ENERGY ALIGNMENT FOR BIAS RECTIFICATION IN CLASS INCREMENTAL LEARNING

Bowen Zhao[†], *Chen Chen*[*,✉], *Xi Xiao*[†,‡], *Qi Ju*[*], *Shutao Xia*[†,‡,✉]

[†]Tsinghua University, [*]Tencent TEG AI, [‡]Peng Cheng Laboratory

ABSTRACT

In class incremental learning (CIL), models are expected to be able to learn new categories continuously. However, the standard DNNs suffer from catastrophic forgetting. Recent studies show class imbalance is an essential factor that causes catastrophic forgetting in CIL. In this paper, from the perspective of energy-based model, we demonstrate that the free energies of categories are aligned with the label distribution theoretically, thus the energies of different classes are expected to be close to each other when aiming for “balanced” performance. However, we discover a severe energy-bias phenomenon in the models trained in CIL. To eliminate the bias, we propose a simple and effective method named Energy Alignment by merely adding the calculated shift scalars onto the output logits, which does not require to (i) modify the network architectures, (ii) intervene the standard learning paradigm. Experimental results show that energy alignment can achieve good performance on several CIL benchmarks.

Index Terms— Class Incremental Learning, Energy, Imbalance

1. INTRODUCTION

In class incremental learning (CIL), an ideal intelligent system is expected to be able to learn new categories from streaming data without forgetting old categories that they have mastered. When new data comes, the whole of (or the most of) old data can not be seen again in CIL. In such scenario, Deep neural networks (DNNs) undergo a serious problem known as catastrophic forgetting [1, 2], which reveals that DNNs can hardly recollect old knowledge after learning new knowledge.

Standard DNNs are usually reliable and powerful when the training data is representative of the evaluation data, i.e., the joint distribution of the observed training data $p_o(x, y)$ is consistent with that of the test data $p_t(x, y)$ (x represents the data variable, and y stands for the label variable). However, the ideal condition $p_o(x, y) = p_t(x, y)$ is not satisfied in CIL. As even with the rehearsal strategy (which uses a very small amount of old data to complement training data), the class imbalance problem between the data volume of old-classes and new-classes is severe. The distribution shift $p_o(y) \neq p_t(y)$ leads to a biased model, which shows a strong tendency towards the majority classes. This is proved to be a vital factor for catastrophic forgetting in CIL [3, 4, 5].

This paper endeavours to rectify the biased models in class incremental learning from the theoretical perspective of energy-based models (EBMs) [6], leading to a simple and effective approach called Energy Alignment. EBMs assign a scalar energy to each configuration of the variables (e.g., the pair (x, y)) — low energies for compatible configurations while high energies for incompatible configurations conventionally. In this work, we reveal that the free energy of a specific category is theoretically aligned with the probability density of the class. Therefore, under the goal of pursuing “balanced” performance, the free energies of different classes are potentially expected to be similar since the target label distribution $p_t(y)$ is actually uniform for each class. However, as to a model trained on imbalanced data, we discover that a severe bias is hidden in the free energies: namely the minority classes commonly possess higher free energies while the majority classes usually hold lower free energies. Consequently, we put forward an algorithm — Energy Alignment (EA) to combat the bias. EA attempts to align the free energies of the minority classes to those of the majority classes, which is simply achieved by correcting the predicted logits with the theoretically calculated shift scalars. We delineate that energy alignment can be applied to a purely discriminative classification model, meaning that one can easily obtain a corrected model without modifying the network architectures or altering the standard training procedure. We ameliorate the heavily biased models in CIL by the proposed energy alignment. Based on the rehearsal [7] and distillation [8, 9] strategies, we rectify the current model with energy alignment after training of each incremental step. Experimental results demonstrate the effectiveness of energy alignment for class incremental learning.

Our key contributions can be summarized as follows:

- We alleviate catastrophic forgetting in class incremental learning by address class imbalance problem from the perspective of energy-based models.
- We interpret the bias in models trained on imbalanced data from the view of energy values. Starting from this observation, we propose energy alignment for correcting biased models, which is not only a simple but also a principled approach.
- We demonstrate the effectiveness of energy alignment on class incremental learning benchmarks.

