Example Weighting for Deep Representation Learning

Example weighting in learning to rank and classify

Guest: (Amos) Xinshao Wang
https://xinshaoamoswang.github.io/about/
https://scholar.google.com/citations?user=yOBhB7UAAAAJ&hl=en

2020/05/17@KAUST, VISION-CAIR GROUP
http://www.mohamed-elhoseiny.com/home

Outline

Outline

Research Summary

- DML:
  1. CVPR 2019 Poster: "Ranked List Loss for Deep Metric Learning," Github & Slide & Poster.
  2. AAAI 2019 Oral: "Deep Metric Learning by Online Soft Mining and Class-Aware Attention," Github & Slide & Poster.
  3. Preprint: "Instance Cross Entropy for Deep Metric Learning".
- Noise-robust DNNs:
  1. Preprint: "Derivative Manipulation for General Example Weighting" . Github.
  2. Preprint: "IMAE for Noise-Robust Learning: Mean Absolute Error Does Not Treat Examples Equally and Gradient Magnitude's Variance Matters" . Github & Poster.
  3. Preprint: "ProSelfLC: Progressive Self Label Correction for Target Revising in Label Noise" .
- Application Work:
  1. Preprint: "ID-aware Quality for Set-based Person Re-identification" .

Outline

- Learning to rank: also well known as deep (distance) metric learning (DML).
  The objective is to learn a discriminative embedding function.

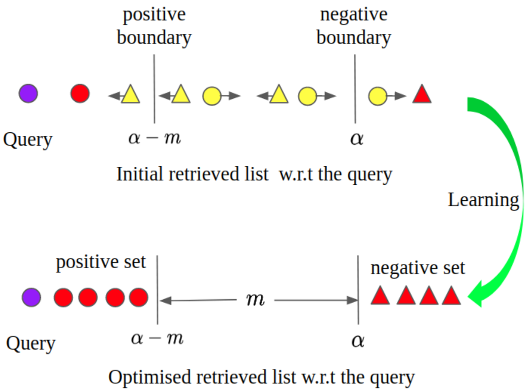


Figure: The optimisation objective of learning to rank.

- Learning to rank.
- Learning to classify.

$$\mathbf{z} = g_{\theta,\mathbf{W}} = \mathbf{W}^\top f_\theta$$

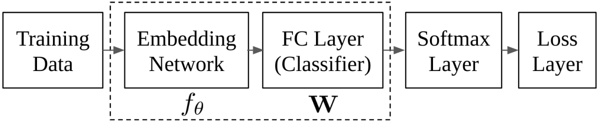


Figure: The pipeline of learning to classify.

Example weighting is universal in deep learning

- Learning to rank.
- Learning to classify.
- Example weighting

We define our interpretation of example weighting [8]:

> **Definition** (Example Weighting). *In gradient-based optimisation, the loss's derivative of an example can be interpreted as its effect on the update of a model [3, 1]. Therefore, a derivative's magnitude function equals to a weighting scheme from the viewpoint of example weighting.*

Accordingly, one technique that leads to a change of the derivative magnitude function, is implicitly equivalent to, modifying an example weighting scheme.

- The 1st row shows synthetic abnormal examples from corrupted CIFAR-10 [7].

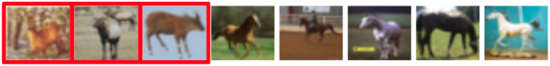
Robust deep learning

Adverse cases

- The 2nd and 3rd rows present realistic abnormal examples from video person re-identification benchmark MARS [12].

Adverse cases

Out-of-distribution anomalies: Know the unknown



Horse class: The first three images are deer semantically.

This video is labelled as the person wearing black skirt.

This video is labelled as the person wearing green shirt.

- 1) The first image in the 3rd row contains only background and no semantic information at all.
- 2) The 2nd first image or the last one in the 3rd row may contain a person that does not belong to any person in the training set.

Adverse cases

In-distribution anomalies: Detect => Correct

- 1) Some images of deer class are wrongly annotated to horse class.
- 2) We cannot decide the object of interest without any prior when an image contains more than one object, e.g., some images contain two persons in the 2nd row.

