

Derivative Manipulation for General Example Weighting

Anonymous CVPR 2021 submission

Paper ID 11436

Abstract

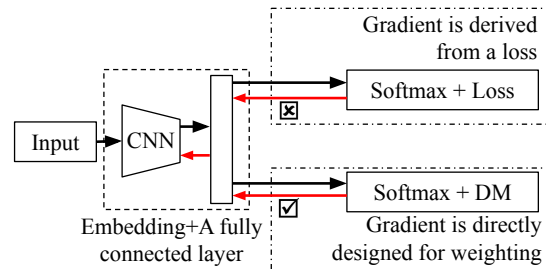
Real-world large-scale datasets usually contain noisy labels and are imbalanced. Therefore, we propose derivative manipulation (DM), a novel and general example weighting approach for training robust deep models under these adverse conditions.

DM has two main merits. First, loss function and example weighting are common techniques in the literature. DM reveals their connection (a loss function does example weighting) and is a replacement of both. Second, despite that a loss defines an example weighting scheme by its derivative, in the loss design, we need to consider whether it is differentiable. Instead, DM is more flexible by directly modifying the derivative so that a loss can be a non-elementary format too. Technically, DM defines an emphasis density function by a derivative magnitude function. DM is generic in that diverse weighting schemes can be derived.

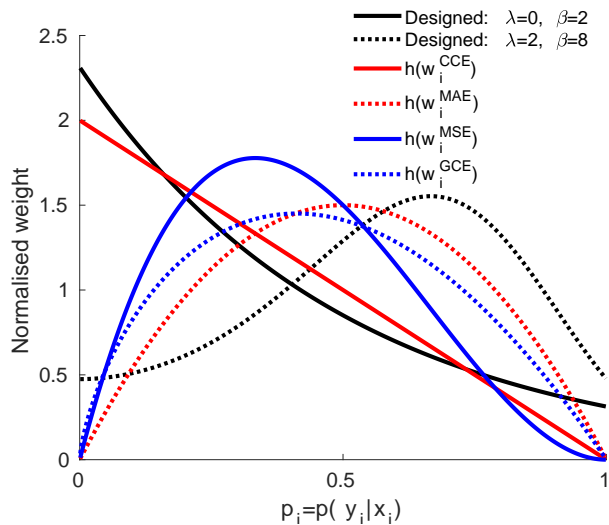
Extensive experiments on both vision and language tasks prove DM's effectiveness.

1. Introduction

In large-scale deep learning tasks, addressing label noise and sample imbalance is fundamental and has been widely studied [42, 5, 54]. There are two popular approaches, i.e., loss design [8, 61, 54] and example weighting design [5, 43, 46, 23, 24, 22], due to their easy implementation and widely demonstrated effectiveness. We reveal that a loss function does example weighting and propose DM to replace them. On the one hand, the derivative magnitude of an example decides how much impact it has on updating a model [11, 3]. Hence, a derivative magnitude function defines a weighting scheme over training examples. On the other hand, in gradient-based optimisation, the role of a loss function is to provide gradient used for back-propagation. DM designs the derivative directly so that we do not need to derive it from a loss function.



(a) Common practice and DM. Black and red arrows denote forward process and gradient back-propagation, respectively.



(b) An EDF is an example-level weighting function normalised by its integral over $[0, 1]$.

Figure 1: Illustration of DM in terms of optimisation and example weighting.

1.1. The relationship between derivative magnitude and loss value

We present two partially incompatible perspectives on the robustness of a loss function, which motivates us to design the derivative other than a loss function:

(1) *Loss value*. From this viewpoint, a loss function, which is less sensitive to large errors (i.e., residuals), is

more robust and preferred [15, 19]. For example, absolute error is considered more robust than the squared error. An outlier has a larger error by definition, but its loss value should not increase dramatically when a robust loss function is applied.

(2) *Derivative magnitude*¹. A more robust model is less affected by noisy data than clean data. Therefore, a noisy example should have a smaller derivative magnitude. However, as proved by [21] and our mathematical analysis illustrated in Figure 1b, *loss functions may have non-monotonic partial derivatives*. In such cases, *two viewpoints are not entirely consistent and the first one becomes misleading when inconsistency occurs*. We support the second because the role of a loss is offering the gradient to back-propagate as shown in Figure 1a. Therefore, derivative magnitude is a direct angle, thus being reliable.

Although it may be feasible to design a loss function which offers the desired derivative, we need to consider whether it is differentiable and derive its derivative to get its underlying weighting scheme. This “two-step” procedure makes the design of a loss more complicated than DM.

1.2. Defining example-level weighting schemes by emphasis density functions

DM nonlinearly transforms the derivative magnitude, followed by derivative normalisation (DN) so that the total emphasis (weight) is one unit. We term a normalised derivative magnitude function an Emphasis Density Function (EDF), which explicitly defines an example-level weighting scheme over data points. An EDF considers emphasis mode and variance, being analogous to a probability density function (PDF), thus we can design an EDF according to existing PDFs. *Emphasis mode* represents examples whose weight values are the largest. *Emphasis variance* decides the spread of emphasis (weight), i.e., the variance of an EDF curve. We show some EDFs in Figure 1b. Intuitively, *examples of greater “interest” should contribute more to a model’s update*. Emphasis variance controls how significantly they are emphasised. For example, depending on a scenario, it can be defined to focus on easy, semi-hard, or hard examples. Therefore, conceptually, DM is a simplified generic form of many existing heuristically-designed example weighting methods [32, 39, 43, 14, 23, 24, 22].

We summarise our contributions: (1) We propose a novel approach that unifies the design of example weighting and loss function in a single framework, named DM. (2) We demonstrate the effectiveness of DM for training robust deep networks on diverse tasks: (a) Image classification under synthetic and real-world label noise; (b) Video retrieval with unknown and diverse abnormal examples; (c) Senti-

¹Throughout this paper, derivative refers to the derivative with respect to logits instead of probabilities, inspired by [54]. We propose to design it, instead of deriving it from a loss, thus being special.

ment classification of movie reviews when label noise and sample imbalance exist. (3) We show DM with diverse network architectures and stochastic optimisers.

2. Related Work

The effects of example weighting and loss function are generally overlapped. Therefore, we need:

Rethinking existing robustness theorems on loss functions. In some prior work, when judging the robustness of a loss function, its weighting scheme is not analysed. Instead, its robustness is judged according to its sensitivity to large errors [19, 15, 51, 8, 6]. Especially in multi-class classification, [8] proposed theorems showing that a deep model is robust to label noise when the used loss function is symmetric and bounded. Accordingly, they claim that Mean Absolute Error (MAE), Mean Square Error (MSE), Categorical Cross Entropy (CCE) are decreasingly robust. However, in this work, we find CCE is very competitive with MAE, MSE and Generalised Cross Entropy (GCE) [61]. But CCE is neither bounded nor symmetric, thus being sensitive to large errors. This contrast indicates the theory may not be applicable for guiding real-world practice. For example, normalised loss functions [36] are recently proposed following IMAE [54] and this work. We remark that IMAE [54] is unpublished, and generalised in this work.

Rethinking proposed example weighting schemes. Many weighting schemes have been proposed for different purposes: (1) Easier examples are preferred and examples with lower error/loss are inferred to be easier [5, 12, 58]. For example, curriculum learning [4] picks easier examples in early training. Self-paced learning [28, 23] increases the weights of more difficult examples gradually. (2) Harder examples are emphasised: Hard example mining is demonstrated to accelerate convergence and improve performance in some cases [45, 10, 35]. However, we remark a loss’s derivative magnitude defines a weighting scheme. Then it becomes the interaction between a proposed weighting scheme and the one from a loss function works in those prior work. Instead, in DM, there is only one weighting scheme.

3. Derivative Manipulation

Let a training set be N training examples $\mathbf{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where (\mathbf{x}_i, y_i) denotes i th sample with input $\mathbf{x}_i \in \mathbb{R}^D$ and label $y_i \in \{1, 2, \dots, C\}$. C is the number of classes. Consider a deep neural network z composed of an embedding network $f(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^K$ and a linear classifier $g(\cdot) : \mathbb{R}^K \rightarrow \mathbb{R}^C$, i.e., $\mathbf{z}_i = z(\mathbf{x}_i) = g(f(\mathbf{x}_i)) : \mathbb{R}^D \rightarrow \mathbb{R}^C$. Generally, a linear classifier is the last fully-connected layer which outputs logits $\mathbf{z} \in \mathbb{R}^C$. To predict the probabilities of \mathbf{x}_i belonging to different classes, \mathbf{z} is normalised by a softmax function: $p(j|\mathbf{x}_i) = \exp(\mathbf{z}_{ij}) / \sum_{m=1}^C \exp(\mathbf{z}_{im})$,

where $p(j|\mathbf{x}_i)$ is the probability of \mathbf{x}_i belonging to class j . For brevity, we define $p_i = p(y_i|\mathbf{x}_i)$.

We study CCE, MAE, MSE and GCE. They take only p_i as input and have a minimum when p_i is 1, which makes $p(j|\mathbf{x}_i), j \neq y_i$ be zeros automatically². Their derivatives share one direction. *They perform differently only due to the derivate magnitude, which theoretically proves it is a key.* Therefore, we propose DM to systematically study derivative's magnitude with its direction kept the same as those losses. We choose L_1 norm to measure the magnitude because its expression contains only p_i , being simpler to analyse. For clarity, we formally define the emphasis mode and variance:

Definition 1 (Emphasis Mode ψ). In DM, an example's weight w_i is a function of p_i merely: $w_i = t(p_i)$, t is a mapping from p_i to w_i . Therefore, we define $\psi = \arg \max_{p_i} w_i$, $\psi \in [0, 1]$.

Definition 2 (Emphasis Variance σ). It is the weight variance over all data points in a batch and is defined by the variance of their weights. $\sigma = E((w_i - E(w_i))^2)$, where $E(\cdot)$ denotes the expectation.

