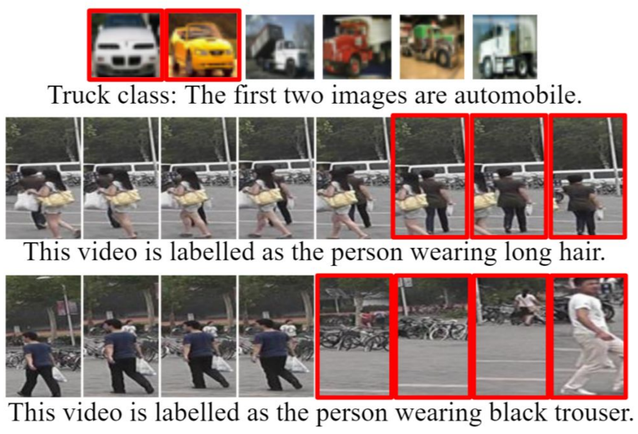
Improved Mean Absolute Error for Robust Learning

https://arxiv.org/pdf/1903.12141.pdf

Xinshao Wang, Elyor Kodirov, Yang Hua, Neil M. Robertson
Queen's University Belfast, UK AnyVision, UK

1 Introduction

- The challenge we aim to solve: **abnormal training examples**. We regard a training example as abnormal unrestrictedly whenever **an observation and its label are semantically unmatched**.



Three representatives:

1) The abnormal images with no person in 3rd row contain no semantic information at all.

2) The last abnormal image in 2nd or 3rd row may contain a person that does not belong to any person in the training set.

3) We cannot decide the object of interest without any prior when an image contains more than one object, e.g., the 2nd and 3rd last images in 2nd row contain two persons.

Fig 1: Display of abnormal training examples highlighted by red boxes.

The 1st row shows synthetic abnormal examples from corrupted CIFAR-10.

The 2nd and 3rd rows present realistic abnormal examples from video person re-identification benchmark MARS (Zheng et al. 2016).

- MAE's underfitting problem followed with New Analysis and Interpretation
  - **Underfitting observations**: In Table 1, when 40% noise exists, compared with CCE, MAE underfits to clean training data points, thus fitting much fewer abnormal examples.
  - **Conclusion**: In Fig 2, MAE emphasises more on uncertain examples, whose probabilities of being classified to its labelled class are 0.5, thus being noise-robust.
  - **Interpretation**: According to Fig 2, MAE differentiates samples in a noise-robust way, but its differentiation degree over training examples is too small.

Table 1: Classification accuracy (%) of CCE, MAE, and IMAE on CIFAR-10 (Krizhevsky 2009). 40% of training examples, i.e., the noisy subset, have wrong labels. We test each model's performance on test set, noisy subset and clean subset of training set. The backbone is ResNet56 owning enough capacity (He et al. 2016).

Loss	Test set (Generalisation)	Noisy subset (Noise-ignorance)	Clean subset (Learning ability)
CCE	63.3	75.0	96.2
MAE	66.9	8.1	**74.3** (worst)
IMAE	**81.5** (best)	**6.5** (best)	93.1

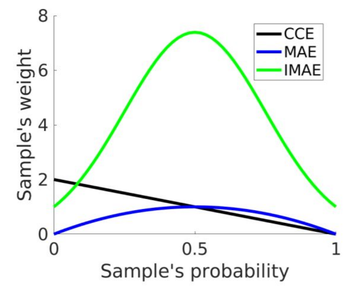
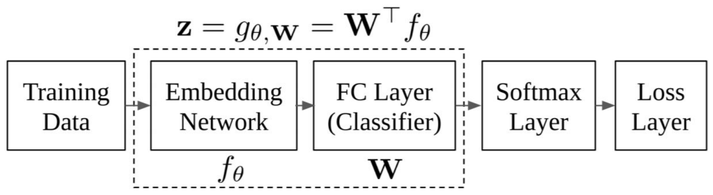


Fig 2: Sample's weight along with sample's probability being classified to its labelled class (input-to-label relevance score).

