MixMatch: A Holistic Approach to Semi-Supervised Learning

Desheulin Oleg

Higher School of Economics

https://arxiv.org/abs/1905.02249

21 ноября 2019 г.

Semi-Supervised Learning

- Часто сложно собрать большой полностью размеченный датасет, это дорого, требует экспертов, сами по себе данные могут содержать конфиденциальную информацию
- Поэтому возникает задача:
  X - размеченные данные, U - неразмеченные
  Необходимо обучить модель, которая будет использовать для обучения X и U, повышая свою обобщающую способность с помощью дополнительных данных.

Что делать с нерамеченными данными?

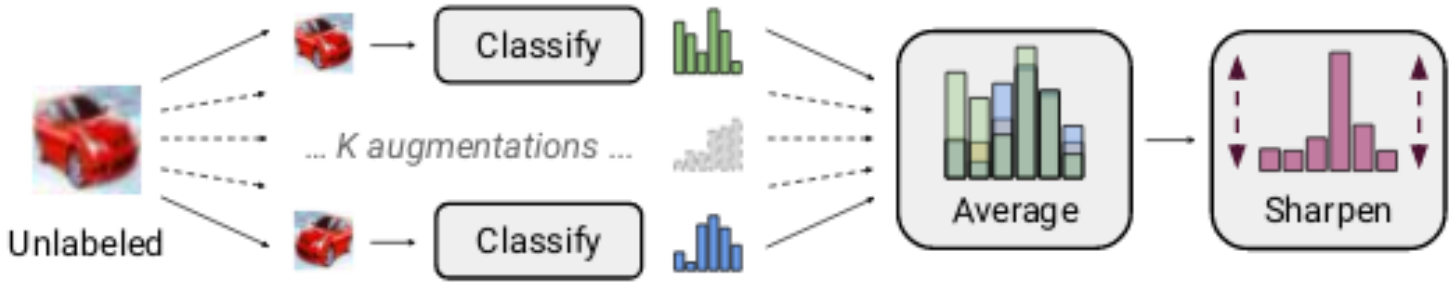


Figure 1: Diagram of the label guessing process used in MixMatch. Stochastic data augmentation is applied to an unlabeled image K times, and each augmented image is fed through the classifier. Then, the average of these K predictions is "sharpened" by adjusting the distribution's temperature. See algorithm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

Что делать с нерамеченными данными?

В формулах:

- Пусть K отвечает за количество аугментаций, тогда:

$$q_i = \frac{1}{K} \sum_{k=1}^{K} p(y|Augment_k(u_i), \theta)$$

- Sharpening:

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

Функции потерь

Пусть L_x отвечает за размеченную выборку, а L_u за неразмеченную. Тогда введем такие функции потерь:

$$L_x = \frac{1}{\|X\|} \sum_{x,y} H(y, p(y|x;\theta))$$

$$L_u = \frac{1}{\|Y\|\|U\|} \sum_{u_i,q_i} \|q_i - p(y|x;\theta)\|_2^2$$

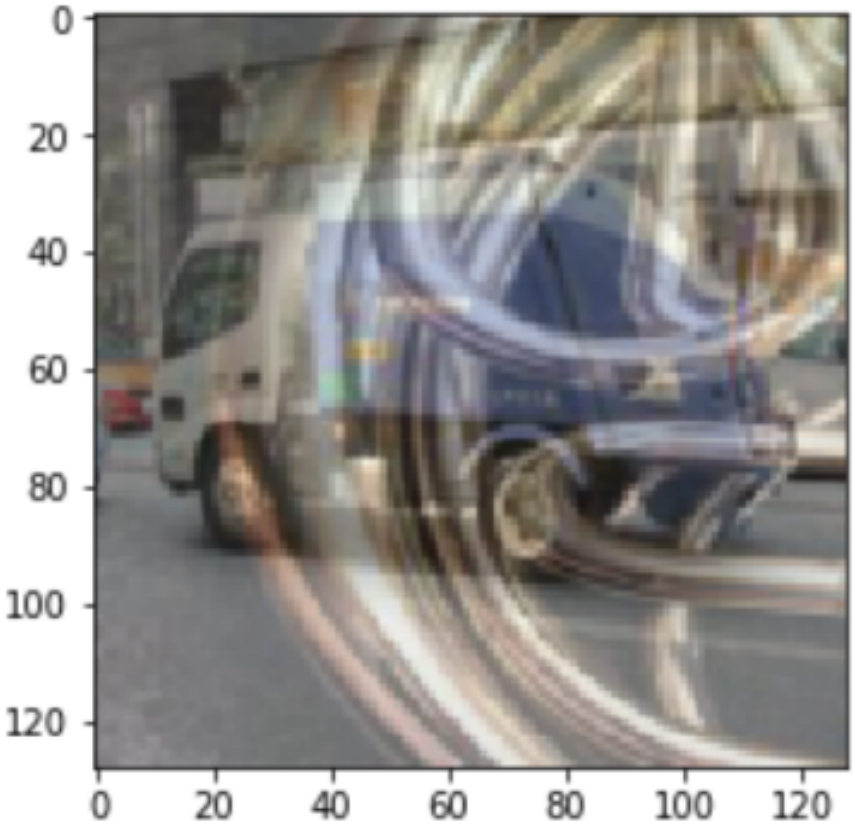
MixUp

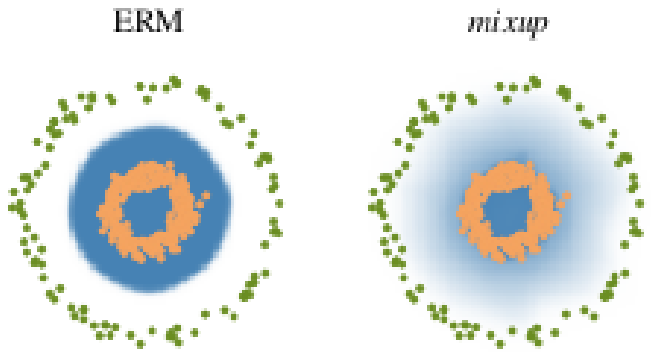
Пусть $(x_1, p_1); (x_2, p_2)$ - два примера из выборки. Тогда MixUp это процедура, которая позволяет получить новую пару (x', p'):