3.1. The derivative direction and magnitude of common losses

CCE.³ For a given (\mathbf{x}_i, y_i) , CCE and its derivative with respect to \mathbf{z}_i are:

$$L_{\text{CCE}}(\mathbf{x}_i, y_i) = -\log p(y_i|\mathbf{x}_i) \Rightarrow \frac{\partial L_{\text{CCE}}}{\partial \mathbf{z}_{ij}} = \begin{cases} p(y_i|\mathbf{x}_i) - 1, & j = y_i \\ p(j|\mathbf{x}_i), & j \neq y_i \end{cases} \quad (1)$$

With L_1 norm, we have $\|\frac{\partial L_{\text{CCE}}}{\partial \mathbf{z}_i}\|_1 = 2(1 - p(y_i|\mathbf{x}_i)) = 2(1 - p_i)$. The weight of \mathbf{x}_i is $w_i^{\text{CCE}} = \|\frac{\partial L_{\text{CCE}}}{\partial \mathbf{z}_i}\|_1 = 2(1 - p_i)$, meaning examples with smaller p_i get higher weights.

MAE. Similarly as above:

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

The weight of \mathbf{x}_i is $w_i^{\text{MAE}} = \|\frac{\partial L_{\text{MAE}}}{\partial \mathbf{z}_i}\|_1 = 2p(y_i|\mathbf{x}_i)(1 - p(y_i|\mathbf{x}_i)) = 2p_i(1 - p_i)$.

MSE. Similarly as above:

$$L_{\text{MSE}}(\mathbf{x}_i, y_i) = (1 - p(y_i|\mathbf{x}_i))^2 \Rightarrow \frac{\partial L_{\text{MSE}}}{\partial \mathbf{z}_{ij}} = \begin{cases} -2p(y_i|\mathbf{x}_i)(p(y_i|\mathbf{x}_i) - 1)^2, & j = y_i \\ -2p(y_i|\mathbf{x}_i)(p(y_i|\mathbf{x}_i) - 1)p(j|\mathbf{x}_i), & j \neq y_i \end{cases} \quad (3)$$

²In the standard cases of MAE and MSE, $p(j|\mathbf{x}_i), j \neq y_i$ are also taken as inputs and pushed towards zeros. However, this is unnecessary because purely maximising p_i towards one naturally does that too.

³The derivation details of all losses are given in the supplementary material.

The weight of \mathbf{x}_i is $w_i^{\text{MSE}} = \|\frac{\partial L_{\text{MSE}}}{\partial \mathbf{z}_i}\|_1 = 4p(y_i|\mathbf{x}_i)(1 - p(y_i|\mathbf{x}_i))^2 = 4p_i(1 - p_i)^2$.

GCE. Similarly as above:

$$L_{\text{GCE}}(\mathbf{x}_i, y_i) = \frac{1 - p(y_i|\mathbf{x}_i)^q}{q} \Rightarrow \frac{\partial L_{\text{GCE}}}{\partial \mathbf{z}_{ij}} = \begin{cases} p(y_i|\mathbf{x}_i)^q(p(y_i|\mathbf{x}_i) - 1), & j = y_i \\ p(y_i|\mathbf{x}_i)^q p(j|\mathbf{x}_i), & j \neq y_i \end{cases} \quad (4)$$

where $q \in [0, 1]$ is a hyperparameter. The weight of \mathbf{x}_i is $w_i^{\text{GCE}} = \|\frac{\partial L_{\text{GCE}}}{\partial \mathbf{z}_i}\|_1 = 2p(y_i|\mathbf{x}_i)^q(1 - p(y_i|\mathbf{x}_i)) = 2p_i^q(1 - p_i)$.

Derivative direction. We note $\frac{\partial L_{\text{CCE}}}{\partial \mathbf{z}_i}$, $\frac{\partial L_{\text{MAE}}}{\partial \mathbf{z}_i}$, $\frac{\partial L_{\text{MSE}}}{\partial \mathbf{z}_i}$ and $\frac{\partial L_{\text{GCE}}}{\partial \mathbf{z}_i}$ share the direction. Concretely:

$$\begin{aligned} \frac{\partial L_{\text{MAE}}}{\partial \mathbf{z}_i} &= p_i \times \frac{\partial L_{\text{CCE}}}{\partial \mathbf{z}_i}; \\ \frac{\partial L_{\text{MSE}}}{\partial \mathbf{z}_i} &= 2p_i \times (1 - p_i) \times \frac{\partial L_{\text{CCE}}}{\partial \mathbf{z}_i}; \\ \frac{\partial L_{\text{GCE}}}{\partial \mathbf{z}_i} &= p_i^q \times \frac{\partial L_{\text{CCE}}}{\partial \mathbf{z}_i}. \end{aligned} \quad (5)$$

Derivative magnitude. We summarise the weighting schemes of all losses as follows:

$$\begin{aligned} w_i^{\text{CCE}} &= 2(1 - p_i) \Rightarrow \psi_{\text{CCE}} = 0; \\ w_i^{\text{MAE}} &= 2p_i(1 - p_i) \Rightarrow \psi_{\text{MAE}} = 0.5; \\ w_i^{\text{MSE}} &= 4p_i(1 - p_i)^2 \Rightarrow \psi_{\text{MSE}} = \frac{1}{3}; \\ w_i^{\text{GCE}} &= 2p_i^q(1 - p_i) \Rightarrow \psi_{\text{GCE}} = \frac{q}{q+1}. \end{aligned} \quad (6)$$

3.2. Defining an example weighting scheme by an emphasis density function

3.2.1 Example weighting via derivative manipulation

Common losses perform differently only due to the derivate magnitude as shown in the previous section. Therefore, we manipulate the derivative magnitude directly. Concretely, given a weighting function w_i^{DM} , we scale CCE's derivative by $w_i^{\text{DM}}/(2(1 - p_i))$:

$$\nabla \mathbf{z}_i = w_i^{\text{DM}}/(2(1 - p_i)) \times \frac{\partial L_{\text{CCE}}}{\partial \mathbf{z}_i}. \quad (7)$$

Then the gradient magnitude of \mathbf{z}_i is: $\|\nabla \mathbf{z}_i\|_1 = \|w_i^{\text{DM}}/(2(1 - p_i)) \times \frac{\partial L_{\text{CCE}}}{\partial \mathbf{z}_i}\|_1 = w_i^{\text{DM}}$.

Treating p_i as a continuous variable, w_i^{DM} can be interpreted as an emphasis density function (EDF). Correspondingly, the integral (area under the curve of w_i^{DM}) between a range, e.g., $[\psi_{\text{DM}} - \Delta, \psi_{\text{DM}} + \Delta]$, denotes the accumulative weight of examples whose p_i is in this range. 2Δ is the length of this range, denoting the examples of interest. As

$p_i \in [0, 1]$, we normalise an EDF by its integral over $[0, 1]$, termed a derivative normalisation (DN):

$$h(w_i^{\text{DM}}) = \frac{w_i^{\text{DM}}}{\int_0^1 w_i^{\text{DM}} dp_i} \Rightarrow \int_0^1 h(w_i^{\text{DM}}) dp_i = 1. \quad (8)$$

We name w_i^{DM} and h weighting function and EDF, respectively. In Eq. (8), the DN operator is trivial. Therefore, for brevity, we discuss the variants of w_i^{DM} instead of the normalised h . Next, we discuss how we express an example weighting function w_i^{DM} mathematically.

3.2.2 Design of example weighting schemes

We derive w_i^{DM} according to the PDFs of probability distributions of the exponential family.

Normal distribution variant. $\psi \geq 0$ denotes the emphasis mode while β adjusts the variance:

$$v_{\text{ND}}(w_i^{\text{DM}}; \psi, \beta) = \exp(-\beta p_i(p_i - 2\psi)). \quad (9)$$

Exponential distribution variant. Harder examples have larger (smaller) weights if $\beta > 0$ ($\beta < 0$):

$$v_{\text{ED}}(w_i^{\text{DM}}; \beta) = \exp(\beta(1 - p_i)). \quad (10)$$

Beta distribution variant. It covers all weighting schemes of the common losses shown in Eq. (6). We remark the difference of coefficients can be ignored since they are gone after DN. $\alpha, \eta \geq 0$.

$$v_{\text{BD}}(w_i^{\text{DM}}; \alpha, \eta) = p_i^{\alpha-1} (1 - p_i)^{\eta-1} \begin{cases} w_i^{\text{CCE}}/2 = (1 - p_i), & \alpha = 1, \eta = 2 \\ w_i^{\text{MAE}}/2 = p_i(1 - p_i), & \alpha = 2, \eta = 2 \\ w_i^{\text{MSE}}/4 = p_i(1 - p_i)^2, & \alpha = 2, \eta = 3 \\ w_i^{\text{GCE}}/2 = p_i^q(1 - p_i), & \alpha = q + 1, \eta = 2 \end{cases} \quad (11)$$

Both emphasis mode and variance matter. However, adjusting the variance is inconvenient in $v_{\text{BD}}(w_i^{\text{DM}}; \alpha, \eta)$. Although $v_{\text{ND}}(w_i; \psi, \beta)$ controls both of them, its mathematical generality to other weighting schemes is not good. Therefore, we design the other variant of w_i^{DM} as follows:

$$w_i^{\text{DM}} = \exp(\beta p_i^\lambda (1 - p_i)), \lambda \geq 0 \Rightarrow \psi_{\text{DM}} = \frac{\lambda}{\lambda + 1} \in [0, 1), \quad (12)$$

where λ and β are the parameters to control the emphasis mode and variance, respectively.

Design reasons. By varying λ and β in Eq (12), we can show that (1) if $\lambda = 0$, $w_i^{\text{DM}} = \exp(\beta(1 - p_i))$, it becomes the same as an exponential distribution variant $v_{\text{ED}}(w_i^{\text{DM}}; \beta)$; (2) If $\lambda = 1$, w_i^{DM} is a normal distribution variant; (3) Eq (12) can be viewed as an exponential transformation of $v_{\text{BD}}(w_i^{\text{DM}}; \alpha, \eta)$, where $\alpha = \lambda + 1, \eta = 2$

and scale it by β followed by an exponential transformation; (4) w_i^{DM} is an extension of w_i^{GCE} by making $\lambda \geq 0$, linear scaling and exponential transformation. As a result, the emphasis variance can be easily adjusted based on tasks.

Finally, DM's loss expression is not an elementary function and can be represented as $\int_{p_i}^1 \frac{w_i^{\text{DM}}}{2p_i(1-p_i)} dp_i$, which is unbounded and non-symmetric in multi-class cases.

