Quantized Residuals to Continuous Prompts for Few-Shot Class Incremental Learning in Vision-Language Models

Supplementary Material

9. Manifold Curvature Recovery via Residuals

In this section, we provide a simplified theoretical framework showing that the visual-textual residual captures second-order geometric information corresponding to fine-grained attributes that are suppressed by contrastive pretraining.

Setup and Assumptions:

Assumption 1 (Smooth Visual Manifold with Attribute Structure): Let $\mathcal{M} \subset \mathbb{R}^D$ be a smooth manifold representing natural images. Each image $\mathbf{X} \in \mathcal{M}$ can be parameterized by a set of coarse attributes $s \in S$ and fine attributes $a \in A$. We write $\mathbf{X} = \mathbf{X}(s, a)$, where s determines the class and a encodes fine-grained variations within the class.

Assumption 2 (Contrastive Pretraining Suppresses Fine-Grained Gradients): Let δ be represent a small perturbation along fine-grained attribute directions (e.g., slight color variation, minor shape differences). The visual encoder $f_v : \mathcal{M} \to \mathbb{R}^D$ learned via contrastive training satisfies:

$$\|f_v(\mathbf{X} + \delta) - f_v(\mathbf{X})\| \leq \epsilon,$$

where ϵ is small. This occurs because Contrastive learning uses augmentations that introduce fine-grained variations and the loss enforces invariance to these variations, suppressing sensitivity to fine attributes. The network learns to emphasize class-discriminative features while ignoring within-class variations. Moreover, since all the fine-grained information are not suppressed in the feature space, this assumption only holds for those fine-grained details to which the feature is invariant.

Assumption 3 (Text Captures Coarse Semantics Only): The textual embedding $\mathbf{t}_X = f_t(\text{Text for } \mathbf{X})$. primarily encodes coarse semantic attributes s but lacks fine-grained attribute information. Formally, $f_t(\text{Text for } \mathbf{X}(s, a)) \approx f_t(\text{Text for } s) = g(s)$, where $g(s)$ is independent of fine attributes.

Proposition 1: Under Assumptions 1-3, for an image $\mathbf{X} = \mathbf{X}(s, a)$, where a represents fine-grained attributes deviating from the class mean $\bar{a}$ by $\delta = a - \bar{a}$, the visual-textual residual satisfies,

$$\mathbf{r} = f_v(\mathbf{X}) - f_t(\text{Text for } s) \approx +\frac{1}{2} H_{f_v}(\bar{\mathbf{X}})[\delta, \delta] + \mathcal{O}(\|\delta\|^3) + \mathbf{c}_0,$$

where c_0 is the systematic cross-modal misalignment, $\bar{\mathbf{X}}(s, \bar{a})$ is the prototype image with mean attributes, and $H_{f_v}(\bar{\mathbf{X}})$ is the Hessian sensitivity to attribute variations.

Proof:

Using Taylor expansion,

$$\begin{aligned} f_v(\mathbf{X}) &= f_v(\bar{\mathbf{X}} + \delta) \\ &= f_v(\bar{\mathbf{X}}) + \nabla f_v(\bar{\mathbf{X}})\delta + \frac{1}{2} H_{f_v}(\bar{\mathbf{X}})[\delta, \delta] + \mathcal{O}(\|\delta\|^3) \end{aligned} \tag{6}$$

Based on Assumption 2, we have $\mathbb{E}_\delta[f_v(\bar{\mathbf{X}})^T \delta] = 0$. For typical CLIP training with strong augmentations, $\mathcal{O}(\epsilon) << \|\delta\|^3$. This makes sense since if the gradient weren't small, the encoder would be sensitive to fine-grained attribute variations, contradicting the contrastive invariance property. The network achieves invariance by flattening the manifold along these directions and removing first-order sensitivity. Define $\mathbf{r} = f_v(\mathbf{X}) - f_t(\text{Text for } s)$. Therefore,

$$\begin{aligned} \mathbf{r} &= f_v(\mathbf{X}) - f_t(\text{Text for } s) \\ &= \nabla f_v(\bar{\mathbf{X}})^T \delta + \frac{1}{2} H_{f_v}(\bar{\mathbf{X}})[\delta, \delta] + \mathcal{O}(\|\delta\|^3 \\ &+ \underbrace{f_v(\bar{\mathbf{X}}) - f_t(\text{Text for } s)}_{\mathbf{c}_0} \\ &= \frac{1}{2} H_{f_v}(\bar{\mathbf{X}})[\delta, \delta] + \mathcal{O}(\|\delta\|^3) + \mathbf{c}_0(s) \end{aligned}$$

