Coarse-To-Fine Incremental Few-Shot Learning - *Appendix*

Xiang Xiang[1], Yuwen Tan[1], Qian Wan[1], Jing Ma[1], Alan L. Yuille[2], and Gregory D. Hager[2]

[1] MoE Key Lab of Image Processing and Intelligent Control,
School of AI and Automation, Huazhong Univ. of Science and Tech., China
[2] Department of Computer Science, Johns Hopkins University, USA
xex@hust.edu.cn

1 Full Proof of Proposition 1

Proposition 1 (Normalizing or freezing weights improves stability; doing both improves the most). *Given $\mathbf{\Theta}_a$, if we only normalize weights of a linear FC classifier, we obtain $\mathbf{\Theta}_b$; if we only freeze them, we obtain $\mathbf{\Theta}_c$; if we do both, we obtain $\mathbf{\Theta}_d$. Then, $\mathcal{D}_d < \mathcal{D}_b < \mathcal{D}_a$ and $\mathcal{D}_d < \mathcal{D}_c < \mathcal{D}_a$.*

Proof. (1) Stability Degree of model $\mathbf{\Theta}_a$.

It is assumed that the training for all sessions will reach the minimum loss. For the training sample m in the 0-th session, the probability that m belongs to superclass is one, i.e., $p^m_{t,c_{super}} = 1$ and $p^m_{t,i} = 0(i \neq c_{super})$. According to $p^m_i = \frac{\exp(o^m_i)}{\sum_{j=1}^{I}\exp(o^m_j)}$, the following conditions are satisfied,

$$\tilde{\mathbf{o}}^{(t)}_{c_{super}} = a(a \in \mathbb{R}), \tilde{\mathbf{o}}^{(t)}_i(i \neq c_{super}) = -\infty. \tag{1}$$

After training of T-th session has reached the minimum loss, $\tilde{\mathbf{o}}^{(T)}_{c_{sub}} = b(b \in \mathbb{R}), \tilde{\mathbf{o}}^{(T)}_i(i \neq c_{sub}) = -\infty$, then,

$$\mathcal{D}_a = \sum_i (\frac{\tilde{\mathbf{o}}^{(T)}_i - \tilde{\mathbf{o}}^{(t)}_i}{\tilde{\mathbf{o}}^{(t)}_i})^2 = \quad (\frac{-\infty - a}{a})^2 + (\frac{b - (-\infty)}{-\infty})^2 = \infty. \tag{2}$$

(2) Stability Degree of model $\mathbf{\Theta}_b$.

Under the same conditions above, the following conditions are satisfied according to $p^m_i = \frac{\exp(\cos\theta^m_i)}{\sum_{j=1}^{I}\exp(\cos\theta^m_j)}$,

$$\tilde{\mathbf{o}}^{(t)}_{c_{super}} = 1, \tilde{\mathbf{o}}^{(t)}_i(i \neq c_{super}) = -1. \tag{3}$$

After training of T-th session has reached minimum loss, $\tilde{\mathbf{o}}^{(T)}_{c_{sub}} = 1, \tilde{\mathbf{o}}^{(T)}_i(i \neq c_{sub}) = -1$, then the following holds:

$$\mathcal{D}_b = \sum_i (\frac{\tilde{\mathbf{o}}^{(T)}_i - \tilde{\mathbf{o}}^{(t)}_i}{\tilde{\mathbf{o}}^{(t)}_i})^2 = \quad (\frac{-1 - 1}{1})^2 + (\frac{1 - (-1)}{-1})^2 = 8. \tag{4}$$

(3) Stability Degree of model $\mathbf{\Theta}_c$.