4. Experiments

To prove DM's value as a useful example weighting framework, we conduct diverse experiments. Apart from label noise, all real-world datasets are highly imbalanced, e.g., the number of videos per person ranges from 1 to 271 in MARS [63], while the number of images per class varies between 18,976 and 88,588 in Clothing 1M [57]. In all experiments, we fix the random seed and do not apply any random computational accelerator for an exact comparison. To search optimisation hyperparameters, we use a separate trusted validation set on Clothing 1M. We create one when it is unavailable on CIFAR-10/100 [27]: we train on 80% training data which is corrupted in synthetic noisy cases and use 20% clean training data only for validation. This differs from the methods [50, 52, 24, 43, 32, 17, 59, 62] that exploit a clean dataset to optimise a model's learning parameters. After searching, we retrain a final model on the entire training data to fairly compare with some prior results.

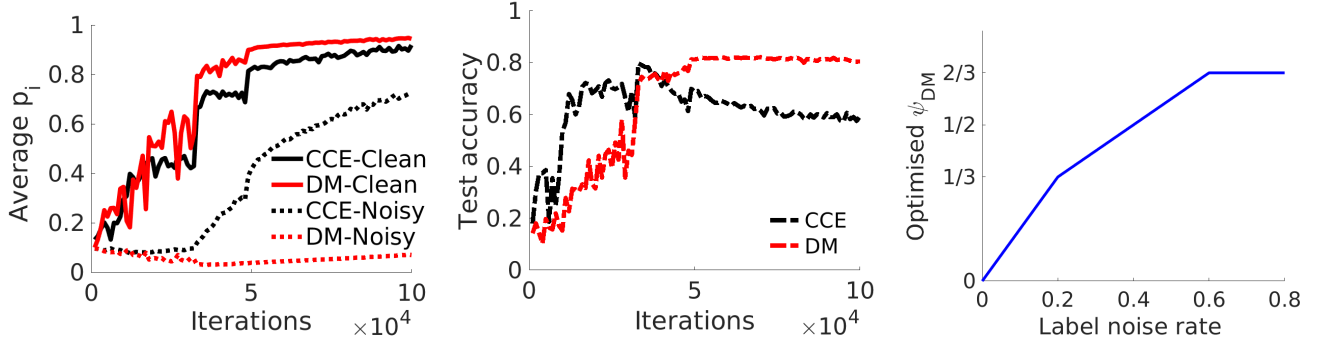
4.1. Discussion on setting the emphasis mode when noisy labels exist

We optimise the emphasis mode according to an intuition: e.g., when there exists more noise, we use a relatively larger emphasis mode to emphasise on easier examples [2, 24]. After an emphasis mode is set, we can search for the best emphasis variance on a validation set. Concretely, when training data is clean, we set $\lambda = 0$ so that $w_i^{\text{DM}} = \exp(\beta(1 - p_i)) \Rightarrow \psi_{\text{DM}} = 0$. When label noise exists, we increase λ so that ψ_{DM} increases. For clarity, we summarise:

Premise 1. *Difficult examples have smaller probabilities of being predicted to its annotated labels, i.e., smaller p_i . Abnormal examples, including noisy ones and outliers, belong to those difficult ones.*

This premise is justified in recent work: [2] shows that DNNs do not fit real datasets by brute-force memorisation. Instead, DNNs learn simple shared patterns before memorising difficult abnormal data points. Consequently, abnormal samples have smaller p_i than easier ones.

In addition, we present our empirical evidence in Figure 2. The learning dynamics prove this premise and DM is superior to CCE: (a) The p_i of clean examples increases while that of noisy ones has no noticeable rise in DM, which means DM hinders fitting of abnormal examples and preserves DNNs' ability to learn on clean data; (b) DM has the



(a) The average p_i of different examples as training goes in CCE and DM.

(b) The test accuracy of CCE and DM as the iteration increases.

(c) We optimise ψ_{DM} over four settings from $\{0, 1/3, 1/2, 2/3\}$.

Figure 2: We train ResNet-56 on CIFAR-10. In (a) and (b), we observe noisy examples have much less p_i than clean ones, thus being more difficult examples in both CCE and DM. In (c), for each label noise rate, we show the optimised ψ_{DM} from $\{0, 1/3, 1/2, 2/3\}$, i.e., $\lambda \in \{0, 1/2, 1, 2\}$.

Table 1: Test accuracy (%) on CIFAR-100. The best results on each block are bolded. The italic row is the most basic baseline where examples have identical weights.

	Method	Clean Labels	Symmetric Noisy Labels			Asymmetric Noisy Labels		
			$r=0.2$	$r=0.4$	$r=0.6$	$r=0.2$	$r=0.3$	$r=0.4$
Results From SL [56]	LS	63.7	58.8	50.1	24.7	63.0	62.3	61.6
	Boot-hard	63.3	57.9	48.2	12.3	63.4	63.2	62.1
	Forward	64.0	59.8	53.1	24.7	64.1	64.0	60.9
	D2L	64.6	59.2	52.0	35.3	62.4	63.2	61.4
	SL	66.8	60.0	53.7	41.5	65.6	65.1	63.1
Our Trained Results	CCE	70.0	60.4	53.2	42.1	66.4	64.7	60.3
	GCE	63.6	62.4	58.6	50.6	62.8	62.2	58.7
	MAE	8.2	6.4	7.3	5.2	7.3	6.3	7.3
	MSE	28.0	24.6	21.3	18.0	24.5	24.3	23.0
	CCE-DN	69.1	60.7	54.2	44.6	65.9	64.0	60.5
	GCE-DN	65.8	62.5	58.3	48.4	64.1	62.1	60.3
	MAE-DN	7.5	5.4	4.5	4.8	5.8	3.5	3.9
	MSE-DN	25.8	28.4	27.0	26.5	27.4	23.9	25.3
	$DM(\beta = 0)$	67.2	56.2	50.9	44.4	64.4	62.5	60.4
	$DM(\lambda = 0)$	70.1	60.9	55.2	44.6	65.5	63.1	60.6
	$DM(\lambda = 0.5)$	69.3	65.7	61.0	52.9	67.4	65.0	60.8
	$DM(\lambda = 1)$	69.2	63.4	54.7	43.9	67.5	65.8	63.3

best test accuracy. Furthermore, it is robust, i.e., early stopping is unnecessary. (c) As we increase the noise rate, the optimised emphasis mode also increases showing a positive correlation between them. A thorough ablation study on the emphasis mode and variance is reported in the Appendix E.

Remark 1. In the early training phase, p_i is awfully random and non-informative, what does DM do? At this phase, DM randomly weights examples since p_i is random. It is the same in all common losses (weighting schemes). Although nothing makes sense at the beginning, including random initialisation, loss values and weighting schemes, DNNs grad-

ually learn meaningful patterns.

4.2. Robust image classification

Datasets. (1) CIFAR-10/100 [27], which contain 10 and 100 classes, respectively. The image size is 32×32 . In CIFAR-10, the training data contains 5k images per class while the test set includes 1k images per class. CIFAR-100 has 500 images per class for training and 100 images per class for testing. We generate synthetic label noise on them: (a) Symmetric noise. With a probability of r , the original label of an image is changed to one of the other class labels

Table 2: Accuracy (%) of DM and other baselines on CIFAR-10 under symmetric label noise. The best results on each block are bolded. Number format of this table follows MentorNet. ‘-’ denotes result was not reported. The italic row is the most basic baseline where examples have identical weights.

	Method	$r=0$	$r=0.2$	$r=0.4$	$r=0.8$
Results From MentorNet	Forgetting	-	0.76	0.71	0.44
	Self-paced	-	0.80	0.74	0.33
	Focal Loss	-	0.77	0.74	0.40
	Boot-soft	-	0.78	0.73	0.39
	MentorNet PD	-	0.79	0.74	0.44
	MentorNet DD	-	0.79	0.76	0.46
Our Trained Results	CCE	0.85	0.74	0.74	0.35
	GCE	0.83	0.81	0.77	0.18
	MAE	0.57	0.50	0.45	0.19
	MSE	0.80	0.78	0.73	0.29
	CCE-DN	0.84	0.75	0.76	0.18
	GCE-DN	0.77	0.82	0.79	0.19
	MAE-DN	0.10	0.10	0.10	0.19
	MSE-DN	0.85	0.51	0.76	0.53
	<i>DM($\beta = 0$)</i>	0.86	<i>0.66</i>	<i>0.56</i>	<i>0.18</i>
	DM($\lambda = 0.0$)	0.86	0.77	0.75	0.18
	DM($\lambda = 0.5$)	0.86	0.83	0.80	0.57

uniformly following [37, 56]. r denotes the noise rate. We remark that some work randomly flips an image’s label to one of all labels including the ground-truth [49, 25]. The actual noise rate becomes quite different when the number of classes is small, e.g., CIFAR-10. We do not compare with those results. (b) Asymmetric noise. We generate asymmetric noise for CIFAR-100 following [56]. CIFAR-100 has 20 superclasses and every superclass has 5 subclasses. In each superclass, two subclasses are randomly selected and their labels are flipped to each other with a probability of r . The overall noise rate is less than r . (2) Clothing 1M [57] is a real-world dataset and contains about 1 million images of 14 classes from shopping websites. Its noise rate is about 38.46% and noise distribution is agnostic.

4.2.1 Results on CIFAR-100

Training details. We follow the settings of [56] for a fair comparison. We use SGD with a momentum of 0.9 and a weight decay of $1e-4$. We train 30k iterations. The learning rate starts at 0.1, and is divided by 10 at 15k and 22k iterations. Standard data augmentation is used: padding images with 4 pixels on every side, followed by a random crop of 32×32 and horizontal flip. The batch size is 256. All methods use ResNet-44 [16].

Competitors. We briefly introduce the compared baselines: (1) Analysed losses (CCE, GCE, MAE and MSE) and their variants after DN (CCE-DN, GCE-DN, MAE-DN and MSE-DN); (2) Bootstrapping trains a model with new labels generated by a convex combination of the original ones and their predictions. A convex combination can be either soft (Boot-soft) or hard (Boot-hard) [42]; (3) Forward (Backward) uses a noise-transition matrix to multiply the network’s predictions (losses) for label correction [41]; (4) D2L addresses noise-robustness by restricting the dimensionality expansion of learned subspaces during training; (5) SL modifies CCE symmetrically with a reverse cross entropy; (6) LS denotes label smoothing [18]. Note that we do not compare with [30]. First, its backbone is not ResNet-44 after checking with the authors. Second, their algorithm is orthogonal to ours because it targets at the inference stage and is a generative classifier on top of deep representations. **Results.** From Table 1, our observations are: (1) When training data is clean, CCE (CCN-DN) is the best against other common losses. Besides, DM($\lambda = 0$) is the best compared with other variants. We conclude by $\psi = 0$, harder examples have higher weights, leading to better performance. (2) When label noise exists, we obtain better performance by increasing λ so that ψ is larger. This demonstrates that we should focus on easier data points as label noise increases.

