient_{S}emi - Supervised_{L}earning_{M}ethod_{f}or_{D}eep_{N}eural_{N}etworks$