2. RELATED WORK

Class Incremental Learning. To alleviate catastrophic forgetting in incremental learning, a number of strategies have been proposed recently. A group of methods alleviate catastrophic forgetting through parameter control [10, 11, 12], which attempts to keep the parameters that are important to old knowledge unchanged. Knowledge distillation [9, 7, 13] is also adopted to remit catastrophic forgetting. However, most of them are hard to manage the scenario of class incremental learning [14], while the simple rehearsal strategy has been demonstrated to be effective in CIL [14, 15]. Based on the rehearsal strategy, another series of studies, like BiC [4], IL2M [16], WA [5] and ScaIL [17], view class imbalance as an essential factor that causes catastrophic forgetting in class incremental learning, and

[✉]corresponding authors; email: zbw18@mails.tsinghua.edu.cn, beckhamchen@tencent.com, xiast@sz.tsinghua.edu.cn; work done by bowen during internship at tencent; this work is supported in part by the National Natural Science Foundation of China (62171248,61972219), the R&D Program of Shenzhen (JCYJ20180508152204044, JCYJ20190813174403598, SGDX20190918101201696), the PCNL KEY project (PCL2021A07), and the Overseas Research Cooperation Fund of Tsinghua Shenzhen International Graduate School (HW2021013).

put forward different solutions to address it. This work is also based on the rehearsal strategy and distillation strategy, but attempts to rectify the bias in class incremental learning from the perspective of energy-based models.

Energy-Based Models and Discriminative models. Considering a model with two sets of variables x (e.g., images) and y (e.g., labels) ($x \in \mathcal{X}$, $y \in \mathcal{Y}$, $\mathcal{X}$ and $\mathcal{Y}$ denote the whole space of x and y, respectively), energy-based models (EBMs) capture compatibility by associating a non-probabilistic, scalar energy to each configuration of the variables [6]. Generally, EBMs build an energy function $E_\theta(x, y)$ that maps each configuration (x, y) to an energy value, where θ is a trainable parameter set. Conventionally, the small energies represent highly compatible configurations of the variables (when y is suitable for x), while large energies signify highly incompatible configurations of the variables (when y is incorrect for x). The energies are usually required to be normalized in the scenario that the model is going to be used in combination with others. The most common approach for this purpose is utilizing the Boltzmann distribution to turn the collection of energy values into a probability density:

$$p_\theta(y|x) = \frac{e^{-E_\theta(x,y)}}{\int_{y' \in \mathcal{Y}} e^{-E_\theta(x,y')}}. \tag{1}$$

Recently, a classification task with C classes is typically optimized with a discriminative deep neural network $f_\theta(x)$ where θ means the learnable parameter set, by which the input variable x is mapped to C scalar values, known as logits. Then, these logits are employed to parameterize a posterior probability distribution over C classes via the softmax function:

$$p_\theta(y|x) = \frac{e^{f_\theta(x)[y]}}{\sum_{y'=1}^{C} e^{f_\theta(x)[y']}}, \tag{2}$$

where $f_\theta(x)[y]$ represents the y^{th} index of $f_\theta(x)$.

Compared Eq. (2) with Eq. (1), EBMs are inherently connected with discriminative models [18, 19]. Without changing the parameterization of the neural network $f_\theta(\cdot)$, we can re-use the logits to represent an energy of the configuration (x, y) [19] as

$$E_\theta(x, y) = -f_\theta(x)[y]. \tag{3}$$

In fact, the negative logit $-f_\theta(x)[y]$ does reflect the degree of compatibility between x and y — small value of $-f_\theta(x)[y]$ corresponds to a "good" configuration while large value of $-f_\theta(x)[y]$ corresponds to a "bad" one, which is exactly what energy-based models require.