Compared with $\mathbf{\Theta}_a$, model $\mathbf{\Theta}_c$ freezes weights of neurons corresponding to previously-seen classes. After training of T-th session has reached its minimum loss, $\tilde{\mathbf{o}}_{c_{super}}^{(T)} = a, \tilde{\mathbf{o}}_{c_{sub}}^{(T)} = \infty^+, \tilde{\mathbf{o}}_i^{(T)}(i \neq c_{super} \vee i \neq c_{sub}) = -\infty$, where $\infty^+ > \infty$ in order to offset the influence of $\tilde{\mathbf{o}}_{c_{super}}^{(T)}$, then,

$$\begin{aligned} \mathcal{D}_c = \sum_i (\frac{\tilde{\mathbf{o}}_i^{(T)} - \tilde{\mathbf{o}}_i^{(t)}}{\tilde{\mathbf{o}}_i^{(t)}})^2 = \quad (\frac{\infty^+ - (-\infty)}{-\infty})^2, \\ 9 > \mathcal{D}_c > 4. \end{aligned} \tag{5}$$

(4) Stability Degree of model $\mathbf{\Theta}_d$.

Compared with $\mathbf{\Theta}_b$, model $\mathbf{\Theta}_d$ freezes weights of neurons corresponding to previously-seen classes. After training of T-th session has reached its minimum loss, $\tilde{\mathbf{o}}_{c_{super}}^{(T)} = 1, \tilde{\mathbf{o}}_{c_{sub}}^{(T)} = 1, \tilde{\mathbf{o}}_i^{(T)}(i \neq c_{super} \vee i \neq c_{sub}) = -1$, then,

$$\mathcal{D}_d = \sum_i (\frac{\tilde{\mathbf{o}}_i^{(T)} - \tilde{\mathbf{o}}_i^{(t)}}{\tilde{\mathbf{o}}_i^{(t)}})^2 = \quad (\frac{1 - (-1)}{-1})^2 = 4. \tag{6}$$

Comparing the stability degree of different models, we have $\mathcal{D}_{max} = \mathcal{D}_a$, $\mathcal{D}_{min} = \mathcal{D}_d$ and $\mathbf{\Theta}_d$ is the most stable.

2 Full Analysis of Conjecture 2

Conjecture 2 (Sufficient & necessary condition of no impact of freezing embedding-weights). $p \vee q \Leftrightarrow \neg r$ *where p: classifier-weights are normalized, q: classifier-weights are frozen, r: freezing embedding-weights improves the performance.*

The 'only if' part: $\neg p \wedge \neg q \Rightarrow r$

Analysis. We have 4 propositions that are all true according to our observations:

① $\neg p \wedge \neg q \rightarrow r$
② $p \wedge q \rightarrow \neg r$
③ $p \wedge \neg q \rightarrow \neg r$
④ $\neg p \wedge q \rightarrow \neg r$.

They share a similar composition pattern. We summarize them as Table 1.

P	Q	$P \wedge Q$	R	$P \wedge Q \rightarrow R$
$\neg p$	$\neg q$	$\neg p \wedge \neg q$	r	$\neg p \wedge \neg q \rightarrow r$
p	q	$p \wedge q$	$\neg r$	$p \wedge q \rightarrow \neg r$
p	$\neg q$	$p \wedge \neg q$	$\neg r$	$p \wedge \neg q \rightarrow \neg r$
$\neg p$	q	$\neg p \wedge q$	$\neg r$	$\neg p \wedge q \rightarrow \neg r$

Table 1: Compound propositions.

Let us make p, q, r an realization of general propositions
P: classifier-weights are normalized,

Q: classifier-weights are frozen,
R: freezing embedding-weights improves the performance, respectively. We want to construct a common proposition for all the four cases all to be true. Namely, we need to solve for a comopsitive proposition $\mathcal{C}(P,Q) \to R$ that satisfies the truth table with ①, ②, ③, ④ ordered top-down.

P	Q	$\mathcal{C}(P,Q)$	R	$\mathcal{C}(\mathcal{P},\mathcal{Q}) \to R$
0	0		1	1
1	1		0	1
1	0		0	1
0	1		0	1

Table 2: A truth table that is not completed.

Note that $A \to B$ is 0 *iff* A is 1 and B is 0. Therefore, we want $\mathcal{C}(P,Q)$'s truth value of the $2,3,4$ line never to be 1. Given the value pairs of P and Q, the only way to make that happen is to let $\mathcal{C}(P,Q)$ be $\neg P \wedge \neg Q \to R$, which is a solution that satisfies all four cases, and thus is always true.