The Hessian term $H_{f_v}(\bar{\mathbf{X}})[\delta, \delta]$ measures the curvature of the visual manifold along fine-grained attribute directions. If H is large, the manifold curves sharply, and fine attributes create significant feature differences, resulting in high discriminability. In addition, the modality gap c_0 provides additional information that can be used to improve the alignment between text and visual domains.

10. Hyper parameter settings

In this section, we provide the details of hyper-parameters in Table 3 that we used for our experiments on each dataset.

11. Proof of Theorem 1

Let Q be any posterior over hypotheses (here hypotheses correspond to continuous weights ϕ). McAllester's PAC-Bayes inequality [18] implies that for any fixed prior P and any posterior Q, with probability at least $1 - \delta$,

$$D_{KL}(\mathbb{E}_Q[\hat{R}_{N_i}] \| \mathbb{E}_Q[R_{N_i}]) \leq \frac{D_{KL}(Q\|P) + \log 2\sqrt{N_i}/\delta}{N_i} \tag{7}$$

Hyper-Parameter	CIFAR100	CUB200	miniImageNet
M	32	32	32
K	32	32	32
DSQ training iters.	15	15	15
No. of prompts	1	1	1
HPE+PC lr	5×10^{-4}	5×10^{-4}	5×10^{-4}
HPE+PC pre-train iters.	50	50	50
Batch size	16	16	16
PC self-attention temp.	2×10^{-3}	1×10^{-3}	2×10^{-3}
Info-NCE Loss Temp.	0.07	0.001	0.07

Table 3. Details of hyper-parameters to training and evaluate the method.

The binary KL-divergence satisfies Pinsker's inequality in the form $|a-b| \leq \sqrt{1/2 D_{KL}(a\|b)}$ for $a, b \in [0,1]$. We get,
$\mathbb{E}_Q[R] - \mathbb{E}_Q[\hat{R}_{N_i}] \leq \sqrt{\frac{D_{KL}(Q\|P)+\log 2\sqrt{N_i}/\delta}{N_i}}$.
Apply factorization to the distribution, and we get,

$$D_{KL}(Q^i\|P) = \sum_{\mathbf{c}\in C} p_C^i(\mathbf{c}) KL(Q_{cont}^i(.|\mathbf{c})\|P_{cont}(.|\mathbf{c})) + D_{KL}(p_C^i\|p_C^0) \tag{8}$$

Now, we can write $D_{KL}(p_C^i\|p_C^0) = \sum_{j=1}^i D_{KL}(p_C^j\|p_C^{j-1}) = \sum_{j=1}^i \Delta_j^{disc}$.
Now, let $K_i^{cont} = D_{KL}(Q_{cont}^i(.|\mathbf{c})\|P_{cont}(.|\mathbf{c}))$, which can be written as $\Delta_i = K_{cont}^i - K_{cont}^{i-1}$. Combining these together, we get

$$\mathbb{E}_{\phi\sim Q^i}[R(\phi)] \leq \mathbb{E}_{\phi\sim Q^i}\hat{R}(\phi) + \sqrt{\frac{K_{cont}^0 + \sum_{j=1}^i \Delta_j + \sum_{j=1}^i \Delta_j^{disc} + \log(\frac{2}{\delta})}{2N_i}} \tag{9}$$

Now, $\Delta_i^{disc} = D_{KL}(p_C^i\|p_C^{i-1})$, which can become infinite if denominator inside log becomes zero. Therefore, we adopt standard smoothing assumption.
Smoothing assumption: Let u be a uniform distribution over $\mathcal{C}$. For small ϵ, we write $\hat{p}^i = (1-\epsilon)p_C^i + \epsilon u$. Now $\hat{p}^i(\mathbf{c})$ is finite for all $\mathbf{c}$.
Lemma: Let $\hat{p}^i(\mathbf{c})$ and $\hat{p}^{i-1}(\mathbf{c})$ are smoothed distribution. Suppose task i introduces at most s_i new codes. Then,

$$D_{KL}(\hat{p}^i(\mathbf{c})\|\hat{p}^{i-1}(\mathbf{c})) \leq \log(1+\frac{s_i}{Z_{i-1}}) + \log(\frac{1}{1-\epsilon}) \tag{10}$$