2 Methodology

- New perspectives and new conclusion about MAE.



$$p(j|\mathbf{x}_i) = \frac{\exp(\mathbf{z}_{ij})}{\sum_{m=1}^{C}\exp(\mathbf{z}_{im})}$$

Fig 3: Pipeline of a deep feature embedding network. The output of softmax layer is interpreted as classification probabilities.

$$L_{\mathrm{CCE}}(\mathbf{X};f_\theta,\mathbf{W}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} q(j|\mathbf{x}_i)\log p(j|\mathbf{x}_i) = -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|\mathbf{x}_i).$$

$$L_{\mathrm{MAE}}(\mathbf{X};f_\theta,\mathbf{W}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}|p(j|\mathbf{x}_i) - q(j|\mathbf{x}_i)| = \frac{2}{N}\sum_{i=1}^{N}(1-p(y_i|\mathbf{x}_i)),$$

$$\frac{\partial L_{\mathrm{CCE}}(\mathbf{x}_i)}{\partial p(j|\mathbf{x}_i)} = \begin{cases} -p(y_i|\mathbf{x}_i)^{-1}, & j = y_i \\ 0, & j \neq y_i \end{cases} \qquad \frac{\partial L_{\mathrm{MAE}}(\mathbf{x}_i)}{\partial p(j|\mathbf{x}_i)} = \begin{cases} -2, & j = y_i \\ 0, & j \neq y_i \end{cases}$$

Gradient magnitude w.r.t. probabilities--Prior conclusion in Zhang and Sabuncu 2018:
CCE is sensitive to abnormal data points while MAE treat all examples equally, thus being robust.

$$\frac{\partial L_{\mathrm{CCE}}(\mathbf{x}_i)}{\partial \mathbf{z}_{ij}} = \begin{cases} p(y_i|\mathbf{x}_i) - 1, & j = y_i \\ p(j|\mathbf{x}_i), & j \neq y_i \end{cases} \qquad \frac{\partial L_{\mathrm{MAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_{ij}} = \begin{cases} 2p(y_i|\mathbf{x}_i)(p(y_i|\mathbf{x}_i) - 1), & j = y_i \\ 2p(y_i|\mathbf{x}_i)p(j|\mathbf{x}_i), & j \neq y_i \end{cases}$$

$$w_{\mathrm{CCE}}(\mathbf{x}_i) = ||\frac{\partial L_{\mathrm{CCE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i}||_1 = 2(1-p(y_i|\mathbf{x}_i))$$

$$w_{\mathrm{MAE}}(\mathbf{x}_i) = ||\frac{\partial L_{\mathrm{MAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i}||_1 = 4p(y_i|\mathbf{x}_i)(1-p(y_i|\mathbf{x}_i))$$

Gradient magnitude w.r.t. logits--Our conclusion is illustrated in Fig 2.

- Interpretation of MAE's underfitting?
  MAE's differentiation degree over data points is too small, which means the relative contribution of one example versus another is not recognised well and the majority contribute almost equally. Therefore, MAE generally underfits training data especially when noise rate is high, as justified in Table 1.

- How to solve it?

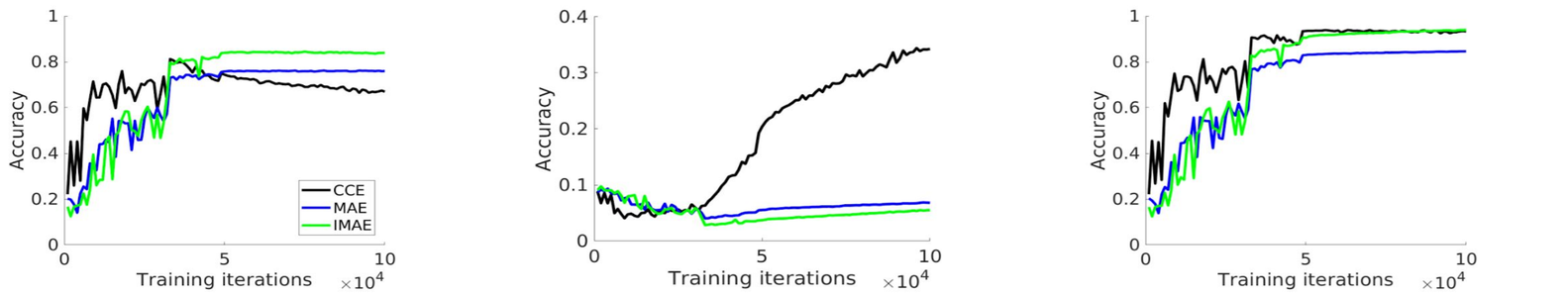
$$w_{\mathrm{IMAE}}(\mathbf{x}_i) = \exp(T\cdot p(y_i|\mathbf{x}_i)(1-p(y_i|\mathbf{x}_i)))$$