Challenge:

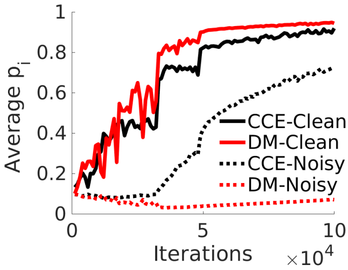
How to train a deep model robustly when some observations and annotations are not semantically matched?

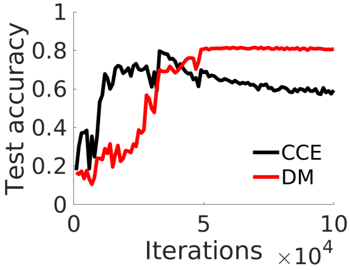
What is the meaning of robustness here?

- Objective 1: **To learn meaningful patterns** on those semantically clean training data (where noise may exist, however, the semantic matching from observations to annotations is correct).
- Objective 2: **To ignore wrong patterns** on those semantically noisy training data, so that the learning process of a model is not contaminated.

Blueprint:

What should an ideal learning process look like?





- $p_i = p(y_i|\mathbf{x}_i)$: predicted relevance between an observation $\mathbf{x}_i$ and it label y_i.
- ResNet-56 on CIFAR-10 with 40% symmetric label noise. In both CCE and our DM, noisy examples have much less p_i than clean ones, thus being more difficult examples.

Classification accuracy (%) on CIFAR-10 [7].

- 40% of training examples, i.e., the noisy subset, have wrong labels.
- The backbone is ResNet56 owning enough capacity [5].

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

IMAE for Noise-Robust Learning

- MAE Does Not Treat Examples Equally.
- Gradient Magnitude's Variance Matters.

Outline

- Ghosh et al., 2017 [2]: CCE is sensitive to label noise while MAE is noise-tolerant.

$$\begin{aligned} L_{\mathrm{CCE}}(\mathbf{x}_i, y_i) &= -\log p(y_i|\mathbf{x}_i) \\ L_{\mathrm{MAE}}(\mathbf{x}_i, y_i) &= 2(1 - p(y_i|\mathbf{x}_i)) \end{aligned} \tag{1}$$

$$\begin{aligned} \sum_{c=1}^{C} L_{\mathrm{CCE}}(\mathbf{x}_i, c) &= \sum_{c=1}^{C} \log \frac{1}{p(c|\mathbf{x}_i)} \\ \sum_{c=1}^{C} L_{\mathrm{MAE}}(\mathbf{x}_i, c) &= \sum_{c=1}^{C} (1 - p(y_i|\mathbf{x}_i)) = 2C - 2 \end{aligned} \tag{2}$$

L_{CCE} : unbounded => non-symmetric => sensitive

L_{MAE} : constant => symmetric => non-sensitive

- Ghosh et al., 2017 [2]: CCE is sensitive to label noise while MAE is noise-tolerant.
- Zhang & Sabuncu, 2018 [11]: Generalised cross entropy (GCE) concludes MAE treats training samples equally, thus being noise-robust.

$$\frac{\partial L_{\mathrm{CCE}}(\mathbf{x}_i)}{\partial p(j|\mathbf{x}_i)} = \begin{cases} -p(y_i|\mathbf{x}_i)^{-1}, & j = y_i \\ 0, & j \neq y_i \end{cases}. \tag{3}$$

$$\frac{\partial L_{\mathrm{MAE}}(\mathbf{x}_i)}{\partial p(j|\mathbf{x}_i)} = \begin{cases} -2, & j = y_i \\ 0, & j \neq y_i \end{cases}. \quad \text{constant} \tag{4}$$

IMAE for Noise-Robust Learning

MAE Treats Examples Equally?

- Zhang & Sabuncu, 2018 [11]: Generalised cross entropy (GCE) concludes MAE treats training samples equally.
- Our observation: compared with CCE, MAE underfits to clean training data points, thus fitting much fewer abnormal examples [9].
  =>MAE's fitting ability is much weaker.