Proof:
Let S denote the old support with $|S| = Z$, and T be the newly set codes $|T| = s$, and U bet the remaining codewods.

$$D_{KL}(\hat{q}\|\hat{p}) = \sum_{\mathbf{c}\in S} \hat{q}(\mathbf{c}) \log \frac{\hat{q}(\mathbf{c})}{\hat{p}(\mathbf{c})} + \sum_{\mathbf{c}\in T} \hat{q}(\mathbf{c}) \log \frac{\hat{q}(\mathbf{c})}{\hat{p}(\mathbf{c})} + \sum_{\mathbf{c}\in U} \hat{q}(\mathbf{c}) \log \frac{\hat{q}(\mathbf{c})}{\hat{p}(\mathbf{c})} \tag{11}$$

Also, $\hat{p}(\mathbf{c}) \geq (1-\epsilon)p(\mathbf{c})$. For $\mathbf{c} \in T$, $p(\mathbf{c}) = 0$, which gives $\hat{p}(\mathbf{c}) = \epsilon u = \epsilon/W$, where $W = K^M$.
The worst-case $\hat{q}$ (maximizing D_{KL} w.r.t these constraints) will (intuitively) place as much mass as possible on the new code T, and spread it uniformly there, since for fixed total mass on T, making $\hat{q}$ uniform maximizes KL divergence if denominator is equal.

Let $q(c)$ be uniform over $S \cup T$: $q(\mathbf{c}) = 1/(Z+s)$ for $\mathbf{c} \in S \cup T$. Then,

$$\hat{q}(\mathbf{c}) = (1-\epsilon)/(Z+s) + \epsilon/W$$

For $\mathbf{c} \in S$,

$$\hat{p}(\mathbf{c}) = (1-\epsilon)/(Z) + \epsilon/W$$

For $\mathbf{c} \in T$,

$$\hat{p}(\mathbf{c}) = \epsilon/W$$

Substituting all these terms in 11, and considering U is small since original Q assigns zero mass there for worst case bound, followed by using $D_{KL}(Unif(Z+s)\|Unif(Z)) = \log(1+s/Z)$ from [5]. we get the inequality.

12. Empirical Analysis of Theorem 1

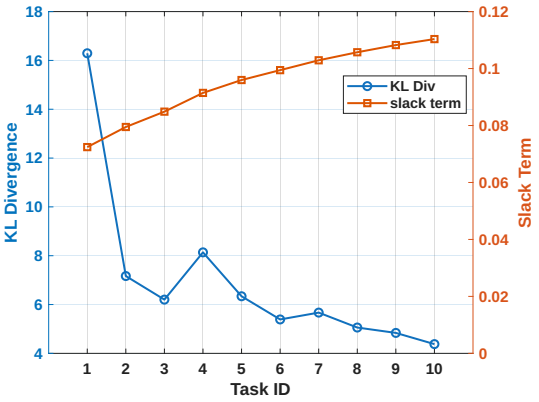

Figure 7. Plot to illustrate the variation of continuous KL divergence and slack term for every iteration

To study the implications of Theorem 1, we compute continuous KL divergence for every session. Since, direct computation is not feasible, we construct a proxy based

on the first and second order statistic of parameter distribution. Assuming that the parameters admit Gaussian distribution with means $\mu 1$ & μ_2 and variances σ_1^2 & σ_2^2, the analytical expression of KL divergence is $D_{KL}(P\|Q) = \log(\frac{\sigma_2}{\sigma_1}) + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} - 0.5$. It can be observed in the Figure 7 that the KL divergence between the parameters of current and preceding session gradually decays, which indicates that the learned parameters of the HPE and PC do not drift significantly, and effectively minimizes catastrophic forgetting. Furthermore, the minimal drift in the parameters results into nearly flattened slack term Δ meaning that the generalization gap does not arbitrarily increase as the session progresses. This observation is direct evidence of stability of the proposed approach in incremental setting.