3. ENERGY ALIGNMENT

Bias in Energies. Similar to Eq. (1), for EBMs, the joint distribution of x and y can be formulated by using the Boltzmann distribution again, which is expressed as

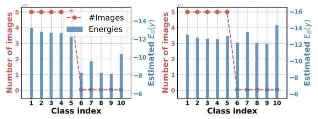
$$p_\theta(x, y) = \frac{e^{-E_\theta(x,y)}}{Z_\theta}, \quad Z_\theta = \int_{x' \in \mathcal{X}, y' \in \mathcal{Y}} e^{-E_\theta(x',y')}. \tag{4}$$

Then, by marginalizing the joint distribution Eq. (4) over x, we obtain

$$p_\theta(y) = \frac{\int_{x' \in \mathcal{X}} e^{-E_\theta(x',y)}}{Z_\theta} = \frac{e^{-E_\theta(y)}}{Z_\theta}, \tag{5}$$

where the Helmholtz free energy $E_\theta(y) = -\log \int_{x' \in \mathcal{X}} e^{-E_\theta(x',y)}$. Then, by taking the logarithm of both sides of Eq. (5), we obtain

$$\log p_\theta(y) = -E_\theta(y) - \log Z_\theta. \tag{6}$$



(a) raw (b) corrected

Fig. 1: Number of training images per class and estimated free energy per class.

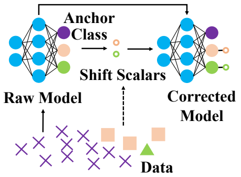


Fig. 2: Framework of energy alignment for biased models.

As shown in Eq. (6), the negative free energy $-E_\theta(y)$ is linearly aligned with $\log p_\theta(y)$, i.e., $-E_\theta(y) \propto \log p_\theta(y)$. When it comes to the discriminative neural network $f_\theta(\cdot)$, combined with Eq. (3), we obtain the negative free energy with model $f_\theta(\cdot)$: $-E_\theta(y) = \log \int_{x' \in \mathcal{X}} e^{f_\theta(x')[y]}$.

Due to class imbalance, the label distribution of the observed training set is different from that of the target (test) set, which are written as $p_o(y)$ and $p_t(y)$ respectively. Though the observed training data is imbalanced, it is still expected to achieve balanced predictions and superior overall results during test, alluding a hidden condition that the target label distribution $p_t(y)$ is exactly uniform for all classes of interest. Ideally, the model parameterized with θ is expected to reflect the target label distribution, i.e., $p_\theta(y) \approx p_t(y)$. Therefore, it is desirable to have

$$p_\theta(y = i) = p_\theta(y = j), 1 \leq i \leq C, 1 \leq j \leq C, i \neq j. \tag{7}$$

Moreover, combined with Eq. (6), we expect to have:

$$-E_\theta(y = i) = -E_\theta(y = j), 1 \leq i \leq C, 1 \leq j \leq C, i \neq j. \tag{8}$$

Unfortunately, the model trained on imbalanced data actually tends to manifest the label distribution of the training data set, i.e., $p_o(y)$ instead of $p_t(y)$. We conduct an experiment on the class-imbalanced version of CIFAR10 with label distribution of its training data illustrated in Fig. 1a. Based on the learned model, we estimate the free energies of the ten categories, bringing to light that the majority classes tend to have lower energies whereas the minority classes usually hold higher energies as shown in Fig. 1a. This phenomenon does not conform to Eq. (8), exposing the bias in the trained model from the perspective of energy-based learning.

Energy Alignment. The above analysis have pointed out that the prior label distribution $p_t(y)$ hidden in the goal is actually uniform for classes of interest. However, the free energies $E_\theta(y)$ of the model trained on imbalanced data does not match the prior knowledge (Eq. (8)). Instead of intervening the training process of $f_\theta(\cdot)$, we propose a simple and effective method, called Energy Alignment (EA), to correct the biased model directly. In EA, the logits are refined to make the energies satisfy Eq. (8).