4.2.2 Results on CIFAR-10

Training details. We follow the same settings as MentorNet [24] and train GoogLeNet V1 to compare fairly with its reported results. Optimiser and data augmentation are the same as CIFAR-100.

Competitors. *Self-paced* [28], *Focal Loss* [33], and *MentorNet* are representatives of example weighting algorithms. Forgetting [2] searches the dropout parameter in the range of (0.2, 0.9). All methods use GoogLeNet V1 [48].

Results. In Table 2, we observe: (1) When looking at common losses, they perform differently in different cases. For example, CCE and MSE-DN are better when $r = 0$ and $r = 0.8$. GCE and GCE-DN are preferred when $r = 0.2$ and $r = 0.4$; (2) For DM, we first set $\lambda = 0.5$, then optimise β manually. We find it performs the best. $\beta = 0$ denotes all samples have identical weights.

4.2.3 Results on Clothing 1M

Training details. We follow [49] to train ResNet-50 [16]: (1) We initialise it by a pretrained model on ImageNet [44]; (2) SGD with a momentum of 0.9 and a weight decay of $2e-5$ is applied. The learning rate starts at 0.01 and is divided by 10 at 10k and 15k iterations. We train 20k iterations; (3) Data augmentation: first resize a raw input image

Table 3: Accuracy (%) on Clothing1M. The leftmost block’s results are from SL while the middle block’s are from Masking. Results of corresponding variants after DN are in the brackets.

Boot-hard	D2L	Forward	SL	S-adaptation	Masking	Joint Optim.	MD-DYR-SH	Our Trained Results				
								CCE(-DN)	GCE(-DN)	MAE(-DN)	MSE(-DN)	DM
68.9	69.5	69.8	71.0	70.3	71.1	72.2	71.0	71.7(72.5)	72.4(64.5)	39.7(16.4)	71.7(69.9)	73.3

Table 4: Video retrieval results on MARS dataset. All other methods use GoogLeNet V2 except that DRSA and CAE use more complex ResNet-50.

Metric	DRSA	CAE	OSM+CAA	Our Trained Results				
				CCE	GCE	MAE	MSE	DM
mAP (%)	65.8	67.5	72.4	58.1	31.6	12.0	19.6	72.8
CMC-1 (%)	82.3	82.4	84.7	73.8	51.5	26.0	39.3	84.3

to 256×256 , and then crop it randomly at 224×224 followed by random horizontal flipping. We set $\lambda = 1, \beta = 2$ for DM.

Competitors. We compare with recent algorithms: (1) S-adaptation applies an auxiliary softmax layer to estimate a noise-transition matrix [9]; (2) Masking is a human-assisted approach that conveys human cognition to speculate the structure of a noise-transition matrix [13]; (3) Joint Optim. [49] learns latent true labels and model’s parameters iteratively. Two regularisation terms are added for label estimation and adjusted in practice; (4) MD-DYR-SH [1] combines dynamic mixup (MD), dynamic bootstrapping plus regularisation (DYR) from soft to hard (SH).

Results. In Table 3, under real-world agnostic noise, DM outperforms the state-of-the-art. It is worth noting that the burden of noise-transition matrix estimation in Forward, S-adaptation, Masking and Joint Optim. is heavy, whereas DM is simple and effective.

4.3. Robust video retrieval

Dataset and evaluation settings. MARS [63] contains 20,715 videos of 1,261 persons. There are 1,067,516 frames in total. Because person videos are collected by tracking and detection algorithms, abnormal examples exist as shown in Figure 3 in the supplementary material: *Some frames contain only background or an out-of-distribution person. Exact noise type and rate are unknown.* We use 8,298 videos of 625 persons for training and 12,180 videos of the other 636 persons for testing. We report the cumulated matching characteristics (CMC) and mean average precision (mAP) results [63].

Implementation details. Following [34, 53], we train GoogLeNet V2 [20] and process a video as an image set, which means we use only appearance information without exploiting latent temporal information. A video’s representation is simply the average fusion of its frames’ representations. The learning rate starts from 0.01 and is divided by 2 every 10k iterations. We stop training at 50k iterations.

We apply an SGD optimiser with a weight decay of $5e - 4$ and a momentum of 0.9. The batch size is 180. Data augmentation is the same as Clothing 1M. At testing, we L_2 normalise videos’ features and calculate the cosine similarity.

Results. The results are displayed in Table 4. Although DRSA [31] and CAE [7] exploit extra temporal information by incorporating attention mechanisms, DM is superior to them in terms of both effectiveness and simplicity. OSM+CAA [53] is the only competitive method. However, OSM+CAA combines CCE and weighted contrastive loss to address anomalies, thus being more complex. We highlight that one query may have multiple matching instances in the MARS benchmark so that mAP is a more reliable and accurate performance assessment. DM is the best in terms of mAP.

4.4. Sentiment analysis of movie reviews

We report results on the IMDB dataset of movie reviews [38] following [40]. Paragraph Vector (PV-DBOW) is used as a document descriptor [29]. We train a neural network with one 8-neuron hidden layer and display the results in Table 5. IMDB contains 25,000 positive movie reviews and 25,000 negative ones. We follow [38, 29, 40] to split them evenly for training and testing, respectively. PV-DBOW represents every movie review using a fixed-length feature vector. It is a binary classification problem. Our implementation benefits from the codes publicly available at <https://github.com/mesnilgr/iclr15> and https://github.com/shaform/experiments/tree/master/caffe_sentiment_analysis.

4.4.1 Label noise

We generate symmetric label noise. It is a binary classification problem so that the maximum noise rate that we can generate is 50%. We test three cases: $r = 0.0, 0.2, 0.4$. We choose an exponential distribution variant as an EDF, i.e.,

Table 5: Experiments on sentiment classification of movie reviews. We test on two adverse cases: label noise and sample imbalance. P-N Ratio denotes the ratio of positive reviews to negative ones.

		CCE	GCE	MAE	MSE	CCE-DN	GCE-DN	MAE-DN	MSE-DN	DM
Label Noise Rate r	0.0	88.9	88.9	88.9	88.9	88.5	87.7	74.5	88.2	89.1
	0.2	87.7	88.6	88.6	88.5	87.2	85.5	72.8	87.0	88.7
	0.4	75.5	83.6	84.9	83.7	75.0	75.2	80.2	75.3	86.4
P-N Ratio	10 : 1	78.9	77.0	75.4	77.6	77.5	59.1	0.5	78.5	80.6
	50 : 1	63.4	0.5	0.5	58.6	61.4	0.5	0.5	64.4	65.0

Table 6: Results of DM and other standard regularisers on CIFAR-100. We set $r = 40\%$, i.e., the label noise is severe but not belongs to the majority. We train ResNet-44. We report the average test accuracy and standard deviation (%) over 5 trials. Baseline is CCE without regularisation.

Baseline	L2	Dropout	Dropout+L2	DM	DM+L2	DM+Dropout	DM+L2+Dropout
44.7 \pm 0.1	51.5 \pm 0.4	46.7 \pm 0.5	52.8 \pm 0.4	55.7 \pm 0.3	59.3\pm0.2	54.3 \pm 0.4	58.3 \pm 0.3

Table 7: Results of common stochastic optimisers. We report three settings of the Adam optimiser [26].

	CCE	GCE	MAE	MSE	DM
SGD (lr: 0.01)	64.6	68.8	39.3	58.4	82.0
SGD + Momentum (lr: 0.01)	61.7	80.7	64.7	76.7	83.8
Nesterov (lr: 0.01)	57.3	80.0	63.9	76.8	84.0
Adam (lr: 0.01, delta: 0.1)	39.3	75.7	57.5	66.8	78.2
Adam (lr: 0.005, delta: 0.1)	44.3	72.6	60.8	67.9	80.8
Adam (lr: 0.005, delta: 1)	52.0	67.7	37.3	58.5	79.2

$\lambda = 0$. Additionally, if $r = 0.0$, $\beta > 0$, and $\beta < 0$ otherwise. Specifically:

- (1) When it is clean, we set $\beta = 0.9$ so that larger weights are assigned to harder examples.
- (2) When noise exists, i.e., $r = 0.2$ and $r = 0.4$, we set $\beta = -0.52$ and $\beta = -0.33$, respectively. Therefore, easier examples have larger weights.

We remark that because IMDB has only two classes, its settings of λ and β are different from other datasets.

4.4.2 Sample imbalance

P-N Ratio denotes the ratio of positive reviews to negative ones. We use all the positive reviews (12500) and randomly sample a small proportion of negative reviews:

- (1) When the P-N Ratio is 10:1, 1250 negative reviews are randomly sampled.
- (2) When the P-N Ratio is 50:1, 250 negative reviews are randomly sampled.

In all experiments, we set $\lambda = 0.3$, $\beta = 6.5$. Due to the space limitation, other details are reported in the Appendix D. Table 5 summarises our results. We remark that when the result is 0.5, the trained model is non-meaningful

as IMDB has only two classes.

4.5. Further analysis

In Table 6, we compare our proposed DM with other standard regularisers, i.e., L2 weight decay and Dropout [47]. We set the dropout rate to 0.2 and L2 weight decay rate to 10^{-4} . For DM, we fix $\beta = 8$, $\lambda = 0.5$. DM is better than those standard regularisers and their combinations significantly. DM works best when it is together with L2 weight decay.

We also experimented with Adam optimiser as shown in Table 7. To explore different networks simultaneously, we train ResNet-56 [16] instead of GoogLeNet V1 on CIFAR-10 with 40% symmetric label noise. We observe that DM's results are consistently the best.

Additionally, the performance on a small-scale dataset and the detailed training analysis under different levels of label noise are presented in the Appendixes C and E, respectively.

5. Conclusion and Future Work

In this work, we propose derivative manipulation for example weighting. DM directly works on gradients bypassing a loss function. As a consequence, it creates great flexibility in designing various example weighting schemes. Extensive experiments on both vision and language tasks empirically show that DM outperforms existing methods despite its simplicity.