13. Proof of Theorem 2

Assumption 1: We model the visual-text residual $\mathbf{r}$ for a class y as a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_w)$, where $\boldsymbol{\mu}_y$ is the class mean residual. The quantizer introduces a zero-mean quantization error $\mathbf{q}$ with covariance $\boldsymbol{\Sigma}_q$. For simplicity, The final class prototype $\mathbf{p}_y$ is defined as a linear combination of the base text prototype $\mathbf{t}_y$ and the mean class residual μ_y, given by $\mathbf{p}_y = \mathbf{t}_y + \beta\boldsymbol{\mu_y}$. The classification margin $\gamma_{y,c}$ (for class y vs. c) is based on the difference in linear scores. Finally, we define the discriminative subspace $\mathbf{U}$ as the linear span of all pairwise class mean differences and $\mathbf{P}_U$ as its orthogonal projection.

For a fixed sample (with residual r), pre- and post-quantization margins for pair (y, c) differ because the prototypes change. Without the loss of generality, we will use notation t_y for textual feature x_y^t of class y Using linear scoring:

$$\gamma_{y,c} - \hat{\gamma}_{y,c} = [(\mathbf{t}_y + \mathbf{r})^\top \mathbf{p}_y - (\mathbf{t}_y + \mathbf{r})^\top \mathbf{p}_c] \tag{12}$$
$$- [(\mathbf{t}_y + \mathbf{r})^\top \hat{\mathbf{p}}_y - (\mathbf{t}_y + \mathbf{r})^\top \hat{\mathbf{p}}_c] \tag{13}$$
$$= (\mathbf{t}_y + r)^\top (\mathbf{p}_y - \hat{\mathbf{p}}_y) - (\mathbf{t}_y + \mathbf{r})^\top (\mathbf{p}_c - \hat{\mathbf{p}}_c) \tag{14}$$

But $\mathbf{p}_y - \hat{\mathbf{p}}_y = \beta(\boldsymbol{\mu}_y - \hat{\boldsymbol{\mu}}_y)$ and similarly for class c. So:

$$\gamma_{y,c} - \hat{\gamma}_{y,c} = \beta(\mathbf{t}_y + \mathbf{r})^\top ((\boldsymbol{\mu}_y - \hat{\boldsymbol{\mu}}_y) - (\boldsymbol{\mu}_c - \hat{\boldsymbol{\mu}}_c))$$

Define the per-class quantization error of the mean residual,

$$\mathbf{e}_y := \boldsymbol{\mu}_y - \hat{\boldsymbol{\mu}}_y \tag{15}$$

Then equation 14 becomes,

$$\gamma_{y,c} - \hat{\gamma}_{y,c} = \beta(\mathbf{t}_y + \mathbf{r})^\top (\mathbf{e}_y - \mathbf{e}_c) \tag{16}$$

We will take the expectation over r and the quantization noise. Note that t_y is fixed and $\mathbb{E}[r] = \mu_y$. Take the expectation over the data and quantization randomness,

$$\mathbb{E}[\gamma_{y,c} - \hat{\gamma}_{y,c}] = \beta\mathbb{E}[(\mathbf{t}_y + \mathbf{r})^\top (\mathbf{e}_y - \mathbf{e}_c)]$$

$$\begin{aligned}\mathbb{E}[|\gamma_{y,c} - \hat{\gamma}_{y,c}|] &\le \beta\mathbb{E}[|(\mathbf{t}_y + \mathbf{r})^\top (\mathbf{e}_y - \mathbf{e}_c)|] \\ &\le \beta\mathbb{E}[||\mathbf{t}_y + \mathbf{r}|| \cdot ||\mathbf{e}_y - \mathbf{e}_c||] \text{ (by Cauchy-Schwarz)} \\ &\le \beta\sqrt{\mathbb{E}[||\mathbf{t}_y + \mathbf{r}||^2]}\sqrt{\mathbb{E}[||\mathbf{e}_y - \mathbf{e}_c||^2]}\end{aligned} \tag{17}$$

Moreover, textual and visual features in our setting are defined to be unit norm vectors. Define,

$$B := \sqrt{\max_y \mathbb{E}[||\mathbf{t}_y + \mathbf{r}||^2]} = 1$$

Using linearity and covariance structure,

$$\mathbb{E}[||\mathbf{e}_y - \mathbf{e}_c||^2] = \mathbb{E}[||\mathbf{e}_y||^2] + \mathbb{E}[||\mathbf{e}_c||^2] - 2\mathbb{E}[\mathbf{e}_y^\top \mathbf{e}_c]$$