P	Q	$\neg P \wedge \neg Q$	R	$\neg P \wedge \neg Q \to R$
0	0	1	1	1
1	1	0	0	1
1	0	0	0	1
0	1	0	0	1

Table 3: The truth table is realized.

Namely, we have $\neg P \wedge \neg Q \Rightarrow R$, which is exactly Conjecture 2, $\neg p \wedge \neg q \Rightarrow r$, with a change of notations.
The 'if' part: $p \vee q \Rightarrow \neg r$.
Analysis. Given propositions ②, ③, ④, we will combine them and derive a logically-equivalent premise.

$$
\begin{aligned}
&(p \wedge q) \vee (p \wedge \neg q) \vee (\neg p \wedge q) \\
\Leftrightarrow\ &(p \vee (p \wedge \neg q) \vee (\neg p \wedge q))) \wedge (q \vee (p \wedge \neg q) \vee (\neg p \wedge q)) \\
\Leftrightarrow\ &\big((p \vee (p \wedge \neg q) \vee \neg p) \wedge (p \vee (p \wedge \neg q) \vee q)\big) \\
\wedge\ &\big((q \vee (p \wedge \neg q) \vee \neg p) \wedge (q \vee (p \wedge \neg q) \vee q)\big) \\
\Leftrightarrow\ &\big((p \vee p \vee \neg p) \vee (p \vee \neg q \vee \neg p) \wedge (p \vee \vee q) \wedge (p \vee \neg q \vee q)\big) \\
\wedge\ &\big((q \vee p \vee \neg p) \wedge (q \vee \neg q \vee \neg p) \wedge (q \vee p \vee q) \wedge (q \vee \neg q \vee q)\big) \\
\Leftrightarrow\ &(1 \wedge 1 \vee (p \vee p \vee q) \wedge 1) \wedge (1 \wedge 1 \wedge (q \vee p \vee q) \wedge 1) \\
\Leftrightarrow\ &(p \vee p \vee q) \wedge (q \vee p \vee q) \\
\Leftrightarrow\ &(p \vee q) \wedge (p \vee q) \Leftrightarrow p \vee q.
\end{aligned}
$$

Similarly, we can derive ② ∨ ③ ∨ ④ as
$(p \wedge q \rightarrow \neg r) \vee (p \wedge \neg q \rightarrow \neg r) \vee (\neg p \wedge q \rightarrow \neg r)$
$\Leftrightarrow (p \wedge q) \vee (p \wedge \neg q) \vee (\neg p \wedge q) \rightarrow \neg r$.

With the premise replaced, we have
② ∨ ③ ∨ ④ $\Leftrightarrow p \vee q \rightarrow \neg r$,

Given ②, ③, ④ are all always true. it holds that $p \vee q \rightarrow \neg r$ is always true, Namely, we have $p \vee q \Rightarrow \neg r$.

3 Insights from Fine-tuning

3.1 Embeddings need to be contrastively learned

As shown in Fig. 1, straightforward training on coarse labels does not help much the subsequent FSL on fine labels (now_acc at $\sim 25\%$), while contrastive learning self-supervised by the fine cues does help (now_acc at $\sim 35\%$). Thus, **coarsely-trained embedding can be generalizable**.

Fig. 2-left shows that freezing embedding-weight outperforms not freezing them. It implies the embedding space without any update is generalizable, and that, *if classifier-weights are not frozen, freezing embedding-weights helps.*

3.2 Freezing weights helps, surely for classifiers

Fig. 2-right implies that, *if classifiers weights are frozen, then freezing embedding-weights does not help.* Comparing left with right of Fig. 2, we find that freezing classifier-weights (right) outperforms not doing so (left), either freezing embedding-weights (circle) or not (triangle).