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

- Our analysis: looking at $\mathbf{z}$ and its gradient magnitude

$$w_{\mathrm{CCE}}(\mathbf{x}_i) = \left|\left|\frac{\partial L_{\mathrm{CCE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i}\right|\right|_1 = 2(1 - p(y_i|\mathbf{x}_i)), \tag{5}$$

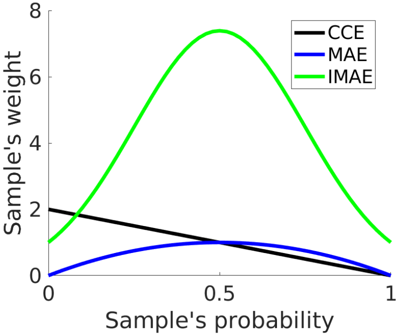
$$w_{\mathrm{MAE}}(\mathbf{x}_i) = \left|\left|\frac{\partial L_{\mathrm{MAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i}\right|\right|_1 = 4p(y_i|\mathbf{x}_i)(1 - p(y_i|\mathbf{x}_i)). \tag{6}$$

IMAE for Noise-Robust Learning

MAE Treats Examples Equally?

- Our analysis: looking at **z** and its gradient magnitude

MAE emphasises more on uncertain examples, whose probabilities of being classified to its labelled class are around 0.5, thus being noise-robust.

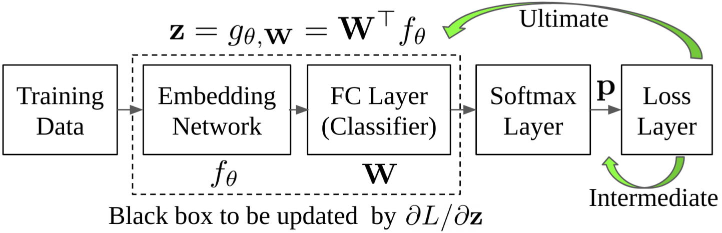


IMAE: Questions you may have

- How to define uncertain examples?

Definition 1 (Uncertain Examples). *We define uncertain examples to be those data points whose $p(y_i|\mathbf{x}_i)$ are around 0.5. Given an example $\mathbf{x}_i$, if its $p(y_i|\mathbf{x}_i)$ is closer to 0.5, its uncertainty is higher.*

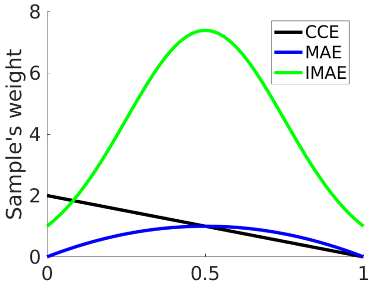
- Why looking at $\mathbf{z}$ and its gradient magnitude instead of $\mathbf{p}$?



IMAE: Questions you may have

- Why does MAE underfit to clean training data points? MAE generally underfits due to its small weights variance (0.09), leading to small impact ratio between even far different examples.

$$
\begin{aligned}
\sigma_{\mathrm{MAE}} &= \int_0^1 w^2_{\mathrm{MAE}}(p)\,\mathrm{d}p - \Big(\int_0^1 w_{\mathrm{MAE}}(p)\,\mathrm{d}p\Big)^2 \\
\sigma_{\mathrm{IMAE}} &= \int_0^1 w^2_{\mathrm{IMAE}}(p)\,\mathrm{d}p - \Big(\int_0^1 w_{\mathrm{IMAE}}(p)\,\mathrm{d}p\Big)^2.
\end{aligned} \tag{7}
$$



If probabilities are uniformly distributed, the variances of CCE's, MAE's and IMAE's weighting curves are 0.33, 0.09 and 4.55, respectively.

IMAE: Questions you may have

- How significantly does gradient magnitude's variance matter?

Outline

IMAE: Gradient Variance Matters

- Our proposed IMAE achieves new state-of-the-art simply by adjusting MAE's weight variance, which is inspiring.
  Revisit MAE:
  $w_{\mathrm{MAE}}(\mathbf{x}_i) = ||\frac{\partial L_{\mathrm{MAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i}||_1 = 4p(y_i|\mathbf{x}_i)(1 - p(y_i|\mathbf{x}_i))$.

$$\begin{aligned} & w_{\mathrm{IMAE}}(\mathbf{x}_i) = \exp(Tp(y_i|\mathbf{x}_i)(1 - p(y_i|\mathbf{x}_i))), \\ & \frac{\partial L_{\mathrm{IMAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i} = \frac{\partial L_{\mathrm{MAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i} \frac{w_{\mathrm{IMAE}}(\mathbf{x}_i)}{w_{\mathrm{MAE}}(\mathbf{x}_i)} \\ & => ||\frac{\partial L_{\mathrm{IMAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i}||_1 = w_{\mathrm{IMAE}}(\mathbf{x}_i). \end{aligned} \tag{8}$$