Without loss of generality, considering the free energies of the i^{th} and j^{th} class ($1 \leq i \leq C, 1 \leq j \leq C, i \neq j$), we view the i^{th} category as the "anchor class" and rebalance the two free energies with a shift scalar α_j, which is formulated as

$$\begin{aligned} & -E_\theta(y = i) = -E_\theta(y = j) + \alpha_j \\ \Rightarrow \quad & \log \int_{x' \in \mathcal{X}} e^{f_\theta(x')[i]} = \alpha_j + \log \int_{x' \in \mathcal{X}} e^{f_\theta(x')[j]} \\ \Rightarrow \quad & \alpha_j = \log \frac{\int_{x' \in \mathcal{X}} e^{f_\theta(x')[i]}}{\int_{x' \in \mathcal{X}} e^{f_\theta(x')[j]}}. \end{aligned} \tag{9}$$

Now, assuming that the shift scalar α_j has been calculated (will be described later), the right side of Eq. (9) can be reformulated as

$$
\begin{aligned}
-E_\theta(y=j)+\alpha_j &= \alpha_j + \log \int_{x' \in \mathcal{X}} e^{f_\theta(x')[j]} \\
&= \log \int_{x' \in \mathcal{X}} e^{f_\theta(x')[j]+\alpha_j}.
\end{aligned} \tag{10}
$$

As shown in Eq. (10), for any input data point, the j^{th} logit can be adjusted by adding a non-learned scalar α_j, leading to the advantage that the negative free energies $-E_\theta(y=i)$ and $-E_\theta(y=j)$ of the i^{th} and j^{th} classes could be equalized. Intuitively, when $\int_{x' \in \mathcal{X}} e^{f_\theta(x')[i]} > \int_{x' \in \mathcal{X}} e^{f_\theta(x')[j]}$, α_j is positive, then the logit $f_\theta(x')[j]$ is augmented by α_j; if $\int_{x' \in \mathcal{X}} e^{f_\theta(x')[i]} < \int_{x' \in \mathcal{X}} e^{f_\theta(x')[j]}$, α_j is negative, then the logit $f_\theta(x')[j]$ is attenuated by α_j.

In terms of any other category j $(j \neq i)$, we can adjust the j^{th} logit with corresponding scalar α_j to satisfy Eq. (8) similarly, hence the model with energy alignment would treat all categories equally and make Eq. (7), Eq. (8) hold. The corrected model is written as $f_{\theta;\{\alpha_j\}_{j=1}^{C-1}}$ here. The proposed energy alignment strategy is illustrated in Fig. 2. As an example, for the above experiment, the negative free energies of the corrected model is depicted in Fig. 1b, which shows that energy alignment can indeed balance the energies.

Approximation of Shift Scalar α_j. In order to acquire α_j, the remaining essential problem is how to calculate the terms $\int_{x' \in \mathcal{X}} e^{f_\theta(x')[i]}$ and $\int_{x' \in \mathcal{X}} e^{f_\theta(x')[j]}$. Avoiding directly computing the intractable integrals, in this work, we employ Monte Carlo integration [20, 21] to estimate the integrals. Specifically, for the i^{th} category, the integral is approximated by $\int_{x' \in \mathcal{X}} e^{f_\theta(x')[i]} \approx \int_{x' \in \mathcal{X}'} e^{f_\theta(x')[i]} \approx \frac{1}{S}\sum_{s=1}^{S} \frac{e^{f_\theta(x_s)[i]}}{q(x_s)}$, where $\mathcal{X}' \subset \mathcal{X}$ is the input space of categories of interest. Since $e^{f_\theta(x)[i]}$ with the trained θ usually outputs very low response to samples that are almost unrelated to the categories of interest, it suffices to estimate the integral on the subspace $\mathcal{X}'$. Data points $\{x_s\}_{s=1}^{S}$ are sampled from the proposal distribution $q(x)$, which is assumed to assign uniform values to samples from the input space $\mathcal{X}'$ (i.e., $q(x)$ is a uniform distribution). The same is true for another integral $\int_{x' \in \mathcal{X}} e^{f_\theta(x')[j]}$. Then, we have

$$
\alpha_j \approx LSE_{s=1}^{S}(f_\theta(x_s)[i]) - LSE_{s=1}^{S}(f_\theta(x_s)[j]), \tag{11}
$$