For the future work, firstly, we will pay attention to how to set the emphasis mode and variance adaptively on different datasets and progressively along with training time. Secondly, some other losses have different derivative directions, e.g., when target modification is applied [55]. We acknowledge that this is also a valuable area for future research, e.g., how to manipulate the derivative direction.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *ICML*, 2019. 7

[2] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *ICML*, 2017. 4, 6

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

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009. 2

[5] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *NeurIPS*, 2017. 1, 2

[6] Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In *ICML*, 2019. 2

[7] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, 2018. 7

[8] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017. 1, 2

[9] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017. 7

[10] Siddharth Gopal. Adaptive sampling for sgd by exploiting side information. In *ICML*, 2016. 2

[11] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and WA Stahel. The approach based on influence functions. In *Robust Statistics*. Wiley, 1986. 1

[12] Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor W Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. In *ICML*, 2020. 2

[13] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. In *NeurIPS*, 2018. 7

[14] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018. 2

[15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 8

[17] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, 2018. 4

[18] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015. 6

[19] Peter J Huber. *Robust statistics*. Wiley, 1981. 2

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 7

[21] Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *Schedae Informaticae*, pages 49–59, 2016. 2

[22] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *ICML*, 2020. 1, 2

[23] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. Self-paced curriculum learning. In *AAAI*, 2015. 1, 2

[24] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018. 1, 2, 4, 6

[25] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *CVPR*, 2019. 6

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 8

[27] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 4, 5, 11, 14

[28] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *NeurIPS*, 2010. 2, 6

[29] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014. 7

[30] Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. In *ICML*, 2019. 6

[31] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 2018. 7

[32] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *ICCV*, 2017. 2, 4

[33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, 2017. 6

[34] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017. 7

[35] Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. In *ICLR Workshop*, 2016. 2

[36] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. 2020. 2

[37] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah M Erfani, Shu-Tao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *ICML*, 2018. 6

[38] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL*, 2011. 7

[39] Eran Malach and Shai Shalev-Shwartz. Decoupling "when to update" from "how to update". In *NeurIPS*, 2017. 2

[40] Grégoire Mesnil, Tomas Mikolov, Marc'Aurelio Ranzato, and Yoshua Bengio. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. In *ICLR Workshop*, 2015. 7

[41] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017. 6

[42] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *ICLR Workshop*, 2015. 1, 6

[43] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018. 1, 2, 4

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 211–252, 2015. 6

[45] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 2

[46] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019. 1

[47] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, pages 1929–1958, 2014. 8

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 6

[49] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, 2018. 6, 7

[50] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *NeurIPS*, 2017. 4

[51] Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. In *NeurIPS*. 2015. 2

[52] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, 2017. 4

[53] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, and Neil M. Robertson. Deep metric learning by online soft mining and class-aware attention. In *AAAI*, 2019. 7

[54] Xinshao Wang, Yang Hua, Elyor Kodirov, and Neil M Robertson. IMAE for noise-robust learning: Mean absolute error does not treat examples equally and gradient magnitude's variance matters. *arXiv preprint arXiv:1903.12141*, 2019. 1, 2

[55] Xinshao Wang, Yang Hua, Elyor Kodirov, and Neil M Robertson. ProSelfLC: Progressive self label correction for training robust deep neural networks. *arXiv preprint arXiv:2005.03788*, 2020. 8

[56] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019. 5, 6

[57] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015. 4, 6

[58] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and J Kwok. Searching to exploit memorization effect in learning with noisy labels. In *ICML*, 2020. 2

[59] Serena Yeung, Vignesh Ramanathan, Olga Russakovsky, Liyue Shen, Greg Mori, and Li Fei-Fei. Learning to learn from noisy web videos. In *CVPR*, 2017. 4

[60] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017. 15

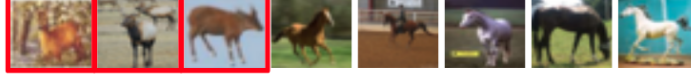
[61] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018. 1, 2

[62] Zizhao Zhang, Han Zhang, Serkan O Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *CVPR*, 2020. 4

[63] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 4, 7, 11

Supplementary Material of Derivative Manipulation

A. Display of Semantically Abnormal Training Examples



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.

Figure 3: Diverse semantically abnormal training examples highlighted by red boxes. The 1st row shows synthetic abnormal examples from corrupted CIFAR-10 [27]. The 2nd and 3rd rows present realistic abnormal examples from video person re-identification benchmark MARS [63].

Out-of-distribution anomalies: 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.

In-distribution anomalies: 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.

B. Derivation Details of Softmax, CCE, MAE and GCE

B.1. Derivation of Softmax Normalisation

We rewrite $p(y_i|\mathbf{x}_i)$ as follows:

$$p(y_i|\mathbf{x}_i)^{-1} = 1 + \sum_{j \neq y_i} \exp(\mathbf{z}_{ij} - \mathbf{z}_{iy_i}). \quad (13)$$

For left and right sides of Eq. (13), we calculate their derivatives w.r.t. \mathbf{z}_{ij} simultaneously.

If $j = y_i$,

$$\begin{aligned} \frac{-1}{p(y_i|\mathbf{x}_i)^2} \frac{\partial p(y_i|\mathbf{x}_i)}{\mathbf{z}_{iy_i}} &= - \sum_{j \neq y_i} \exp(\mathbf{z}_{ij} - \mathbf{z}_{iy_i}) \\ \Rightarrow \frac{\partial p(y_i|\mathbf{x}_i)}{\mathbf{z}_{iy_i}} &= p(y_i|\mathbf{x}_i)(1 - p(y_i|\mathbf{x}_i)). \end{aligned} \quad (14)$$

If $j \neq y_i$,

$$\begin{aligned} \frac{-1}{p(y_i|\mathbf{x}_i)^2} \frac{\partial p(y_i|\mathbf{x}_i)}{\mathbf{z}_{ij}} &= \exp(\mathbf{z}_{ij} - \mathbf{z}_{iy_i}) \\ \Rightarrow \frac{\partial p(y_i|\mathbf{x}_i)}{\mathbf{z}_{ij}} &= -p(y_i|\mathbf{x}_i)p(j|\mathbf{x}_i). \end{aligned} \quad (15)$$

In summary, the derivation of softmax layer is:

$$\frac{\partial p(y_i|\mathbf{x}_i)}{\partial \mathbf{z}_{ij}} = \begin{cases} p(y_i|\mathbf{x}_i)(1 - p(y_i|\mathbf{x}_i)), & j = y_i \\ -p(y_i|\mathbf{x}_i)p(j|\mathbf{x}_i), & j \neq y_i \end{cases} \quad (16)$$

B.2. Derivation of CCE

According to Eq. (1), we have

$$L_{\text{CCE}}(\mathbf{x}_i; f_\theta, \mathbf{W}) = -\log p(y_i|\mathbf{x}_i). \quad (17)$$

Therefore, we obtain (the parameters are omitted for brevity),

$$\frac{\partial L_{\text{CCE}}}{\partial p(j|\mathbf{x}_i)} = \begin{cases} -p(y_i|\mathbf{x}_i)^{-1}, & j = y_i \\ 0, & j \neq y_i \end{cases}. \quad (18)$$

B.3. Derivation of MAE

According to Eq. (2), we have

$$L_{\text{MAE}}(\mathbf{x}_i; f_\theta, \mathbf{W}) = 1 - (p(y_i|\mathbf{x}_i)). \quad (19)$$

Therefore, we obtain

$$\frac{\partial L_{\text{MAE}}}{\partial p(j|\mathbf{x}_i)} = \begin{cases} -1, & j = y_i \\ 0, & j \neq y_i \end{cases}. \quad (20)$$

B.4. Derivation of GCE

According to Eq. (4), we have

$$L_{\text{GCE}}(\mathbf{x}_i; f_\theta, \mathbf{W}) = \frac{1 - p(y_i|\mathbf{x}_i)^q}{q}. \quad (21)$$

Therefore, we obtain

$$\frac{\partial L_{\text{GCE}}}{\partial p(j|\mathbf{x}_i)} = \begin{cases} -p(y_i|\mathbf{x}_i)^{q-1}, & j = y_i \\ 0, & j \neq y_i \end{cases}. \quad (22)$$

B.5. Derivatives w.r.t. Logits \mathbf{z}_i

B.5.1 $\partial L_{\text{CCE}}/\partial \mathbf{z}_i$

The calculation is based on Eq. (18) and Eq. (16).

If $j = y_i$, we have:

$$\begin{aligned} \frac{\partial L_{\text{CCE}}}{\partial \mathbf{z}_{iy_i}} &= \sum_{j=1}^C \frac{\partial L_{\text{CCE}}}{\partial p(j|\mathbf{x}_i)} \frac{\partial p(y_i|\mathbf{x}_i)}{\mathbf{z}_{ij}} \\ &= p(y_i|\mathbf{x}_i) - 1. \end{aligned} \quad (23)$$

If $j \neq y_i$, it becomes:

$$\begin{aligned} \frac{\partial L_{\text{CCE}}}{\partial \mathbf{z}_{ij}} &= \sum_{j=1}^C \frac{\partial L_{\text{CCE}}}{\partial p(j|\mathbf{x}_i)} \frac{\partial p(y_i|\mathbf{x}_i)}{\mathbf{z}_{ij}} \\ &= p(j|\mathbf{x}_i). \end{aligned} \quad (24)$$

In summary, $\partial L_{\text{CCE}}/\partial \mathbf{z}_i$ can be represented as:

$$\frac{\partial L_{\text{CCE}}}{\partial \mathbf{z}_{ij}} = \begin{cases} p(y_i|\mathbf{x}_i) - 1, & j = y_i \\ p(j|\mathbf{x}_i), & j \neq y_i \end{cases}. \quad (25)$$

B.5.2 $\partial L_{\text{MAE}}/\partial \mathbf{z}_i$

The calculation is analogous with that of $\partial L_{\text{CCE}}/\partial \mathbf{z}_i$.