- $\lambda \sim Beta(\alpha, \alpha)$
- $\lambda' = max(1 - \lambda, \lambda)$ *
- $x' = \lambda' x_1 + (1 - \lambda') x_2$
- $p' = \lambda' p_1 + (1 - \lambda') p_2$

* Этой строки нет в оригинальной статье про MixUp, здесь она необходима чтоб поддерживать разделение между размеченной и неразмеченной выборками.

MixUp Examples





(b) Effect of *mixup* ($\alpha = 1$) on a toy problem. Green: Class 0. Orange: Class 1. Blue shading indicates $p(y = 1|x)$.

Процедура MixMatch

- Аугментируем данные в обеих выборках, генерируем X_a, $U_{a,k}$
- Получаем предсказание для неразмеченной части - $\hat{q}_a$
- Делаем Sharpening, получаем q_a
- Смешиваем обе выборки вместе в W
- $X'_i = MixUp(X_{a,i}, W_j)$
- $U'_i = MixUp(U_{a,i}, W_j)$
- Накладываем наши функции потерь отдельно на X' и U'

Эксперименты: Какие идеи есть?

- Consistency Regularization, добавляем вот такой лосс:

$$L = \|p(y|Augment(x), \theta) - p(y|Augment(x), \theta)\|_2^2$$

  На этой идее основаны методы Mean Teacher, усредняем веса по чекпоинтам, и Virtual Adversarial Training, пытаемся наложить такой шум, чтоб максимально поменять лейбл картинки.
- Минимизация энтропии: $H(p(y|u, \theta)) \rightarrow min$
  На этой идее основан метод Pseudo-Label. Минимизируем энтропию и затем дообучаемся на тех данных, для которых она стала мальнькой.

Эксперименты: количество размеченных данных

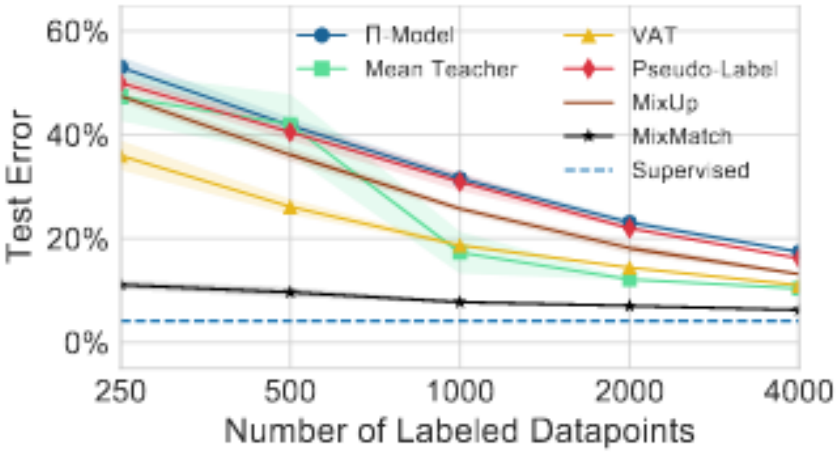


Figure 2: Error rate comparison of MixMatch to baseline methods on CIFAR-10 for a varying number of labels. Exact numbers are provided in table 5 (appendix). "Supervised" refers to training with all 50000 training examples and no unlabeled data. With 250 labels MixMatch reaches an error rate comparable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

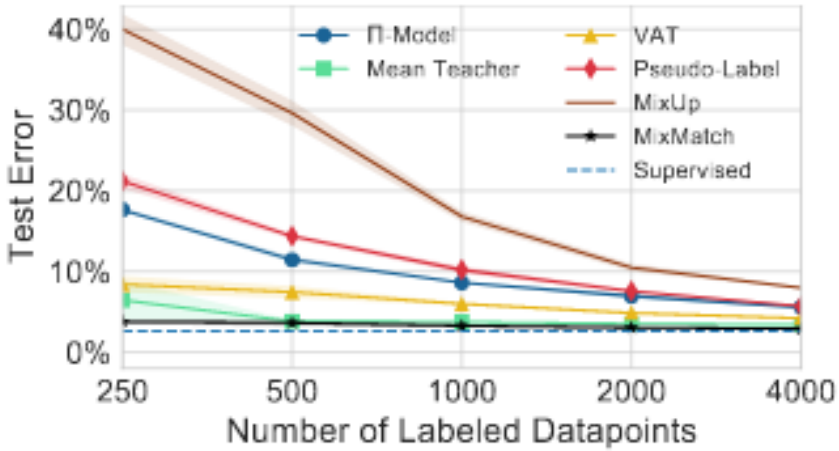


Figure 3: Error rate comparison of MixMatch to baseline methods on SVHN for a varying number of labels. Exact numbers are provided in table 6 (appendix). "Supervised" refers to training with all 73257 training examples and no unlabeled data. With 250 examples MixMatch nearly reaches the accuracy of supervised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

Эксперименты: ablation study

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.