$$\frac{\partial L_{\mathrm{IMAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i} = \frac{\partial L_{\mathrm{MAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i}\cdot\frac{w_{\mathrm{IMAE}}(\mathbf{x}_i)}{w_{\mathrm{MAE}}(\mathbf{x}_i)}$$

$$=> ||\frac{\partial L_{\mathrm{IMAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i}||_1 = w_{\mathrm{IMAE}}(\mathbf{x}_i).$$

- Being the same as MAE, IMAE is a symmetric loss since it does not revise the loss computation of MAE.
  That is to say, IMAE's loss values are bounded.
- Being more flexible and practical, IMAE has a hyper-parameter T to control the weighting variance over training samples.
- The theoretical analysis on MAE's noise-robustness based on loss values (Ghosh, Kumar, and Sastry 2017) can be directly applied to IMAE.

3 Experiments on CIFAR-10 with symmetric label noise (For more experiments, please see https://arxiv.org/pdf/1903.12141.pdf)



(a) ResNet20: Testing set (higher is better). (b) ResNet20: Noisy subset (lower is better). (c) ResNet20: Clean subset (higher is better).

(d) ResNet56: Testing set (higher is better). (e) ResNet56: Noisy subset (lower is better). (f) ResNet56: Clean subset (higher is better).

Figure 4: CIFAR-10 with noise rate $r = 40\%$. The accuracies on testing set, noisy subset and clean subset of training set along with training iterations. The legend on the top left is shared by all subfigures. *Better viewed in colour.*

Backbone	r	Loss	Testing accuracy: Best	Testing accuracy: Final	Training accuracy: Naive fitting: Noisy subset	Training accuracy: Naive fitting: Clean subset	Hybrid accuracy: Meaningful patterns
ResNet20	0%	CCE	91.5	91.3	–	**100**	**98.5**
		MAE	89.3	89.2	–	95.8	94.7
		IMAE	**91.7**	**91.4**	–	99.8	98.4
	40%	CCE	81.2	67.0	34.3	93.3	72.6
		MAE	76.2	75.9	6.8	84.6	79.7
		IMAE	**84.3**	**84.0**	**5.5**	**94.0**	**88.2**
	80%	CCE	43.0	20.3	38.3	57.0	22.0
		MAE	27.7	27.5	**9.7**	29.4	27.8
		IMAE	**52.0**	**41.0**	16.8	**64.8**	**41.5**
ResNet56	0%	CCE	**92.4**	**92.2**	–	**100**	**98.7**
		MAE	89.0	89.0	–	96.1	94.9
		IMAE	92.2	**92.2**	–	99.8	98.5
	40%	CCE	81.6	63.3	75.0	**96.2**	63.6
		MAE	67.0	66.9	8.1	74.3	70.2
		IMAE	**82.2**	**81.5**	**6.5**	93.1	**86.5**
	80%	CCE	**38.2**	16.4	52.5	**62.3**	18.7
		MAE	15.2	15.1	**9.6**	15.6	15.1
		IMAE	37.1	**34.0**	13.0	44.7	**34.8**

4 Summary

- New Findings:
  - MAE emphasises on medium-probability examples, thus being noise-tolerant.
  - MAE underfits to meaningful patterns when severe noise exists.
- Solution: IMAE
  - An effective and simple solution for improving MAE's learning ability while preserving its noise-robustness.
  - It works for different types of abnormal training examples.