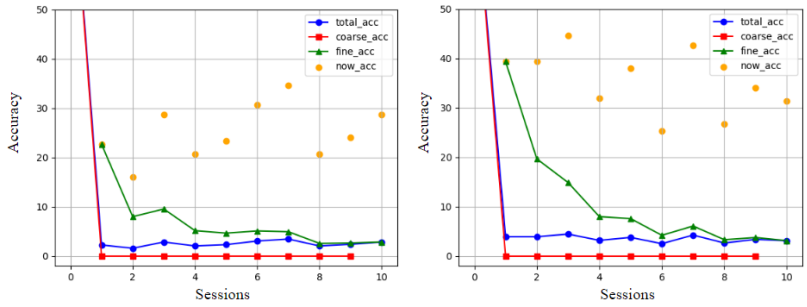


Fig. 1: Ablation study of contrastive learning when fine-tuning ResNet12 w/o IL. Left: w/o; right: w/. (CIFAR-100)

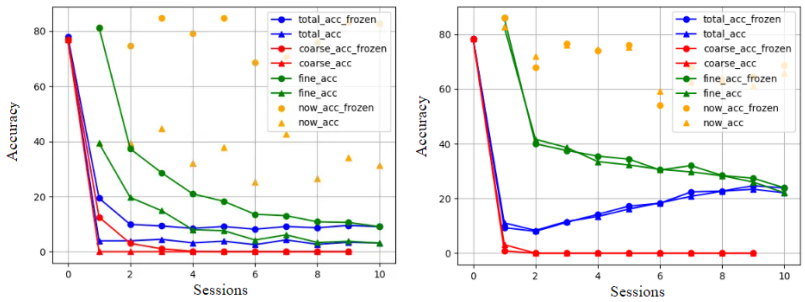


Fig. 2: Ablation study of freezing embedding-weights for fine-tuning a contrastive model. Left: when not freezing classifier-weights. Right: when freezing them.

4 Rethinking C2FSCIL and Knowe with More Results

Given a tree-like product catalog at Amazon.com, there is a class hierarchy per tree, as shown in Fig. 3. Furthermore, Fig. 4 presents the basic idea of Knowe. To learn over time (*i.e.*, sequential learning), it is suggested in [4,5] that neural networks can be limited by catastrophic forgetting (CF) just like Perceptron is unable to solve X-OR. Knowledge forgetting, or called catastrophic forgetting/interference is about a learner's memory (*e.g.*, LSTM) and is a result of the stability–plasticity dilemma regarding how to design a model that is sensitive to, but not radically disrupted by, new input [4,5]. Often, maintaining plasticity results in forgetting old classes while maintaining stability prevents the model from learning new classes, which may be caused by a single set of shared weights. Our setting requires a balance between coarse and fine performance unexplored by existing works, as shown in Fig. 5.

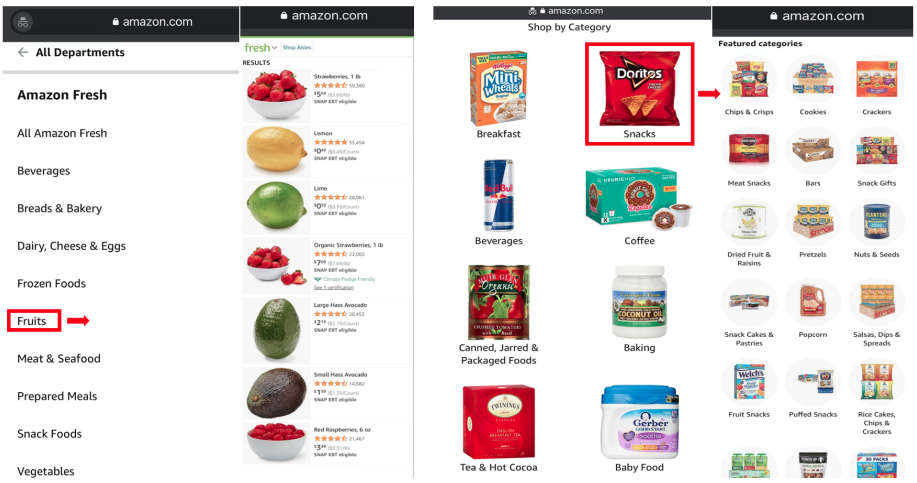


Fig. 3: Two examples of Amazon item catalog. Best seen on computer.

