FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence[1]

February 6, 2020

[1] *FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence* by Sohn et al.

Overview

FixMatch:

Overview

FixMatch:

- is a strong Self-Supervised Learning approach

Overview

FixMatch:

- is a strong Self-Supervised Learning approach
- has a very simple loss function which has an interesting interpretation

Overview

FixMatch:

- is a strong Self-Supervised Learning approach
- has a very simple loss function which has an interesting interpretation
- has quite complex augmentation strategy

Overview

FixMatch:

- is a strong Self-Supervised Learning approach
- has a very simple loss function which has an interesting interpretation
- has quite complex augmentation strategy
- achieves 88.61% accuracy on CIFAR-10 with 4 labels per class!

Notation

Notation

- Imagine we have two datasets of images:

Notation

- Imagine we have two datasets of images:
  - Labeled images $X^l = \left\{ \left(x_n^l, p_n \right) \right\}_{n=1}^{N}$

Notation

- Imagine we have two datasets of images:
  - Labeled images $X^l = \left\{ \left(x_n^l, p_n \right) \right\}_{n=1}^{N}$
  - Unlabeled images $X^u = \{x_n^u\}_{n=1}^{k \times N}$, i.e. X^u is k times larger than X^l

Notation

- Imagine we have two datasets of images:
  - Labeled images $X^l = \left\{\left(x_n^l, p_n\right)\right\}_{n=1}^{N}$
  - Unlabeled images $X^u = \{x_n^u\}_{n=1}^{k \times N}$, i.e. X^u is k times larger than X^l
  - p_n is a one-hot class label

Notation

- Imagine we have two datasets of images:
  - Labeled images $X^l = \left\{ \left(x_n^l, p_n \right) \right\}_{n=1}^{N}$
  - Unlabeled images $X^u = \{x_n^u\}_{n=1}^{k \times N}$, i.e. X^u is k times larger than X^l
  - p_n is a one-hot class label
- Let $H(p, q)$ be a cross-entropy between p and q.

Notation

- Imagine we have two datasets of images:
  - Labeled images $X^l = \left\{\left(x_n^l, p_n\right)\right\}_{n=1}^N$
  - Unlabeled images $X^u = \{x_n^u\}_{n=1}^{k \times N}$, i.e. X^u is k times larger than X^l
  - p_n is a one-hot class label
- Let $H(p, q)$ be a cross-entropy between p and q.
- We also have augmentation functions:

Notation

- Imagine we have two datasets of images:
  - Labeled images $X^l = \left\{\left(x_n^l, p_n\right)\right\}_{n=1}^{N}$
  - Unlabeled images $X^u = \{x_n^u\}_{n=1}^{k \times N}$, i.e. X^u is k times larger than X^l
  - p_n is a one-hot class label
- Let $H(p, q)$ be a cross-entropy between p and q.
- We also have augmentation functions:
  - $\alpha(x)$ is a weak (i.e. simple) augmentation: random horizontal flipping and translations

Notation

- Imagine we have two datasets of images:
  - Labeled images $X^l = \left\{\left(x_n^l, p_n\right)\right\}_{n=1}^{N}$
  - Unlabeled images $X^u = \{x_n^u\}_{n=1}^{k\times N}$, i.e. X^u is k times larger than X^l
  - p_n is a one-hot class label
- Let $H(p, q)$ be a cross-entropy between p and q.
- We also have augmentation functions:
  - $\alpha(x)$ is a weak (i.e. simple) augmentation: random horizontal flipping and translations
  - $\mathcal{A}(x)$ is a strong (i.e. sophisticated) augmentation: color inversion, translation, contrast adjustment, etc

FixMatch

FixMatch

FixMatch loss is a combination of two losses:

$$\mathcal{L}_{\mathsf{FM}} = \mathcal{L}_{\mathsf{cls}} + \lambda_{\mathsf{pl}} \mathcal{L}_{\mathsf{pl}} \tag{1}$$

FixMatch

FixMatch loss is a combination of two losses:

$$\mathcal{L}_{\mathsf{FM}} = \mathcal{L}_{\mathsf{cls}} + \lambda_{\mathsf{pl}}\mathcal{L}_{\mathsf{pl}} \tag{1}$$

- $\mathcal{L}_{\mathsf{cls}}$ is a usual cross-entropy classification loss on X^l dataset.