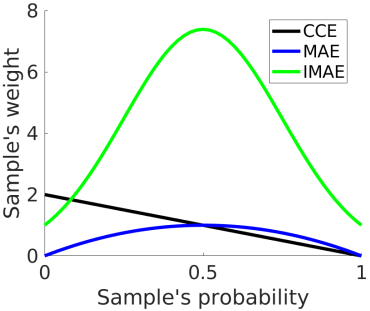
IMAE: Gradient Variance Matters

- Our proposed IMAE achieves new state-of-the-art simply by adjusting MAE's weight variance, which is inspiring.
  Revisit MAE:
  $w_{\mathrm{MAE}}(\mathbf{x}_i) = ||\frac{\partial L_{\mathrm{MAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i}||_1 = 4p(y_i|\mathbf{x}_i)(1 - p(y_i|\mathbf{x}_i))$.

$$
\begin{aligned}
&w_{\mathrm{IMAE}}(\mathbf{x}_i) = \exp(Tp(y_i|\mathbf{x}_i)(1 - p(y_i|\mathbf{x}_i))), \\
&\frac{\partial L_{\mathrm{IMAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i} = \frac{\partial L_{\mathrm{MAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i} \frac{w_{\mathrm{IMAE}}(\mathbf{x}_i)}{w_{\mathrm{MAE}}(\mathbf{x}_i)} \\
&=> ||\frac{\partial L_{\mathrm{IMAE}}(\mathbf{x}_i)}{\partial \mathbf{z}_i}||_1 = w_{\mathrm{IMAE}}(\mathbf{x}_i).
\end{aligned}
\tag{8}
$$

  (1) Direction is the same;
  (2) Magnitude variance is adjusted.

IMAE: Gradient Variance Matters

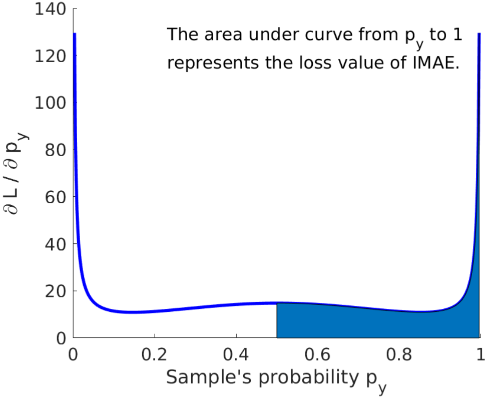


$w_{\mathrm{IMAE}}(\mathbf{x}_i) =$
$\exp(Tp(y_i|\mathbf{x}_i)(1 - p(y_i|\mathbf{x}_i)))$
$T = 8$

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

IMAE: Gradient Variance Matters

- ***IMAE is neither symmetric nor bounded***, which challenges the robustness theories studied in [2, 11, 10].
- $L_{\mathrm{IMAE}}(\mathbf{x}_i, y_i) = \int_{p_y}^{1} \frac{\exp(Tp_y(1-p_y))}{2p_y(1-p_y)} dp_y$
  For notation simplicity, $p_y = p(y_i|\mathbf{x}_i)$.



Although the loss expression of IMAE is not an elementary function, we visualise it by integral.

Outline

Why Derivative Manipulation?

Intuition and Principles

- In gradient-based optimisation, manipulating the derivative directly is more straightforward than designing loss functions, and it has a direct impact on the update of a model.
- A loss function's derivative magnitude function can be understood as a weighting scheme: the loss's derivative of an example defines how much impact it has on the update of a model.

Why Derivative Manipulation?

Two Incompatible Perspectives on Robustness

- Robustness according to loss value:
  A more robust and preferred loss function is less sensitive to large errors [4, 6].
  Under label noise, theoretically, a robust loss function should be symmetric or at least bounded [2].
- Robustness according to derivative magnitude:
  An outlier should have a smaller derivative magnitude.

Whether a larger loss value corresponds to a larger derivative depends on the particular loss functions!

When an example has a very large loss (non-robust), its derivative may be so small that its effect is negligible (robust derivative magnitude).

Outline

QUEEN'S UNIVERSITY BELFAST

Insights

- Existing robustness theorems on loss functions are not applicable.
- A loss function has a built-in example weighting scheme defined by its derivative's magnitude function.
- Be careful with your understanding: example weighting and loss functions are overlapped in the prior work.