According to Eq. (20) and Eq. (16), if $j = y_i$:

$$\begin{aligned} \frac{\partial L_{\text{MAE}}}{\partial \mathbf{z}_{iy_i}} &= \sum_{j=1}^C \frac{\partial L_{\text{MAE}}}{\partial p(j|\mathbf{x}_i)} \frac{\partial p(y_i|\mathbf{x}_i)}{\mathbf{z}_{ij}} \\ &= -p(y_i|\mathbf{x}_i)(1 - p(y_i|\mathbf{x}_i)). \end{aligned} \quad (26)$$

otherwise ($j \neq y_i$):

$$\begin{aligned} \frac{\partial L_{\text{MAE}}}{\partial \mathbf{z}_{ij}} &= \sum_{j=1}^C \frac{\partial L_{\text{MAE}}}{\partial p(j|\mathbf{x}_i)} \frac{\partial p(y_i|\mathbf{x}_i)}{\mathbf{z}_{ij}} \\ &= p(y_i|\mathbf{x}_i)p(j|\mathbf{x}_i). \end{aligned} \quad (27)$$

In summary, $\partial L_{\text{MAE}}/\partial \mathbf{z}_i$ is:

$$\frac{\partial L_{\text{MAE}}}{\partial \mathbf{z}_{ij}} = \begin{cases} p(y_i|\mathbf{x}_i)(p(y_i|\mathbf{x}_i) - 1), & j = y_i \\ p(y_i|\mathbf{x}_i)p(j|\mathbf{x}_i), & j \neq y_i \end{cases}. \quad (28)$$

B.5.3 $\partial L_{\text{GCE}}/\partial \mathbf{z}_i$

The calculation is based on Eq. (22) and Eq. (16).

If $j = y_i$, we have:

$$\begin{aligned} \frac{\partial L_{\text{GCE}}}{\partial \mathbf{z}_{iy_i}} &= \sum_{j=1}^C \frac{\partial L_{\text{GCE}}}{\partial p(j|\mathbf{x}_i)} \frac{\partial p(y_i|\mathbf{x}_i)}{\mathbf{z}_{ij}} \\ &= p(y_i|\mathbf{x}_i)^q(p(y_i|\mathbf{x}_i) - 1). \end{aligned} \quad (29)$$

If $j \neq y_i$, it becomes:

$$\begin{aligned} \frac{\partial L_{\text{GCE}}}{\partial \mathbf{z}_{ij}} &= \sum_{j=1}^C \frac{\partial L_{\text{GCE}}}{\partial p(j|\mathbf{x}_i)} \frac{\partial p(y_i|\mathbf{x}_i)}{\mathbf{z}_{ij}} \\ &= p(y_i|\mathbf{x}_i)^q p(j|\mathbf{x}_i). \end{aligned} \quad (30)$$

In summary, $\partial L_{\text{GCE}}/\partial \mathbf{z}_i$ can be represented as:

$$\frac{\partial L_{\text{GCE}}}{\partial \mathbf{z}_{ij}} = \begin{cases} p(y_i|\mathbf{x}_i)^q(p(y_i|\mathbf{x}_i) - 1), & j = y_i \\ p(y_i|\mathbf{x}_i)^q p(j|\mathbf{x}_i), & j \neq y_i \end{cases}. \quad (31)$$

C. Small-scale Fine-grained Visual Categorisation of Vehicles

How does DM perform on small datasets, for example, the number of data points is no more than 5,000? We have tested DM on CIFAR-10 and CIFAR-100 in the main paper. However, both of them contain a training set of 50,000 images.

For this question, we answer it from different perspectives as follows:

1. The problem of label noise on CIFAR-10 and CIFAR-100 in Section 4.2 is of similar scale.

- In Table 2, when noise rate is 80% on CIFAR-10, the number of clean training examples is around $50,000 \times 20\% = 5,000 \times 2$. Therefore, this clean set is only two times as large as 5,000. Beyond, the learning process may be interrupted by other noisy data points.

- In Table 1, when noise rate is 60% on CIFAR-100, the number of clean training data points is about $50,000 \times 40\% = 5,000 \times 4$, i.e., four times as large as 5,000.

2. We compare DM with other standard regularisers on a small-scale fine-grained visual categorisation problem in Table 8.

Vehicles-10 Dataset. In CIFAR-100 [27], there are 20 coarse classes, including vehicles 1 and 2. Vehicles 1 contains 5 fine classes: bicycle, bus, motorcycle, pickup truck, and train. Vehicles 2 includes another 5 fine classes: lawn-mower, rocket, streetcar, tank, and tractor. We build a small-scale vehicle classification dataset composed of these 10 vehicles from CIFAR-100. Specifically, the training set contains 500 images per vehicle class while the testing set has 100 images per class. Therefore, the number of training data points is 5,000 in total.

Table 8: The test accuracy (%) of DM and other standard regularisers on Vehicles-10. We train ResNet-44. Baseline denotes CCE without regularisation. We test two cases: symmetric label noise rate is $r = 40\%$, and clean data $r = 0$.

r	Baseline	L2	Dropout	Dropout+L2	DM	DM+L2	DM+Dropout	DM+L2+Dropout
0	75.4	76.4	77.9	78.7	83.8	84.4	84.5	84.7
40%	42.3	44.8	41.6	47.4	45.8	55.7	48.8	58.1

D. Experimental Details of Robust Sentiment Analysis of Movie Reviews

The network architecture is shown in Figure 4 and its SGD solver is as follows:

```
# configuration of solver.prototxt
net: "nn.prototxt"

# test_iter specifies how many forward passes the test should carry out.
# We have test batch size 250 and 100 test iterations,
# covering the full 25,000 testing reviews.
test_iter: 100
# Carry out testing every 500 training iterations.
test_interval: 500

# The base learning rate, momentum and the weight decay of the network.
base_lr: 0.01
momentum: 0.9
weight_decay: 0.002

# The learning rate policy
lr_policy: "inv"
gamma: 0.00001
power: 0.75

# Display every 100 iterations
display: 100
# The maximum number of iterations
max_iter: 10000
snapshot_prefix: "nn"
# solver mode: CPU or GPU
solver_mode: GPU

random_seed: 123
```

E. More Detailed Empirical Results

E.1. Training DNNs under label noise

Practical research question: What training examples should be focused on and how much more should they be emphasised when training DNNs under label noise?

Our proposal: DM incorporates emphasis mode and variance, and serves as explicit regularisation in terms of example weighting.

Important finding: When label noise rate is higher, we can improve a model’s robustness by moving emphasis mode towards relatively less difficult examples.

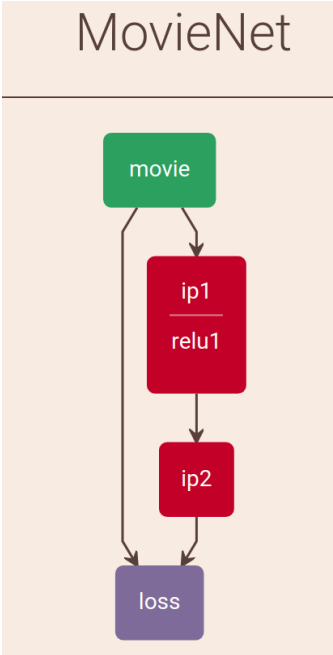


Figure 4: A designed neural network with one 8-neuron hidden layer for sentiment analysis on IMDB. It is quite simple and our objective is not to represent the state-of-the-art.

E.2. Empirical Analysis of DM on CIFAR-10

To empirically understand DM well, we explore the behaviours of DM on CIFAR-10 with $r = 20\%, 40\%, 60\%, 80\%$, respectively. We use ResNet-56 which has larger capacity than ResNet-20.

Design choices. We mainly analyse the impact of different emphasis modes for different noise rates. We explore five emphasis modes: 1) None: $\beta = 0$. There is no emphasis mode since all examples are treated equally; 2) 0: $\lambda = 0$; 3) $\frac{1}{3}$: $\lambda = 0.5$; 4) $\frac{1}{2}$: $\lambda = 1$; 5) $\frac{2}{3}$: $\lambda = 2$. We remark that when λ is larger, the emphasis mode is higher, leading to relatively easier training data points are emphasised. *When emphasis mode changes, emphasis variance changes accordingly.* Therefore, to set a proper spread for each emphasis mode, we try four emphasis variances and choose the best one⁴ to study the impact of emphasis mode.

Results analysis. We show the results in Table 9. The intact training set serves as an indicator of meaningful fitting and we observe that its accuracy is always consistent with the final test accuracy. We display the training dynamics in Figure 5. We summarise our observations as follows:

Fitting and generalisation. We observe that CCE always achieves the best accuracy on corrupted training sets, which indicates that CCE has a strong data fitting ability even if there is severe noise [60]. As a result, CCE has much worse final test accuracy than most models.

Emphasising on harder examples. When there exists label noise, we obtain the worst final test accuracy if emphasis mode is 0, i.e., CCE and DM with $\lambda = 0$. This unveils that in applications where we have to learn from noisy training data, it will hurt the model’s generalisation dramatically if we use CCE or simply focus on harder training data points.

Emphasis mode. When noise rate is 0, 20%, 40%, 60%, and 80%, we obtain the best final test accuracy when $\lambda = 0$, $\lambda = 0.5$, $\lambda = 1$, $\lambda = 2$, and $\lambda = 2$, respectively. This demonstrates that when noise rate is higher, we can improve a model’s robustness by moving emphasis mode towards relatively less difficult examples with a larger λ , which is informative in practice.

Emphasis spread. As displayed in Table 9 and Figures 6-9 in the supplementary material, emphasis variance also matters a lot when fixing emphasis mode, i.e., fixing λ . For example in Table 9, when $\lambda = 0$, although focusing on harder examples similarly with CCE, DM can outperform CCE by modifying the emphasis variance. As shown in Figures 6-9, some models

⁴Since there is a large interval between different β in our four trials, we deduce that the chosen one is not the optimal. The focus of this work is not to optimize the hyper-parameters. Instead, we focus more on the practical research question: *What training examples should be focused on and how much more should they be emphasised when training DNNs under label noise?*

Table 9: Results of CCE, DM on CIFAR-10 under noisy labels. For every model, we show its best test accuracy during training and the final test accuracy when training terminates, which are indicated by ‘Best’ and ‘Final’, respectively. We also present the results on corrupted training sets and original intact one. The overlap rate between corrupted and intact sets is $(1 - r)$. When λ is larger, β should be larger for adjusting emphasis variance. The intact training set serves as an indicator of meaningful fitting and we observe that its accuracy is always consistent with the final test accuracy.