where $LSE_{s=1}^{S}(z_s) = \log \sum_{s=1}^{S} e^{z_s}$. To mitigate the approximation error, we divide the categories into M clusters within which the same shift scalar is shared, based on the number of training samples per class. Specifically, for class clusters $\mathcal{O}_i$ and $\mathcal{O}_j$ with C_i and C_j categories respectively, similar to Eq. (9), Eq. (10) and Eq. (11), we view cluster $\mathcal{O}_i$ as the "anchor cluster" and rebalance the average energies for the two clusters with a shared shift scalar α_j. The estimation of α_j is given by

$$
\alpha_j \approx \frac{1}{C_i}\sum_{i \in \mathcal{O}_i} LSE_{s=1}^{S}(f_\theta(x_s)[i]) - \frac{1}{C_j}\sum_{j \in \mathcal{O}_j} LSE_{s=1}^{S}(f_\theta(x_s)[j]). \tag{12}
$$

Moreover, one can still adjust the logits of classes in cluster $\mathcal{O}_j$ by the shared α_j directly for any input data to obtain a corrected model, written as $f_{\theta;\{\alpha_j\}_{j=1}^{M-1}}$.

Algorithm 1 Class Incremental Learning with Energy Alignment

```
Input: Training data $\{\mathcal{D}^{(b)}\}_{b=1}^{B}$ of $B$ incremental steps
Initialization: Model $f_\theta$
$f_\theta \leftarrow Train(f_\theta, \mathcal{D}^{(1)})$    ▷ Train with loss Eq. (14) (no rehearsal data)
$f_{\theta_t;\alpha} \leftarrow DetachCopy(f_\theta)$    ▷ Save as the teacher model for the next incremental step; a dummy shift scalar $\alpha$ is added here for the convenience of the following statement
$\mathcal{D}_{re}^{(2)} \leftarrow RandomSample(\mathcal{D}^{(1)})$    ▷ Select rehearsal samples for the next incremental step
for $b \in 2, \cdots, B$ do    ▷ For each incremental step
    $f_\theta \leftarrow Train(f_\theta, f_{\theta_t;\alpha}, \mathcal{D}^{(b)}, \mathcal{D}_{re}^{(b)})$    ▷ Train with loss Eq. (13)
    $\mathcal{D}_{ea} \leftarrow RandomSample(\mathcal{D}^{(b)}, \mathcal{D}_{re}^{(b)})$    ▷ Select samples for energy alignment
    $\alpha \leftarrow ShiftScalar(f_\theta, \mathcal{D}_{ea})$    ▷ Calculate shift scalar by Eq. (12)
    $f_{\theta_t;\alpha} \leftarrow DetachCopy(Correct(f_\theta, \alpha))$    ▷ Save the corrected model as teacher model used in the next step
    $\mathcal{D}_{re}^{(b+1)} \leftarrow RandomSample(\mathcal{D}^{(b)}, \mathcal{D}_{re}^{(b)})$    ▷ Select rehearsal samples from current new data and rehearsal data
end for
Output: Model $f_{\theta_t;\alpha}$
```

Class Incremental Learning with Energy Alignment. We present the learning process of class incremental learning with energy alignment in Alg. 1. Assuming that B batches of training data $\{\mathcal{D}^{(b)}\}_{b=1}^{B}$ from different categories are available gradually in which, for the b^{th} incremental step, the new training data $\mathcal{D}^{(b)}$ comes from $C^{(b)}$ new classes, and the rehearsal data $\mathcal{D}_{old}^{(b)}$ (which is selected from the previous data $\{\mathcal{D}^{(1)}, \cdots, \mathcal{D}^{(b-1)}\}$) comes from $C_{old}^{(b)}$ classes, where $C_{old}^{(1)} = 0$ and $C_{old}^{(b)} = \sum_{k=1}^{b-1} C^{(k)}, b > 1$. The number of rehearsal samples in each incremental step (except the first step, as there is no old data in the initial step) is constant, i.e., $|\mathcal{D}_{old}^{(1)}| = 0$, $|\mathcal{D}_{old}^{(2)}| = |\mathcal{D}_{old}^{(3)}| = \cdots = |\mathcal{D}_{old}^{(B)}|$, so that as more categories are encountered, the number of rehearsal samples per old class decreases, and the problem of class imbalance becomes more serious. The model $f_\theta(\cdot)$ is trained with a compound loss:

$$
\mathcal{L}_{CIL}(x, y^*) = (1-\lambda)\mathcal{L}_{CIL-CE}(x, y^*) + \lambda\mathcal{L}_{CIL-KD}(x), \tag{13}
$$

where y^* denotes the true label, $(x, y^*) \in \mathcal{D}^{(b)} \cup \mathcal{D}_{old}^{(b)}$, λ is used to balance the two losses which is set to $\lambda_{base} \cdot \frac{C_{old}^{(b)}}{C^{(b)}+C_{old}^{(b)}}$ with a hyper-parameter λ_{base}. The cross-entropy loss is defined as

$$
\mathcal{L}_{CIL-CE}(x, y^*) = \sum_{i=1}^{C^{(b)}+C_{old}^{(b)}} -\mathbb{I}_{i=y^*} \log(p_\theta(i|x)), \tag{14}
$$

where $(x, y^*) \in \mathcal{D}^{(b)} \cup \mathcal{D}_{old}^{(b)}$, $\mathbb{I}_{i=y^*}$ is the indicator function and $p_\theta(i|x)$ is the predicted probability of the i^{th} class defined as Eq. (2). The knowledge distillation loss is defined as

$$
\mathcal{L}_{CIL-KD}(x) = \sum_{i=1}^{C_{old}^{(b)}} -\hat{q}_{\theta_t;\alpha}(i|x) \log(q_\theta(i|x)), \tag{15}
$$

where $\hat{q}_{\theta_t;\alpha}(i|x) = \frac{e^{f_{\theta_t;\alpha}(x)[i]/T}}{\sum_{j=1}^{C_{old}^{(b)}} e^{f_{\theta_t;\alpha}(x)[j]/T}}$, $q_\theta(i|x) = \frac{e^{f_\theta(x)[i]/T}}{\sum_{j=1}^{C_{old}^{(b)}} e^{f_\theta(x)[j]/T}}$, $x \in \mathcal{D}^{(b)} \cup \mathcal{D}_{old}^{(b)}$, T represents temperature.

During class incremental learning, in each incremental step, the old classes form the "old cluster" which is viewed as "anchor cluster", and the new classes form the "new cluster" (i.e., the number of clusters $M = 2$), then, a shift scalar α can be calculated for the "new cluster". We perform energy alignment with the shift scalar and obtain the corrected model $f_{\theta_t;\alpha}(x)$, which is also served as a teacher model in the next incremental step as shown in Alg. 1.

#step	1	2	3	4	5	6	7	8	9	10	Avg
LwF [9]	99.2	95.4	86.2	74.1	63.9	55.1	50.3	44.5	40.4	36.6	60.7
iCaRL [7]	99.5	97.8	94.1	91.8	88.0	82.7	77.3	73.2	67.3	63.8	81.8
EEIL [22]	99.4	**99.0**	**96.4**	93.8	90.4	88.8	86.6	84.9	82.2	80.2	89.2
BiC [4]	98.4	96.2	94.0	92.9	91.1	89.4	88.1	86.5	85.4	84.4	89.8
RPS [23]	99.4	97.4	94.2	92.6	89.4	86.2	83.7	82.1	79.5	74.0	86.6
WA [5]	98.8	96.8	94.5	93.1	90.5	89.9	88.8	88.0	86.2	84.1	90.2
EA (Ours)	99.1 ±0.3	97.5 ±0.5	95.5 ±0.4	**94.0** ±0.4	**91.3** ±0.2	**90.7** ±0.3	**89.7** ±0.0	**88.7** ±0.3	**87.5** ±0.3	**86.3** ±0.1	**91.2** ±0.2

Table 1: Performance (top-5 accuracy $_{\pm\text{std}}$%) on ImageNet100 with 10 incremental steps.