Assuming that $\mathbf{e}_y$ and $\mathbf{e}_c$ are independent and identically distributed, $\mathbb{E}[\mathbf{e}_y^\top \mathbf{e}_c] = 0$, and $\mathbb{E}[||\mathbf{e}_c||^2]$ and $\mathbb{E}[||\mathbf{e}_c||^2]$ are equal. This gives,

$$\mathbb{E}[||\mathbf{e}_y - \mathbf{e}_c||^2] = 2\mathbb{E}[||\mathbf{e}||^2] \tag{18}$$

Now we want to analyze the error that only affects the classification margin, which lies in the discriminative subspace U. Therefore, the relevant noise is the projection into the discriminative subspace U. Let $\mathbf{P}_U$ project onto U.

$$\mathbb{E}[||\mathbf{e}_y - \mathbf{e}_c||^2] \approx 2\mathbb{E}[||\mathbf{P}_U\mathbf{e}||^2] = 2 \cdot \text{trace}(\mathbf{P}_U\boldsymbol{\Sigma}_q\mathbf{P}_U) \tag{19}$$

Combining equations 17 and 19,

$$\mathbb{E}[|\gamma_{y,c} - \hat{\gamma}_{y,c}|] \le \beta\sqrt{2 \cdot \text{trace}(\mathbf{P}_U\boldsymbol{\Sigma}_q\mathbf{P}_U)}$$

$$\mathbb{E}[|\gamma_{y,c} - \hat{\gamma}_{y,c}|] \le \beta\sqrt{2}\sqrt{\text{trace}(\mathbf{P}_U\boldsymbol{\Sigma}_q\mathbf{P}_U)}$$

Isotropic case:
If $\boldsymbol{\Sigma}_q \le \kappa \cdot \mathbf{I}$ (in the positive semidefinite sense), then,

$$\text{trace}(\mathbf{P}_U\boldsymbol{\Sigma}_q\mathbf{P}_U) \le \text{trace}(\mathbf{P}_U(\kappa \cdot \mathbf{I})\mathbf{P}_U) = \kappa \cdot \text{trace}(\mathbf{P}_U^2) \tag{20}$$
$$= \kappa \cdot \text{trace}(\mathbf{P}_U) = \kappa d, \tag{21}$$

where $d = \dim(U) \le$ Num. of Classes $- 1$.

$$\sqrt{\text{trace}(\mathbf{P}_U\boldsymbol{\Sigma}_q\mathbf{P}_U)} \le \sqrt{\kappa d}$$

This completes the proof of first part.
To guarantee maximum expected margin loss $\le \delta$,

$$\beta\sqrt{2\kappa} \le \delta \implies \kappa \le \frac{\delta^2}{2d\beta^2}. \tag{22}$$

For an isotropic Gaussian source with per-dimension variance σ_r^2 (i.e., $\boldsymbol{\Sigma}_r = \sigma_r^2 I$), the Shannon rate-distortion function is given by,

$$R = \frac{1}{2}\log\left(\frac{\sigma_r^2}{\kappa}\right)$$

No. of Shots	Session-wise Accuracy										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	86.49	84.03	81.57	77.55	76.71	74.27	71.80	70.76	68.92	68.26	68.28
2	86.49	84.63	82.68	79.48	78.50	76.70	75.79	75.81	74.45	74.23	74.51
3	86.62	84.82	83.35	81.25	80.54	78.80	78.26	78.23	77.47	77.48	77.51
4	86.49	84.98	84.14	82.11	82.0	80.35	79.89	79.88	79.18	79.30	79.38
Full-Shot	86.52	85.75	85.54	84.63	84.76	83.89	83.85	83.95	83.62	83.91	84.23

Table 4. Sessionwise accuracy for CUB200 dataset for different number of training samples (shots) using ViT-B/16 backbone.

Substituting κ form equation 22, we get,

$$R_{dim} \geq \frac{1}{2}\log_2\left(\frac{\sigma_r^2}{\frac{\delta^2}{2d\beta^2}}\right)$$

For a D-dimensional vector, total bit rate is,

$$R \geq \frac{D}{2}\log_2\left(\frac{\sigma_r^2}{\frac{\delta^2}{2d\beta^2}}\right) = \frac{D}{2}\log_2\left(\frac{2\beta^2 d\sigma_r^2}{\delta^2}\right)$$

For M subspace with K codes in each, total bit rate is given by $R = M\log_2 K$. This completes the second part of the theorem.