Noise Rate r	Emphasis Mode	Model	Testing Accuracy (%)		Accuracy on Training Sets (%)	
			Best	Final	Corrupted	Intact (Meaningful Fitting)
20%	0	CCE	86.5	76.8	95.7	80.6
	None	DM ($\beta=0$)	83.5	58.1	50.6	60.2
	0 ($\lambda = 0$)	DM ($\beta = 2$)	84.9	76.4	85.3	80.5
	1/3($\lambda = 0.5$)	DM ($\beta = 12$)	89.4	87.8	81.5	95.0
	1/2($\lambda = 1$)	DM ($\beta = 16$)	87.3	86.7	78.4	93.8
	2/3($\lambda = 2$)	DM ($\beta = 24$)	85.8	85.5	76.0	91.4
40%	0	CCE	82.8	60.9	83.0	64.4
	None	DM ($\beta=0$)	71.8	44.9	31.3	45.8
	0 ($\lambda = 0$)	DM ($\beta = 1$)	78.4	65.6	63.3	66.6
	1/3($\lambda = 0.5$)	DM ($\beta = 12$)	85.1	79.9	67.7	85.7
	1/2($\lambda = 1$)	DM ($\beta = 16$)	84.7	83.3	60.3	88.9
	2/3($\lambda = 2$)	DM ($\beta = 20$)	52.7	52.7	35.4	53.6
60%	0	CCE	69.5	37.2	84.1	40.5
	None	DM ($\beta=0$)	69.9	57.9	40.1	58.6
	0 ($\lambda = 0$)	DM ($\beta = 0.5$)	72.3	53.9	42.1	55.1
	1/3($\lambda = 0.5$)	DM ($\beta = 12$)	77.5	58.5	55.5	62.6
	1/2($\lambda = 1$)	DM ($\beta = 12$)	71.9	70.0	41.0	73.9
	2/3($\lambda = 2$)	DM ($\beta = 12$)	80.2	72.5	44.9	75.4
80%	0	CCE	36.1	16.1	54.3	18.4
	None	DM ($\beta=0$)	44.4	28.2	20.6	28.8
	0 ($\lambda = 0$)	DM ($\beta = 0.5$)	46.2	21.3	27.8	23.1
	1/3($\lambda = 0.5$)	DM ($\beta = 8$)	51.6	22.4	46.1	24.4
	1/2($\lambda = 1$)	DM ($\beta = 8$)	35.5	31.5	19.8	32.3
	2/3($\lambda = 2$)	DM ($\beta = 12$)	33.0	32.8	14.2	32.6

even collapse and cannot converge if the emphasis variance is not rational.

E.3. Learning Dynamics on Clean Training Data

The learning dynamics on clean training data are displayed in the Figures 10-11.

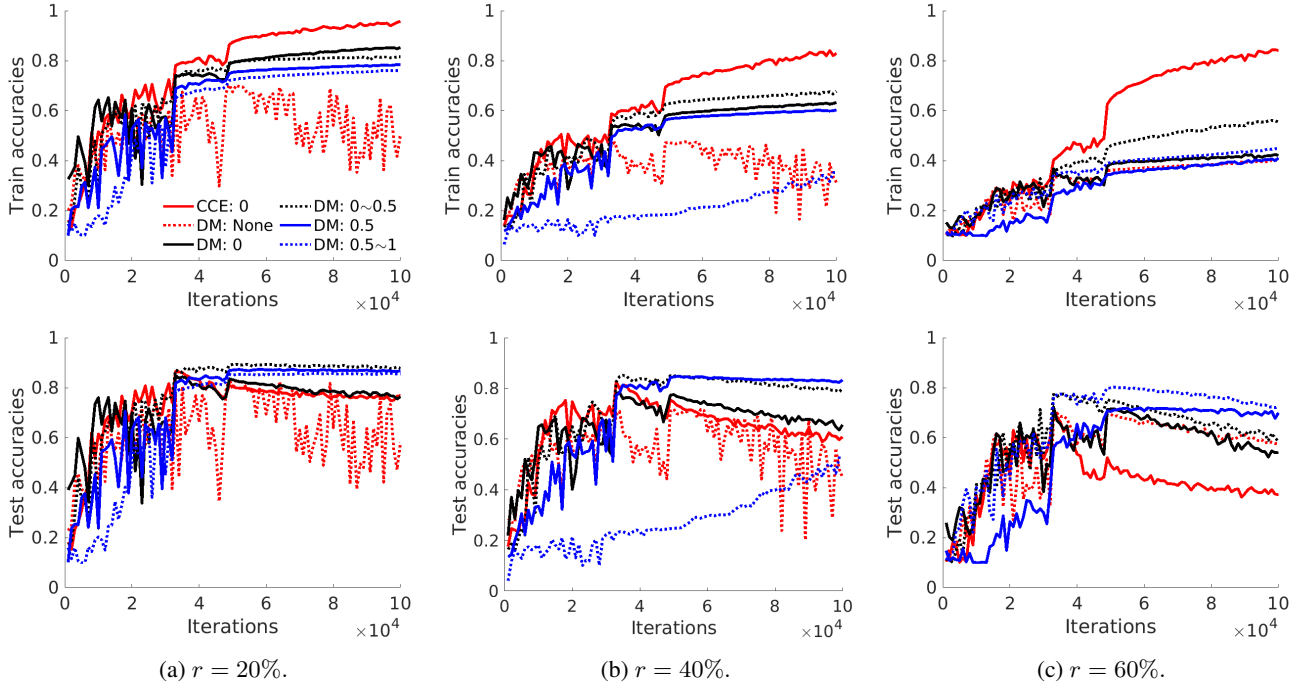


Figure 5: The learning dynamics of ResNet-56 on CIFAR-10, i.e., training and testing accuracies along with training iterations. The legend in the top left is shared by all subfigures. 'xxx: yyy' means 'method: emphasis mode'. We have two key observations: 1) When noise rate increases, better generalisation is obtained with higher emphasis mode, i.e., focusing on relatively easier examples; 2) Both overfitting and underfitting lead to bad generalisation. For example, 'CCE: 0' fits training data much better than the others while 'DM: None' generally fits it unstably or a lot worse. Better viewed in colour.

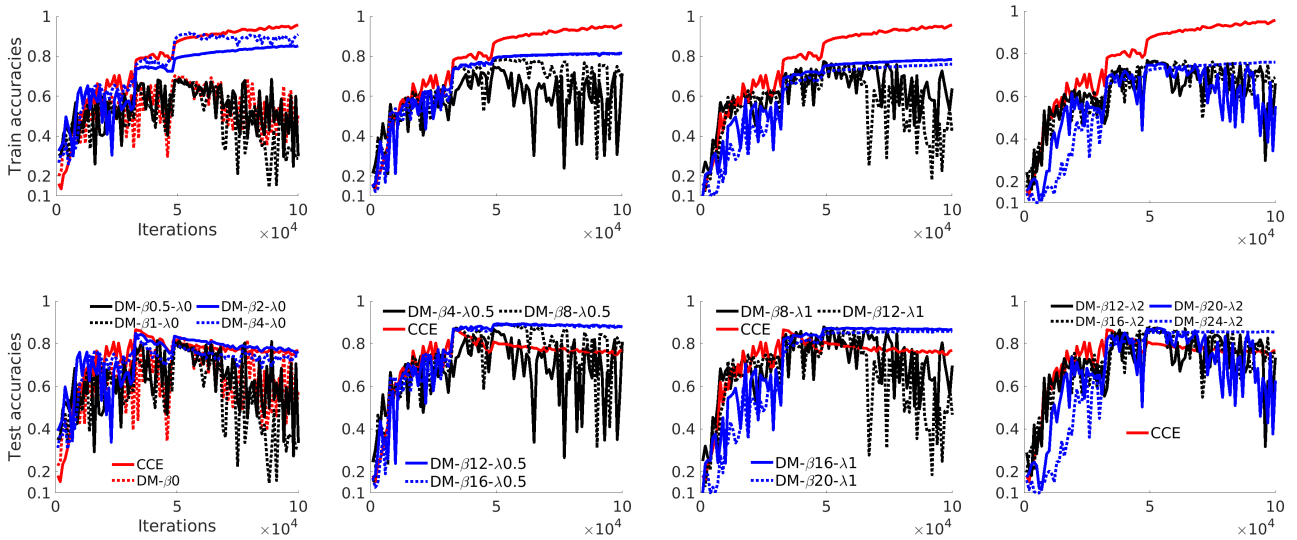


Figure 6: ResNet-56 on CIFAR-10 ($r = 20\%$). From left to right, the results of four emphasis modes 0, $\frac{1}{3}$, 0.5, $\frac{2}{3}$ with different emphasis variances are displayed in each column respectively. When λ is larger, β should be larger. Specifically :

- 1) when $\lambda = 0$: we tried $\beta = 0.5, 1, 2, 4$;
- 2) when $\lambda = 0.5$: we tried $\beta = 4, 8, 12, 16$;
- 3) when $\lambda = 1$: we tried $\beta = 8, 12, 16, 20$;
- 4) when $\lambda = 2$: we tried $\beta = 12, 16, 20, 24$.

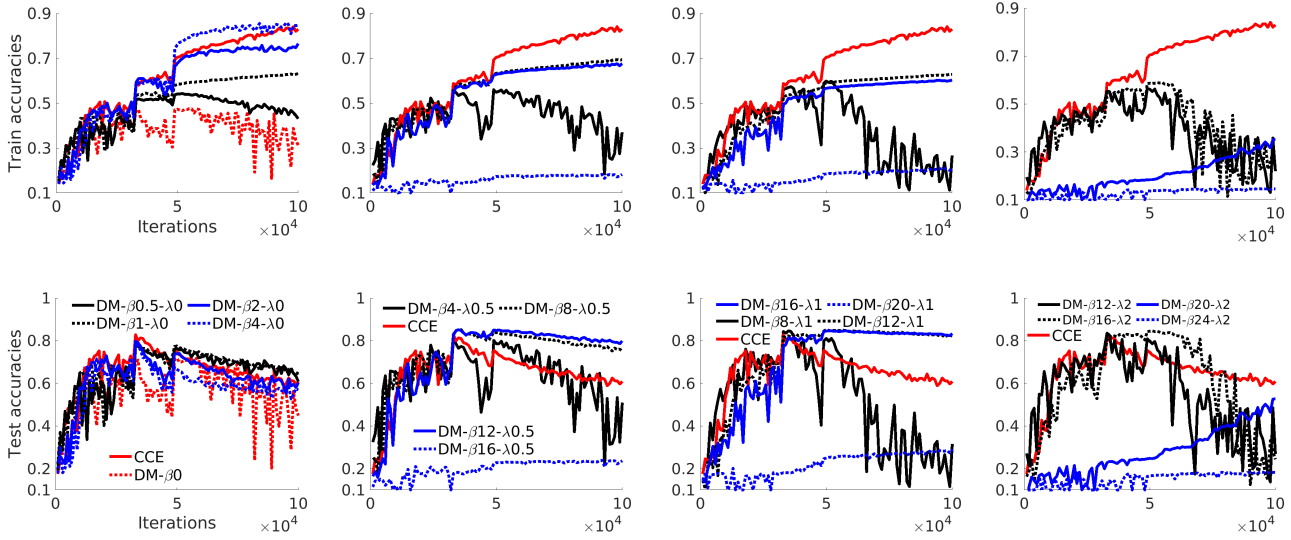


Figure 7: ResNet-56 on CIFAR-10 ($r = 40\%$). From left to right, the results of four emphasis modes $0, \frac{1}{3}, 0.5, \frac{2}{3}$ with different emphasis variances are displayed in each column respectively. When λ is larger, β should be larger. Specifically :

- 1) when $\lambda = 0$: we tried $\beta = 0.5, 1, 2, 4$;
- 2) when $\lambda = 0.5$: we tried $\beta = 4, 8, 12, 16$;
- 3) when $\lambda = 1$: we tried $\beta = 8, 12, 16, 20$;
- 4) when $\lambda = 2$: we tried $\beta = 12, 16, 20, 24$.