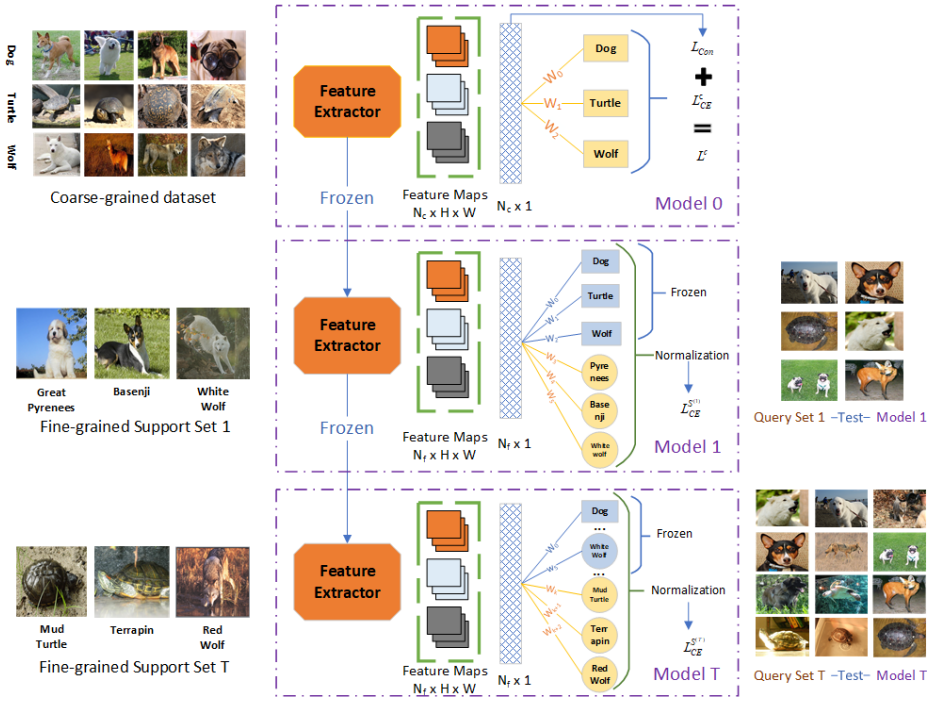


Fig. 4: Basic idea of Knowe. In base session we train $\mathbf{\Theta}$ on $\mathbb{D}$ to get $\mathbf{\Theta}^{(0)}$. Per incremental session, $\mathbf{\Theta}^{(t)}$ is trained on C-way K-shot support set $\mathbb{S}^{(t)}$ based on $\mathbf{\Theta}^{(t-1)}$, $t \geq 1$ and then tested on any class seen in either $\mathbb{D}$ or $\mathbb{S}^{(1)}, ..., \mathbb{S}^{(t)}$.

Method	Class hierarchy	Few-shot Learning	Incremental Learning
LwF [3]			✓
CEC [6]		✓	✓
ANCOR [2]	✓	✓	
IIRC [1]	✓		✓
C2FSCIL (Ours)	✓	✓	✓

Table 4: Comparison of settings with related works.

References

1. Abdelsalam, M., Faramarzi, M., Sodhani, S., Chandar, S.: IIRC: Incremental Implicitly-Refined Classification. In: CVPR (2021)