#step	1	2	3	4	5	6	7	8	9	10	Avg
LwF [9]	90.1	77.7	63.9	51.8	43.0	35.5	31.6	28.4	26.4	24.3	42.5
iCaRL [7]	90.0	83.0	77.5	70.5	63.0	57.5	53.5	50.0	48.0	44.0	60.8
EEIL [22]	94.9	**94.9**	84.7	77.8	71.7	66.8	62.5	59.0	55.2	52.3	69.4
BiC [4]	94.1	92.5	89.6	**89.1**	85.7	83.2	80.2	77.5	75.0	73.2	82.9
IL2M [16]	–	–	–	–	–	–	–	–	–	–	78.3
WA [5]	93.9	91.5	89.4	87.7	86.5	85.6	84.5	83.2	82.1	81.1	85.7
EA (Ours)	94.4 ±0.3	92.5 ±0.1	**90.4** ±0.1	89.0 ±0.1	**87.7** ±0.0	**86.8** ±0.1	**85.7** ±0.0	**84.5** ±0.0	**83.4** ±0.0	**82.6** ±0.1	**87.0** ±0.0

Table 2: Performance (top-5 accuracy $_{\pm\text{std}}$%) on ImageNet1000 with 10 incremental steps.

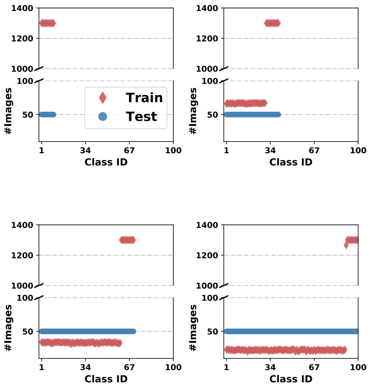


Fig. 3: Frequency distributions in the 1th, 4th, 7th and 10th incremental step (from top to bottom, from left to right) on ImageNet100 with 10 incremental steps in total (10 classes per step).

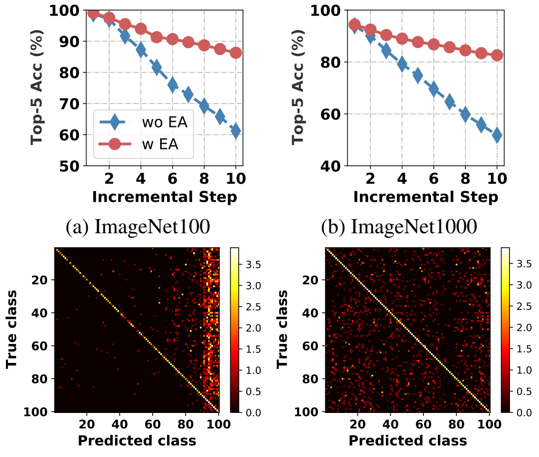
4. EXPERIMENTS

Experiments are conducted on ImageNet ILSVRC 2012 [24]. ImageNet ILSVRC 2012 includes about 1.2 million images for training and 50,000 images for validation in which two settings are provided — ImageNet100 which contains 100 randomly selected classes and ImageNet1000 which consists of the whole 1,000 classes. Our implementation is based on Pytorch [25]. ResNet-18 [26, 27] is employed as backbone.

For fair comparisons, in accordance with the conventional experiment settings proposed in previous work [4, 7, 22, 5], ImageNet100 and ImageNet1000 are split into 10 incremental steps with 10 and 100 classes per step, respectively. In addition, 2,000 and 20,000 images are stored for old classes in the experiments on ImageNet100 and ImageNet1000, respectively. We select rehearsal exemplars randomly. Some frequency distributions are illustrated in Fig. 3, indicating the heavily imbalanced training data. As new classes arrive, the number of samples can be retained for each class decreases gradually. Thus, the class imbalance problem becomes more serious, while the target label distribution remains exact uniform for each class consistently. For each incremental step, the trained model is evaluated on all seen classes and reported on accuracy. After all, the average accuracy (Avg) over all incremental steps except the first step is calculated (as the first step is actually not related to "incremental").