For empirical computation, we omit the factor d because it is a dataset-dependent constant that does not affect how the required bit-rate scales with D and classification margin, and removing it produces a more conservative and easier-to-apply sufficient condition without altering the theorem's practical implications.

14. Additional Experiments

We provide comprehensive analyses for QR-Prompt under different backbone and few-shot settings. We vary the number of training samples (shots) in the incremental sessions from 1 to 4, and full shot samples. Tables 4, 6, and 8 compare the sessionwise accuracies using ViT-B/16 backbone for three datasets. Similarly, Tables 5, 7, and 9 compare the accuracies for ViT-L/14 backbone.

No. of Shots	Session-wise Accuracy										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	87.26	83.55	81.46	77.74	77.76	74.87	73.36	73.04	71.26	70.75	70.90
2	87.26	84.95	84.08	81.04	80.19	77.95	76.96	76.52	75.16	74.99	75.22
3	87.25	85.27	84.72	82.48	82.17	80.00	79.43	79.29	78.29	78.19	78.01
4	87.26	85.30	84.72	82.75	82.94	80.70	80.21	80.23	79.25	79.52	79.49
Full-Shot	87.26	86.61	86.18	85.17	85.50	84.51	84.59	84.54	84.27	84.69	85.10

Table 5. Sessionwise accuracy for CUB200 dataset for different number of training samples (shots) using ViT-L/14 backbone.

Number of Shots	Session-wise Accuracy								
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	98.35	98.01	97.23	96.45	96.22	96.01	95.98	95.85	96.02
2	98.50	98.21	97.31	96.51	96.29	96.12	96.01	95.94	96.10
3	98.52	98.24	97.34	96.58	96.32	96.24	96.12	96.03	96.11
4	98.50	98.26	97.41	96.62	96.37	96.39	96.27	96.20	96.29
Full-Shot	98.52	98.35	97.76	97.76	97.89	97.62	97.38	97.17	97.21

Table 6. Sessionwise accuracy for miniImagenet dataset for different number of training samples (shots) using ViT-B/16 backbone.

Number of Shots	Session-wise Accuracy								
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	98.72	98.63	98.16	98.04	98.05	97.57	97.23	97.20	96.23
2	98.70	98.69	98.23	98.19	98.19	98.01	97.73	97.78	97.67
3	98.72	98.71	98.17	98.08	98.11	97.88	97.86	97.86	97.79
4	98.72	98.71	98.17	98.15	98.15	97.93	97.91	97.91	97.94
Full-Shot	98.71	98.69	98.29	98.27	98.22	98.13	98.03	98.09	98.05

Table 7. Sessionwise accuracy for miniImagenet dataset for different number of training samples (shots) using ViT-L/14 backbone.

Number of Shots	Session-wise Accuracy								
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	82.47	77.47	75.21	71.07	69.04	67.29	65.32	62.59	60.41
2	82.47	79.03	77.86	74.45	73.30	71.71	71.26	69.99	68.43
3	82.47	79.51	78.44	75.33	74.66	73.56	73.52	72.02	69.37
4	82.47	80.05	78.96	75.33	73.85	74.16	73.29	73.33	71.95
Full-Shot	82.47	81.32	81.24	79.71	79.68	79.66	79.67	79.35	78.04

Table 8. Sessionwise accuracy for CIFAR100 dataset for different number of training samples (shots) using ViT-B/16 backbone.

Number of Shots	Session-wise Accuracy								
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	85.23	80.06	77.77	74.35	73.19	71.87	70.80	69.73	67.73
2	85.23	81.15	80.13	77.76	76.86	75.60	75.22	74.35	72.44
3	85.26	82.00	81.06	78.96	78.40	77.96	77.72	76.60	75.43
4	85.26	82.18	81.16	79.43	79.26	78.94	78.70	79.01	77.1
Full-Shot	85.23	84.18	83.89	82.61	82.91	82.75	82.87	82.52	81.74

Table 9. Sessionwise accuracy for CIFAR100 dataset for different number of training samples (shots) using ViT-L/14 backbone.