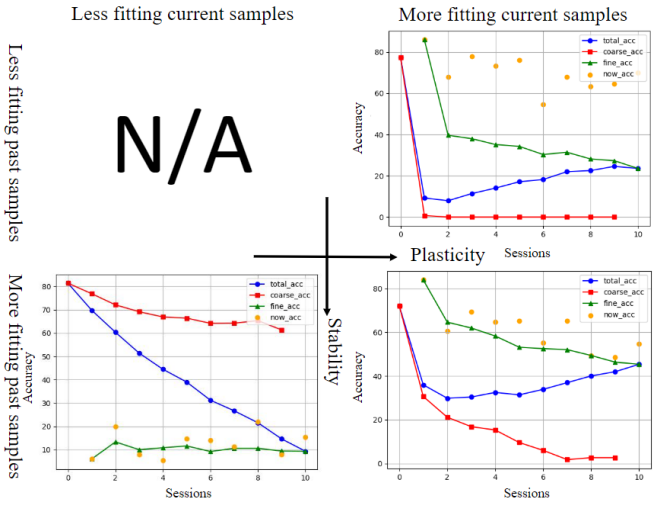
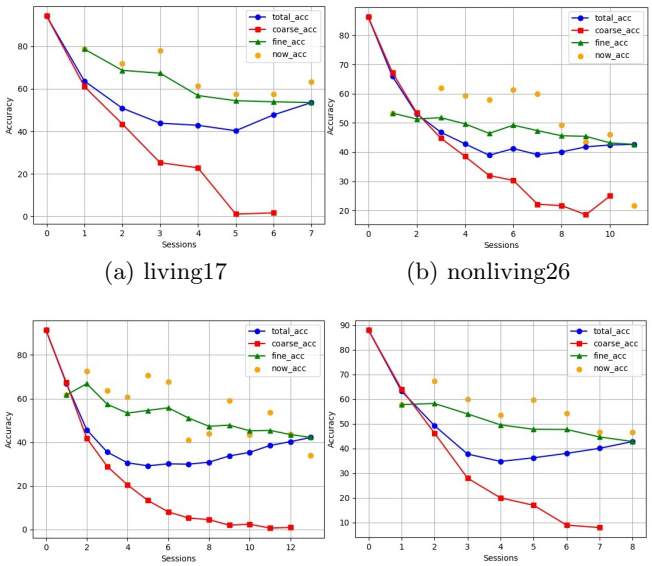


Fig. 5: The stability-plasticity trade-off. Top-right is FT w/o IL; bottom-left represents most IFSL methods; bottom-right is our approach. (CIFAR-100)



(a) living17 (b) nonliving26

(c) entity13 (d) entity30

Fig. 6: Knowe reaches a balance on BREEDS.

2. Bukchin, G., Schwartz, E., Saenko, K., Shahar, O., Feris, R., Giryes, R., Karlinsky, L.: Fine-grained angular contrastive learning with coarse labels. In: CVPR (2021)

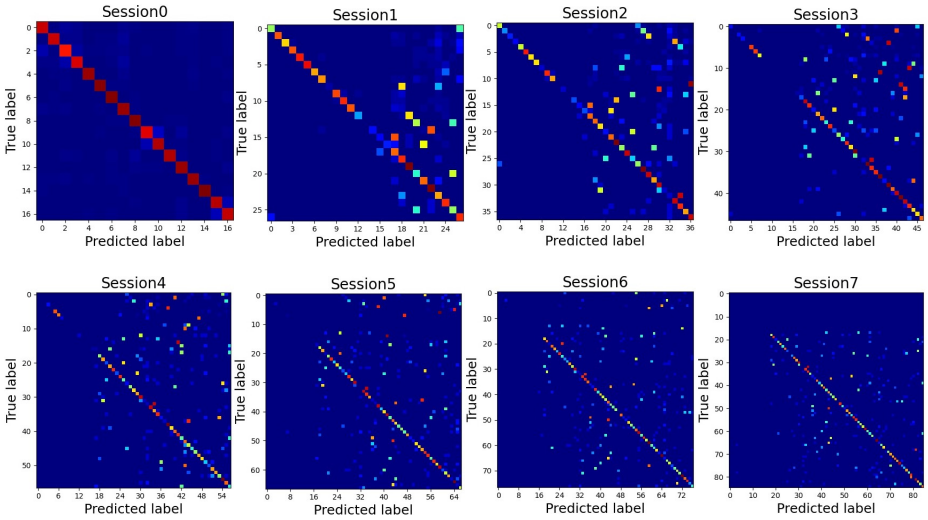


Fig. 7: The visualization of all confusion matrices of Knowe tested on the BREEDS living17 dataset.

3. Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: CVPR (2018)
4. McCloskey, M., Cohen, N.: Catastrophic interference in connectionist networks: the sequential learning problem. The Psychology of Learning and Motivation **24**, 109–164 (1989)
5. M.French, R.: Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences **3** (1999)
6. Zhang, C., Song, N., Lin, G., Zheng, Y., Pan, P., Xu, Y.: Few-shot incremental learning with continually evolved classifiers. In: CVPR (2021)