Effect of Energy Alignment. As shown in Fig. 4 (top), EA significantly improves the performance (the gain in terms of top-5 accuracy in the last incremental step is more than 24% on ImageNet100 and 29% on ImageNet1000). We further plot the confusion matrices (after logarithmic transformation) for the models in the last incremental step on ImageNet100. From Fig. 4 (bottom), plain method (without EA) tends to predict objects as new classes, i.e., many samples from old classes (1˜90) are misclassified as new classes (91˜100). With the help of EA, the model treats new classes and old classes fairly. These results intuitively show that EA can effectively alleviate class imbalance in CIL.

Comparisons with State-of-the-Art. Comparison results with the competitive and representative methods on ImageNet100 and Imagenet1000 are listed in Tab. 1 and Tab. 2, respectively, which apparently substantiate that our proposed algorithm achieves better performance when compared to state-of-the-art approaches, thereby demonstrating the effectiveness of EA again.



(a) ImageNet100

(b) ImageNet1000

(c) wo EA

(d) w EA

Fig. 4: Comparisons of learning with or without EA. (top) Top-5 accuracy on ImageNet100 and ImageNet1000. (bottom) Confusion matrix of model trained with or without EA on ImageNet100.

5. CONCLUSION AND FURTHER WORK

In this paper, we propose a straightforward and effective algorithm to deal with the model bias issue in class incremental learning. From the perspective of energy-based models, we systematically analyze the relationship between the free energies of categories and the label distribution. Based on theoretical calculation, we propose the Energy Alignment (EA) approach to adjust output energies of different classes to achieve better overall performance. The comprehensive experiments conducted on class incremental learning benchmarks demonstrate that it outperforms many state-of-the-art methods.

6. REFERENCES

[1] Robert M. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive Sciences*, vol. 3, pp. 128–135, 1999.

[2] Michael McCloskey and Neal J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *Psychology of Learning and Motivation - Advances in Research and Theory*, vol. 24, no. C, pp. 109–165, 1989.

[3] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin, "Learning a unified classifier incrementally via rebalancing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 831–839.

[4] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu, "Large scale incremental learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 374–382.

[5] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia, "Maintaining discrimination and fairness in class incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang, "A tutorial on energy-based learning," *Predicting structured data*, vol. 1, no. 0, 2006.

[7] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[9] Zhizhong Li and Derek Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[10] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 139–154.

[11] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[12] Friedemann Zenke, Ben Poole, and Surya Ganguli, "Continual learning through synaptic intelligence," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3987–3995.

[13] Peng Zhou, Long Mai, Jianming Zhang, Ning Xu, Zuxuan Wu, and Larry S Davis, "M2kd: Multi-model and multi-level knowledge distillation for incremental learning," *arXiv preprint arXiv:1904.01769*, 2019.

[14] Yen-Chang Hsu, Yen-Cheng Liu, and Zsolt Kira, "Re-evaluating continual learning scenarios: A categorization and case for strong baselines," *arXiv preprint arXiv:1810.12488*, 2018.

[15] Gido M. van de Ven and Andreas S. Tolias, "Three scenarios for continual learning," *CoRR*, vol. abs/1904.07734, 2019.

[16] Eden Belouadah and Adrian Popescu, "Il2m: Class incremental learning with dual memory," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[17] Eden Belouadah and Adrian Popescu, "Scail: Classifier weights scaling for class incremental learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1266–1275.

[18] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky, "Your classifier is secretly an energy based model and you should treat it like one," in *International Conference on Learning Representations*, 2020.

[19] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li, "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems*, 2020.

[20] Russel E Caflisch et al., "Monte carlo and quasi-monte carlo methods," *Acta numerica*, vol. 1998, pp. 1–49, 1998.

[21] John Hammersley, *Monte carlo methods*, Springer Science & Business Media, 2013.

[22] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari, "End-to-end incremental learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 233–248.

[23] Jathushan Rajasegaran, Munawar Hayat, Salman H Khan, Fahad Shahbaz Khan, and Ling Shao, "Random path selection for continual learning," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec 2015.

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.