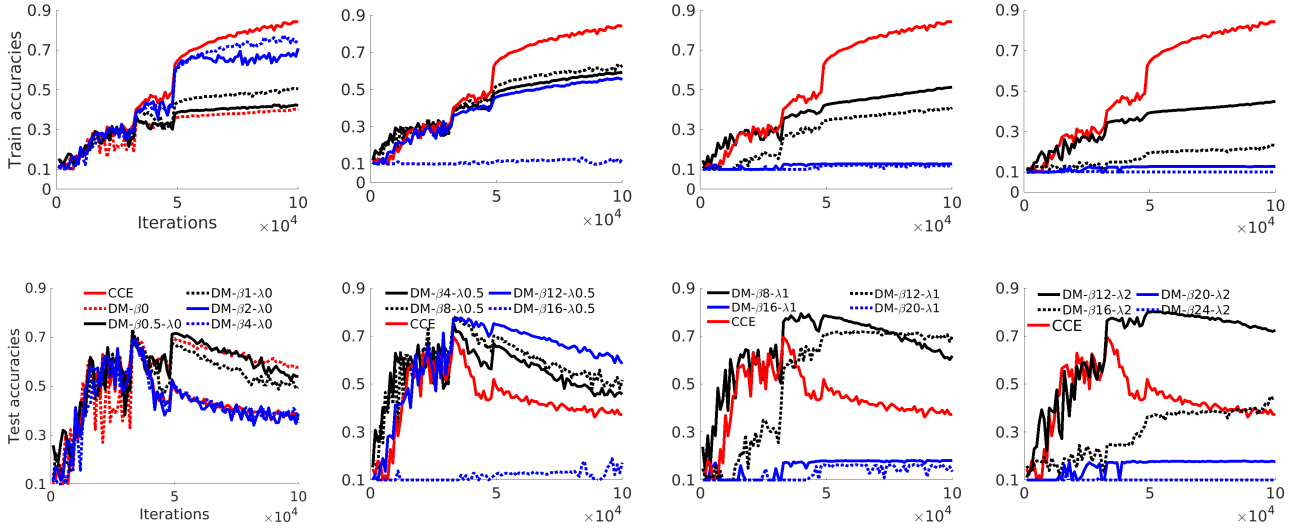


Figure 8: ResNet-56 on CIFAR-10 ($r = 60\%$). From left to right, the results of four emphasis modes $0, \frac{1}{3}, 0.5, \frac{2}{3}$ with different emphasis variances are displayed in each column respectively. When λ is larger, β should be larger. Specifically :

- 1) when $\lambda = 0$: we tried $\beta = 0.5, 1, 2, 4$;
- 2) when $\lambda = 0.5$: we tried $\beta = 4, 8, 12, 16$;
- 3) when $\lambda = 1$: we tried $\beta = 8, 12, 16, 20$;
- 4) when $\lambda = 2$: we tried $\beta = 12, 16, 20, 24$.

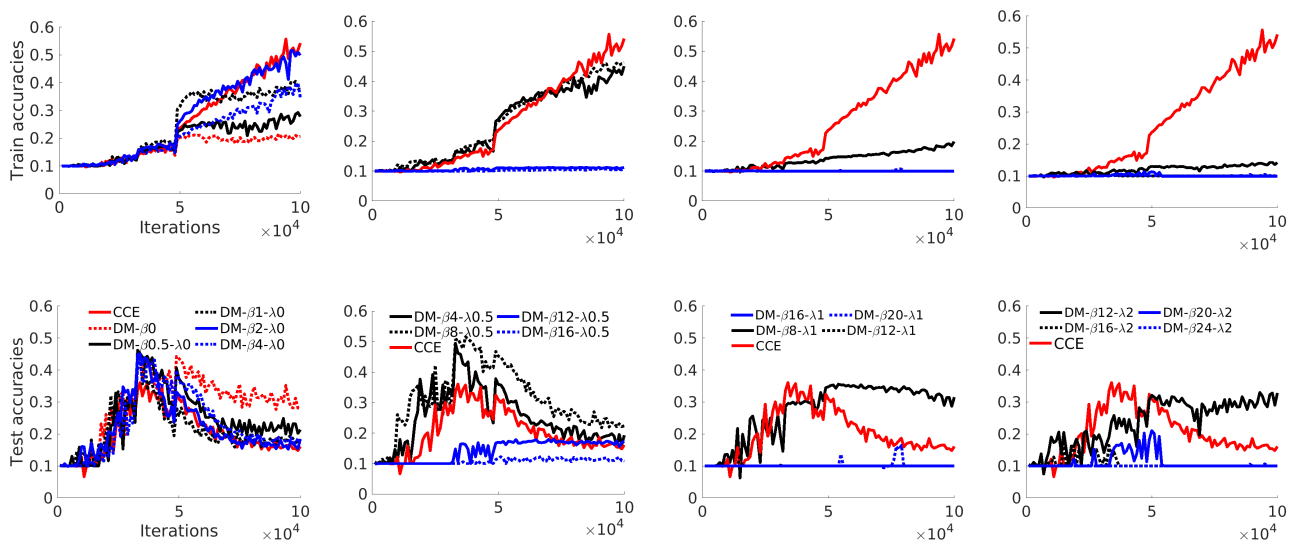


Figure 9: ResNet-56 on CIFAR-10 ($r = 80\%$). From left to right, the results of four emphasis modes 0, $\frac{1}{3}$, 0.5, $\frac{2}{3}$ with different emphasis variances are displayed in each column respectively. When λ is larger, β should be larger. Specifically :

- 1) when $\lambda = 0$: we tried $\beta = 0.5, 1, 2, 4$;
- 2) when $\lambda = 0.5$: we tried $\beta = 4, 8, 12, 16$;
- 3) when $\lambda = 1$: we tried $\beta = 8, 12, 16, 20$;
- 4) when $\lambda = 2$: we tried $\beta = 12, 16, 20, 24$.

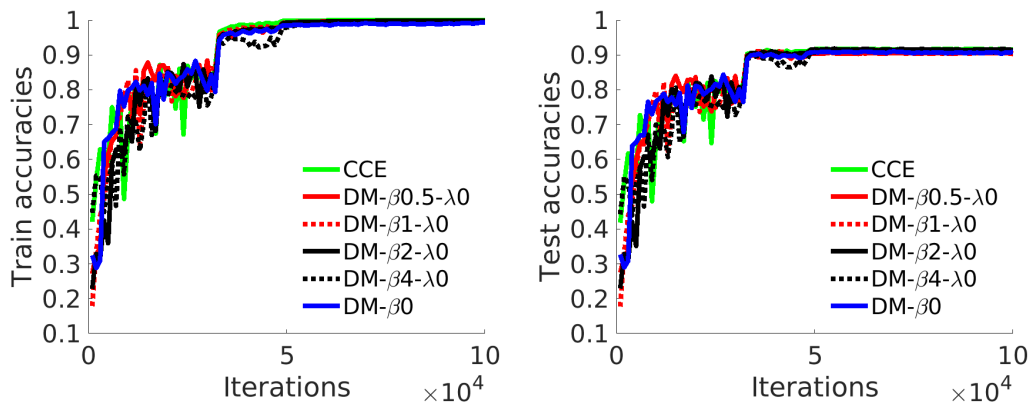


Figure 10: The training and test accuracies on clean CIFAR-10 along with training iterations. The training labels are clean. We fix $\lambda = 0$ to focus on harder examples while changing emphasis variance controller β . The backbone is ResNet-20. The results of ResNet-56 are shown in Figure 11.

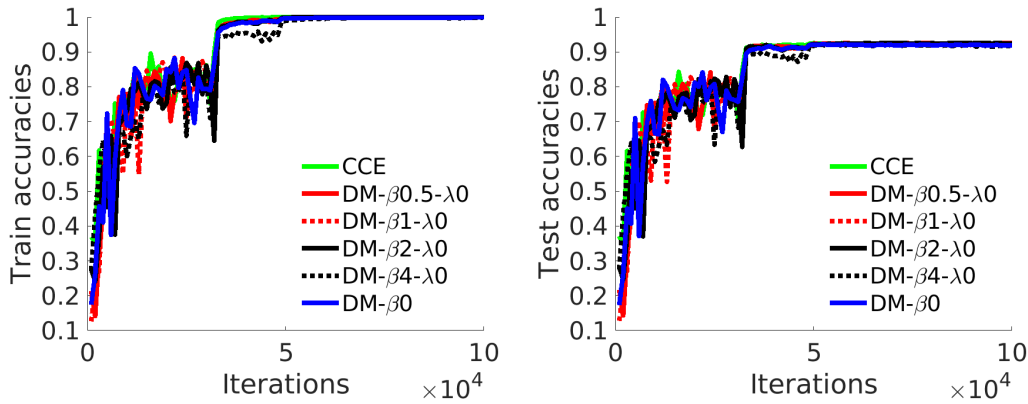


Figure 11: The training and test accuracies on clean CIFAR-10 along with training iterations. The training labels are clean. We fix $\lambda = 0$ to focus on more difficult examples while changing emphasis variance controller β . The backbone is ResNet-56. The results of ResNet-20 are shown in Figure 10.