Outline

Direction: the same as common losses

$$
\begin{aligned}
L_{\mathrm{CCE}}(\mathbf{x}_i, y_i) &= -\log p(y_i|\mathbf{x}_i) \\
L_{\mathrm{MAE}}(\mathbf{x}_i, y_i) &= 1 - p(y_i|\mathbf{x}_i) \\
L_{\mathrm{MSE}}(\mathbf{x}_i, y_i) &= (1 - p(y_i|\mathbf{x}_i))^2, \\
L_{\mathrm{GCE}}(\mathbf{x}_i, y_i) &= \frac{1 - p(y_i|\mathbf{x}_i)^q}{q},
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
\frac{\partial L_{\mathrm{CCE}}}{\partial \mathbf{z}_{ij}} &= \begin{cases} p(y_i|\mathbf{x}_i) - 1, & j = y_i \\ p(j|\mathbf{x}_i), & j \neq y_i \end{cases}. \\
\frac{\partial L_{\mathrm{MAE}}}{\partial \mathbf{z}_i} &= p_i \times \frac{\partial L_{\mathrm{CCE}}}{\partial \mathbf{z}_i}; \\
\frac{\partial L_{\mathrm{MSE}}}{\partial \mathbf{z}_i} &= 2p_i \times (1 - p_i) \times \frac{\partial L_{\mathrm{CCE}}}{\partial \mathbf{z}_i}; \\
\frac{\partial L_{\mathrm{GCE}}}{\partial \mathbf{z}_i} &= p_i^q \times \frac{\partial L_{\mathrm{CCE}}}{\partial \mathbf{z}_i}.
\end{aligned}
\tag{10}
$$

QUEEN'S UNIVERSITY BELFAST

Outline

Magnitude: Emphasis Density Function

- Example weighting in common losses

$$
\begin{aligned}
w_i^{\mathrm{CCE}} &= 2(1-p_i) => \psi_{\mathrm{CCE}} = 0; \\
w_i^{\mathrm{MAE}} &= 2p_i(1-p_i) => \psi_{\mathrm{MAE}} = 0.5; \\
w_i^{\mathrm{MSE}} &= 4p_i(1-p_i)^2 => \psi_{\mathrm{MSE}} = \frac{1}{3}; \\
w_i^{\mathrm{GCE}} &= 2p_i^q(1-p_i) => \psi_{\mathrm{GCE}} = \frac{q}{q+1}.
\end{aligned} \tag{11}
$$

- Our generalised formulation

$$
\begin{aligned}
&\nabla \mathbf{z}_i = w_i^{\mathrm{DM}}/(2(1-p_i)) \times \frac{\partial L_{\mathrm{CCE}}}{\partial \mathbf{z}_i}. \\
&w_i^{\mathrm{DM}} = \exp(\beta p_i^{\lambda}(1-p_i)) => \psi_{\mathrm{DM}} = \frac{\lambda}{\lambda+1}. \\
&\lambda \geq 0 => \psi_{\mathrm{DM}} \in [0,1).
\end{aligned} \tag{12}
$$

Magnitude: Emphasis Density Function

- Emphasis density function:

$$w_i^{\mathrm{DM}} = \exp(\beta p_i^{\lambda}(1 - p_i)) => \psi_{\mathrm{DM}} = \frac{\lambda}{\lambda + 1}.$$

$$h_{\mathrm{DM}}(w_i; \lambda, \beta) = \frac{w_i^{\mathrm{DM}}}{\int_0^1 w_i^{\mathrm{DM}} d_{p_i}} => \tag{13}$$

$$\int_0^1 h_{\mathrm{DM}}(w_i; \lambda, \beta) d_{p_i} = 1.$$

- Another generalisation: alternative EDF formats?

