MixMatch: A Holistic Approach to Semi-Supervised Learning

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

(* work done while at) **Google Brain**

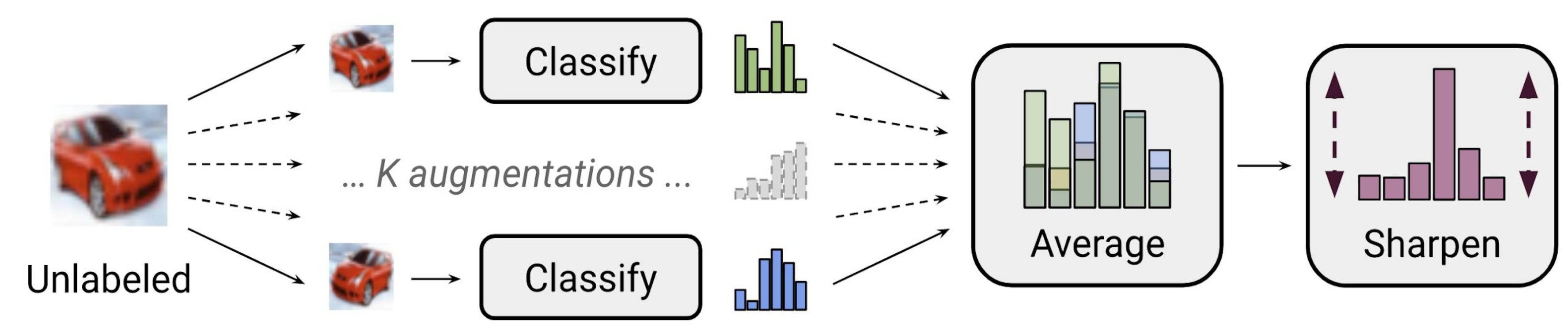
ABSTRACT

Semi-supervised learning improves classification accuracy by using unlabeled data. **MixMatch** is a new technique that unifies several existing semi-supervised learning approaches.

MixMatch guesses low-entropy and high-consistency labels for the unlabeled examples and also mixes the unlabeled data with labeled examples using MixUp.

MixMatch obtains state-of-the-art results across many tasks and fractions of labeled examples, often reducing error rate by a factor of 2 to 4.

MIXMATCH METHOD



Sharpening Temperature (0.5)
Augmentations (2)
MixUp Beta distribution parameter (0.75)

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

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

Labeled loss
Standard Cross-Entropy

$$\mathcal{L}_{\mathcal{U}} = \frac{1}{L|\mathcal{U}'|} \sum_{u,q \in \mathcal{U}'} \|q - \text{p}_{\text{model}}(y \mid u; \theta)\|_2^2$$

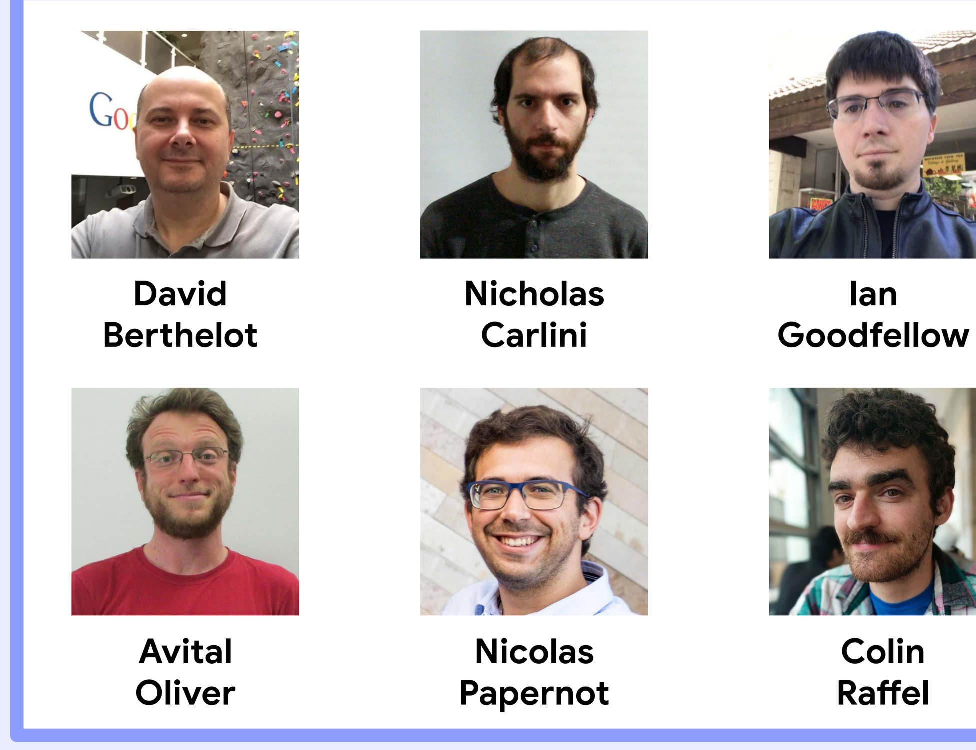
Unlabeled loss
Brier Score (L_2)