FixMatch

FixMatch loss is a combination of two losses:

$$\mathcal{L}_{\mathrm{FM}} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{pl}} \mathcal{L}_{\mathrm{pl}} \tag{1}$$

- $\mathcal{L}_{\mathrm{cls}}$ is a usual cross-entropy classification loss on X^l dataset.
- $\mathcal{L}_{\mathrm{pl}}$ is a *pseudo-labeling loss*, i.e. a cross-entropy loss that uses synthetic targets produced by our model

Pseudo-labeling loss

Pseudo-labeling loss

Pseudo-labeling loss $\mathcal{L}_{\mathrm{pl}}$ equals:

$$\mathcal{L}_{\mathrm{pl}} = \frac{1}{kN} \sum_{n=1}^{kN} 1\left[\max\left(\bar{q}_n\right) \geq \tau\right] H\left(\bar{q}'_n, p_m\left(y | \mathcal{A}\left(x_n\right)\right)\right) \tag{2}$$

Where:

- $\bar{q}_n = p_m(y|\alpha(x_n))$, i.e. class probabilities for weakly augmented x_n
- $\bar{q}'_n = \arg\max \bar{q}_n$, i.e. a pseudo label (one-hot)
- τ is a hyperparameter

Pseudo-labeling loss

Pseudo-labeling loss $\mathcal{L}_{\text{pl}}$ equals:

$$\mathcal{L}_{\text{pl}} = \frac{1}{kN} \sum_{n=1}^{kN} 1\left[\max\left(\bar{q}_n\right) \geq \tau\right] H\left(\bar{q}'_n, p_m\left(y|\mathcal{A}\left(x_n\right)\right)\right) \quad (2)$$

Where:

- $\bar{q}_n = p_m(y|\alpha(x_n))$, i.e. class probabilities for weakly augmented x_n
- $\bar{q}'_n = \arg\max \bar{q}_n$, i.e. a pseudo label (one-hot)
- τ is a hyperparameter

Algorithm:

Pseudo-labeling loss

Pseudo-labeling loss $\mathcal{L}_{\text{pl}}$ equals:

$$\mathcal{L}_{\text{pl}} = \frac{1}{kN} \sum_{n=1}^{kN} 1\left[\max\left(\bar{q}_n\right) \geq \tau\right] H\left(\bar{q}'_n, p_m\left(y | \mathcal{A}\left(x_n\right)\right)\right) \tag{2}$$

Where:

- $\bar{q}_n = p_m(y|\alpha(x_n))$, i.e. class probabilities for weakly augmented x_n
- $\bar{q}'_n = \arg\max \bar{q}_n$, i.e. a pseudo label (one-hot)
- τ is a hyperparameter

Algorithm:

- Pick an unlabeled image x_n, produce two augmentated versions: $\bar{x}_n = \alpha(x)_n$ and $\tilde{x}_n = \mathcal{A}(x_n)$

Pseudo-labeling loss

Pseudo-labeling loss $\mathcal{L}_{\mathrm{pl}}$ equals:

$$\mathcal{L}_{\mathrm{pl}} = \frac{1}{kN} \sum_{n=1}^{kN} 1\left[\max\left(\bar{q}_n\right) \geq \tau\right] H\left(\bar{q}'_n, p_m\left(y|\mathcal{A}\left(x_n\right)\right)\right) \tag{2}$$

Where:

- $\bar{q}_n = p_m(y|\alpha(x_n))$, i.e. class probabilities for weakly augmented x_n
- $\bar{q}'_n = \arg\max \bar{q}_n$, i.e. a pseudo label (one-hot)
- τ is a hyperparameter

Algorithm:

- Pick an unlabeled image x_n, produce two augmentated versions: $\bar{x}_n = \alpha(x)_n$ and $\tilde{x}_n = \mathcal{A}(x_n)$
- Compute class probabilities $\bar{q}_n$ and $\tilde{q}_n$ for $\bar{x}_n$ and $\tilde{x}_n$