Magnitude: Emphasis Density Function

Other alternatives

- Normal Distribution Variant:

$$v_{\mathrm{ND}}(w_i;\psi,\beta) = \frac{\exp(-\beta p_i(p_i - 2\psi))}{\int_0^1 \exp(-\beta p_i(p_i - 2\psi))d_{p_i}}, \quad (14)$$

- Exponential Distribution Variant:
  $\lambda = 0,\ w_i^{\mathrm{DM}} = \exp(\beta(1 - p_i))$
  => Derivative Normalisation (DN).
- Beta Distribution Variant:

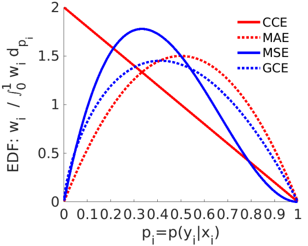
$$v_{\mathrm{BD}}(w_i;\alpha,\eta) = \frac{p_i^{\alpha-1}(1-p_i)^{\eta-1}}{\int_0^1 p_i^{\alpha-1}(1-p_i)^{\eta-1}d_{p_i}}, \alpha,\eta \geq 0$$

$$= \begin{cases} \text{CCE-DN: } \frac{w_i^{\mathrm{CCE}}}{\int_0^1 w_i^{\mathrm{CCE}} d_{p_i}}, & \alpha = 1, \eta = 2 \\ \text{MAE-DN: } \frac{w_i^{\mathrm{MAE}}}{\int_0^1 w_i^{\mathrm{MAE}} d_{p_i}}, & \alpha = 2, \eta = 2 \\ \text{MSE-DN: } \frac{w_i^{\mathrm{MSE}}}{\int_0^1 w_i^{\mathrm{MSE}} d_{p_i}}, & \alpha = 2, \eta = 3 \\ \text{GCE-DN: } \frac{w_i^{\mathrm{GCE}}}{\int_0^1 w_i^{\mathrm{GCE}} d_{p_i}}, & \alpha = q+1, \eta = 2 \end{cases} \quad (15)$$

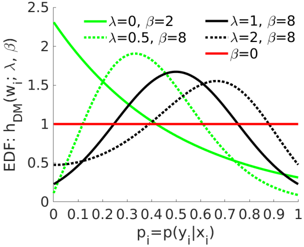
QUEEN'S UNIVERSITY BELFAST

Magnitude: Emphasis Density Function

Figure Illustration



(a) EDFs of CCE, MAE, MSE, and GCE.



(b) EDFs of DM when different λ, β are chose.

Figure: An EDF is a weight function normalised by its integral over $[0, 1]$: $\frac{w_i^{\mathrm{DM}}}{\int_0^1 w_i^{\mathrm{DM}} d_{p_i}}$ so that the total emphasis (weight) is constrained to be one unit.

Magnitude: Emphasis Density Function

Summary and definitions

Definition 1 (Emphasis Mode ψ). The "emphasis mode" refers to those examples that own the largest weight. Since an example's weight is determined by p_i, for simplicity, we define the emphasis mode to be p_i of examples whose weights are the largest, i.e., $\psi = \arg\max_{p_i} w_i, \quad \psi \in [0, 1]$.

For example, by 'emphasis mode is 0 in CCE' we mean those images with $p_i = 0$ own the highest weights.

Definition 2 (Emphasis Variance σ). The emphasis variance is the weight variance over all training instances in a mini-batch, i.e., $\sigma = \mathrm{E}((w_i - \mathrm{E}(w_i))^2)$, where $\mathrm{E}(\cdot)$ denotes the expectation of a variable.

Summary

Insights on Robust losses, Example weighting

- Existing robustness theorems on loss functions are not applicable.
- A loss function has a built-in example weighting scheme defined by its derivative's magnitude function.
- Be careful with your understanding: example weighting and loss functions are overlapped in the prior work.

Example weighting in practice

- Emphasis Mode: What examples get the highest weights?
- Emphasis Variance: How large is the variance over weights?

Thanks for your attention :)
Questions are welcome :)

References

[1] Barron, J. T. A general and adaptive robust loss function. In *CVPR*, 2019.

[2] Ghosh, A., Kumar, H., and Sastry, P. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017.

[3] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. The approach based on influence functions. In *Robust Statistics*. Wiley, 1986.

[4] Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

[5] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

[6] Huber, P. J. *Robust statistics*. Wiley, 1981.

[7] Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

[8] Wang, X., Kodirov, E., Hua, Y., and Robertson, N. M. Derivative manipulation for general example weighting. *arXiv preprint arXiv:1905.11233*, 2019.

[9] Wang, X., Kodirov, E., Hua, Y., and Robertson, N. M. Improving MAE against CCE under label noise. *arXiv preprint arXiv:1903.12141*, 2019.

[10] Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019.

[11] Zhang, Z. and Sabuncu, M. R. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018.

[12] Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., and Tian, Q. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016.