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

Note: We also use Weight Decay but don't model it as a loss since we use Adam.

AUTHORS & LINKS

Paper arxiv.org/abs/1905.02249
Code bit.ly/mixmatch-code
Contact dberth@google.com



Core Concepts

- **Minimal entropy**: the classifier must be confident.
  - Pseudo-Label, Virtual Adversarial Training + Entropy Minimization.
- **Label Consistency**: same class for *weak* augmentations of one image.
  - Pi-Model, Mean Teacher.

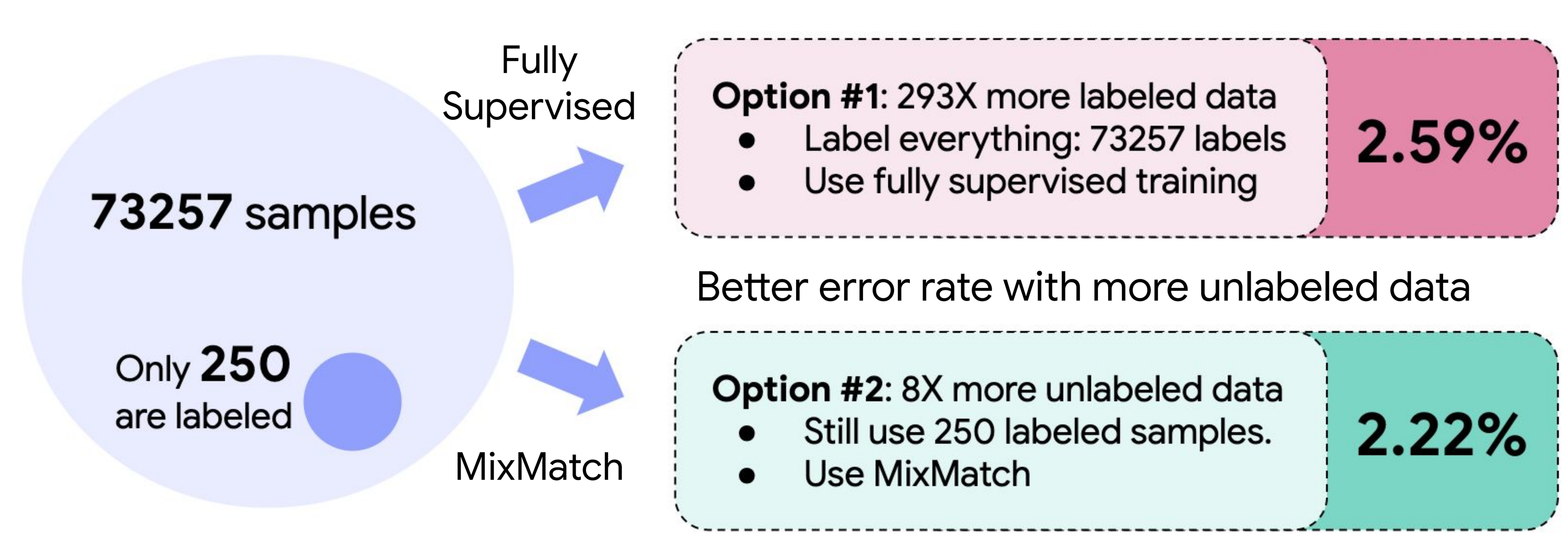
Nice to have

- **Generalizing function**: we want the classifier to generalize.
  - In practice: use popular convex / flat minima methods.
  - Concretely: MixUp, Weight Decay.