Pseudo-labeling loss

Pseudo-labeling loss $\mathcal{L}_{\text{pl}}$ equals:

$$\mathcal{L}_{\text{pl}} = \frac{1}{kN} \sum_{n=1}^{kN} 1 \left[\max\left(\bar{q}_n\right) \geq \tau\right] H\left(\bar{q}'_n, p_m\left(y|\mathcal{A}\left(x_n\right)\right)\right) \tag{2}$$

Where:

- $\bar{q}_n = p_m(y|\alpha(x_n))$, i.e. class probabilities for weakly augmented x_n
- $\bar{q}'_n = \arg\max \bar{q}_n$, i.e. a pseudo label (one-hot)
- τ is a hyperparameter

Algorithm:

- Pick an unlabeled image x_n, produce two augmentated versions: $\bar{x}_n = \alpha(x)_n$ and $\tilde{x}_n = \mathcal{A}(x_n)$
- Compute class probabilities $\bar{q}_n$ and $\tilde{q}_n$ for $\bar{x}_n$ and $\tilde{x}_n$
- Pick only examples with confident class probabilities for weakly-augmented images

Pseudo-labeling loss

Pseudo-labeling loss $\mathcal{L}_{\mathrm{pl}}$ equals:

$$\mathcal{L}_{\mathrm{pl}} = \frac{1}{kN} \sum_{n=1}^{kN} 1\left[\max\left(\bar{q}_n\right) \geq \tau\right] H\left(\bar{q}'_n, p_m\left(y|\mathcal{A}\left(x_n\right)\right)\right) \tag{2}$$

Where:

- $\bar{q}_n = p_m(y|\alpha(x_n))$, i.e. class probabilities for weakly augmented x_n
- $\bar{q}'_n = \arg\max \bar{q}_n$, i.e. a pseudo label (one-hot)
- τ is a hyperparameter

Algorithm:

- Pick an unlabeled image x_n, produce two augmentated versions: $\bar{x}_n = \alpha(x)_n$ and $\tilde{x}_n = \mathcal{A}(x_n)$
- Compute class probabilities $\bar{q}_n$ and $\tilde{q}_n$ for $\bar{x}_n$ and $\tilde{x}_n$
- Pick only examples with confident class probabilities for weakly-augmented images
- Cross-entropy term forces the model to give the same predictions for a weakly-augmented and a strongly-augmented images

An interesting property of this pseudo-labeling loss

An interesting property of this pseudo-labeling loss

- Previously used variants of pseudo-labelling loss required tuning of λ_{pl} weight during training and gradually increase it.

An interesting property of this pseudo-labeling loss

- Previously used variants of pseudo-labelling loss required tuning of λ_{pl} weight during training and gradually increase it.
- In FixMatch model becomes gradually more confident in new images and authors omit λ_{pl} whatsoever!

An interesting property of this pseudo-labeling loss

- Previously used variants of pseudo-labelling loss required tuning of λ_{pl} weight during training and gradually increase it.
- In FixMatch model becomes gradually more confident in new images and authors omit λ_{pl} whatsoever!
- So we get a curriculum learning out of the box!

Notes

Notes

- FixMatch provides very strong scores

Notes

- FixMatch provides very strong scores
- Has a very simple loss with an interesting side-effect of curriculum learning

Notes

- FixMatch provides very strong scores
- Has a very simple loss with an interesting side-effect of curriculum learning
- A disadvantage: strong augmentations are based on CutOut, CTAugment, etc and seem *very* sophisticated

Notes

- FixMatch provides very strong scores
- Has a very simple loss with an interesting side-effect of curriculum learning
- A disadvantage: strong augmentations are based on CutOut, CTAugment, etc and seem *very* sophisticated
- Strongly beats SotA in many setups:

	CIFAR-10			CIFAR-100			SVHN		
Method	40 labels	250 labels	4000 labels	400 labels	2500 labels	10000 labels	40 labels	250 labels	1000 labels
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
Pseudo-Labeling	-	49.78±0.43	16.09±0.28	-	57.38±0.46	36.21±0.19	-	20.21±1.09	9.94±0.61
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
UDA	29.05±5.93	8.82±1.08	4.88±0.18	59.28±0.88	33.13±0.22	24.50±0.25	52.63±20.51	5.69±2.76	**2.46**±0.24
ReMixMatch	**19.10**±9.64	**5.44**±0.05	4.72±0.13	**44.28**±2.06	**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