Problem

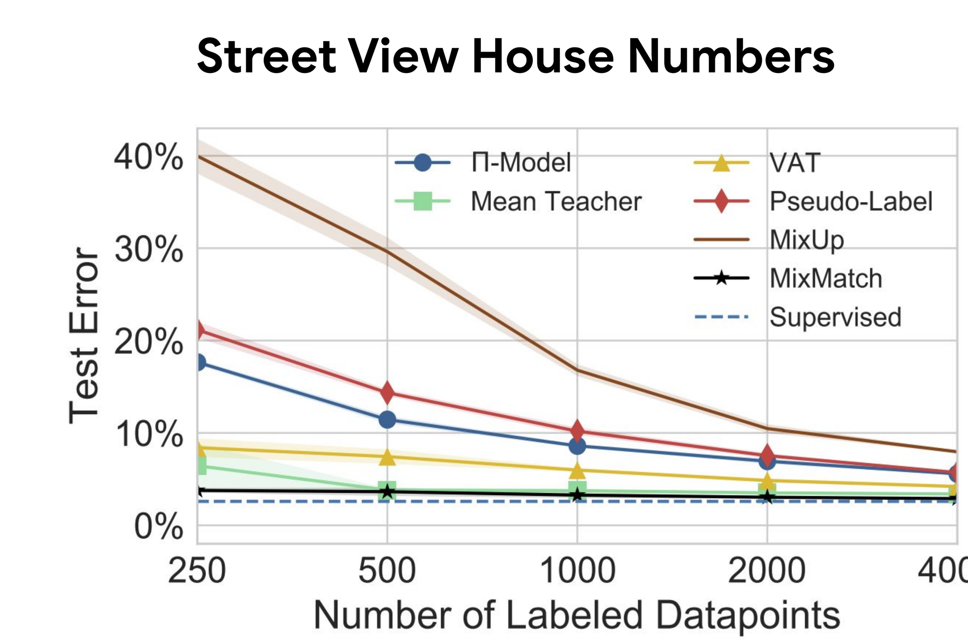
- **Above concepts are hard to combine together.**

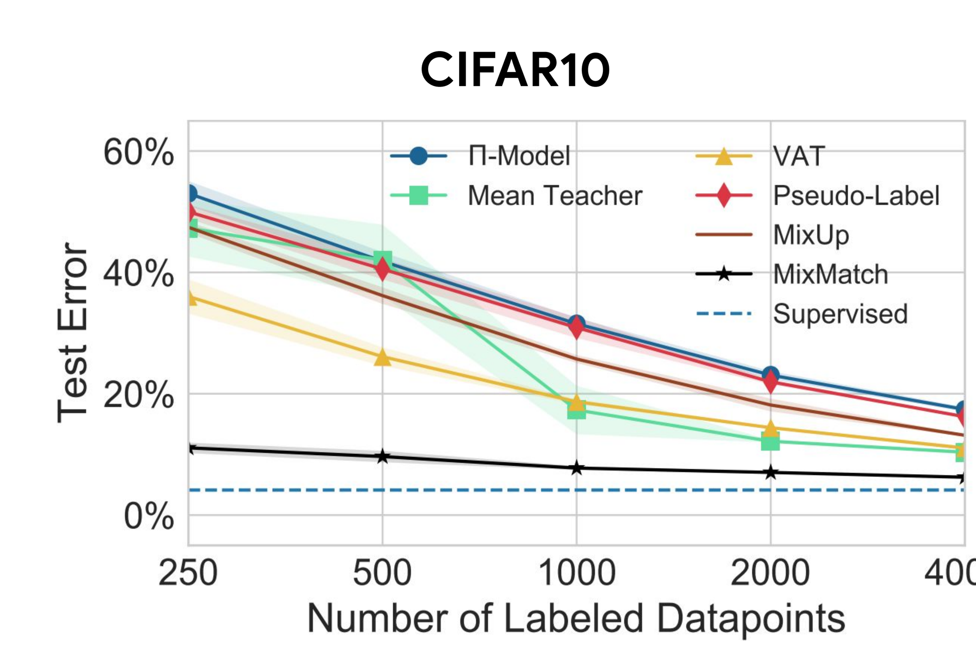
MORE LABELS OR MORE DATA?



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.





Thanks to Monica Dinculescu for help with the poster design!