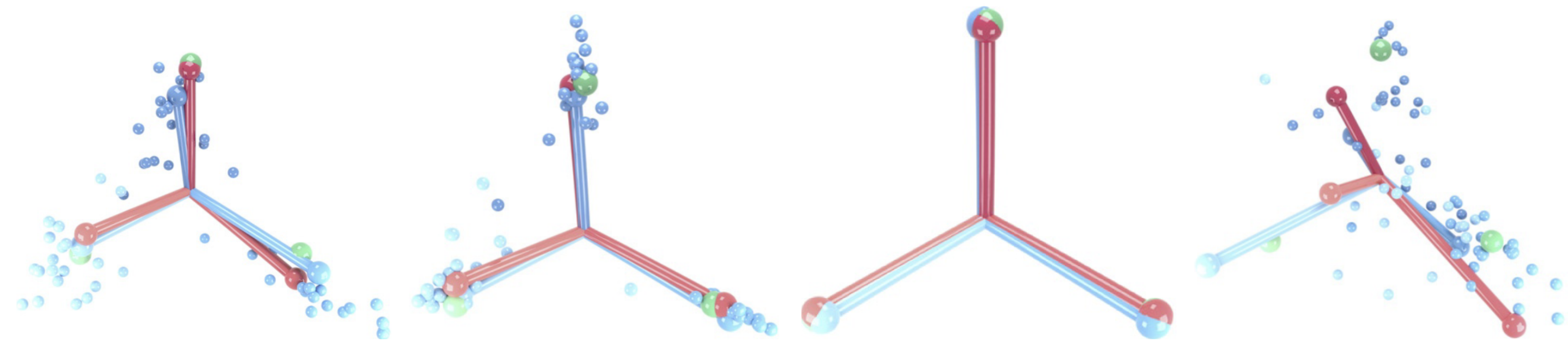
MASIL: Towards Maximum Separable Class Representation for Few Shot Class Incremental Learning

Anant Khandelwal

Abstract

- Learning the maximal separable classifier for FSCIL solving the problem of overfitting and forgetting of seen classes.
- Proposed the idea of concept factorization explaining the collapsed features for base session classes in terms of concept basis and use these to induce classifier simplex for incremental few shot classes.
- Train the classifier jointly on base and novel classes without retaining any base class samples in memory

Neural Collapse



The figure depicts, in three dimensions, Neural Collapse as training proceeds from left to right. Green spheres represent the vertices of the standard simplex, red ball and sticks represent linear classifiers, blue ball and sticks represent class means, and small blue spheres represent last-layer features.

Neural Collapse as defined in (*Papyan et.al*) consists of the following four properties:

- ($\mathcal{NC}1$) **Variability Collapse**: Last layer features of the backbone network for a particular class collapse to within-class mean.
- ($\mathcal{NC}2$) **Convergence**: results in optimal class-means which are equally and maximally pairwise separated forming a simplex Equiangular Tight Frame (ETF).
- ($\mathcal{NC}3$) **Classifier Convergence**: Optimal class means forming ETF are aligned to the corresponding classifier weights uptio rescaling.
- ($\mathcal{NC}4$) **Simplification to nearest class center** When ($\mathcal{NC}1$)-($\mathcal{NC}3$) holds, the model prediction using logits respects nearest class centers.

Concept Factorization

- Class wise feature vectors which are concentrated at higher layers to be recursively broken into multiple concepts moving from highest layer to lower layers tracing back to the input images where it can be explained with regions as concepts, combination of which makes it possible to be able to classify it to particular class.

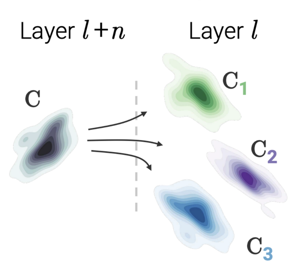


image activations of same class, get progressively merged decompose into multiple concepts

- Activations when collapses to the mean features vector for each class forming the class simplex vector which is composed of concept basis vectors, combining for all classes which gives the overall basis called "*concept bank*".
- We can explain any input in terms on concept bank and computed coefficients using NNLS.

MASIL

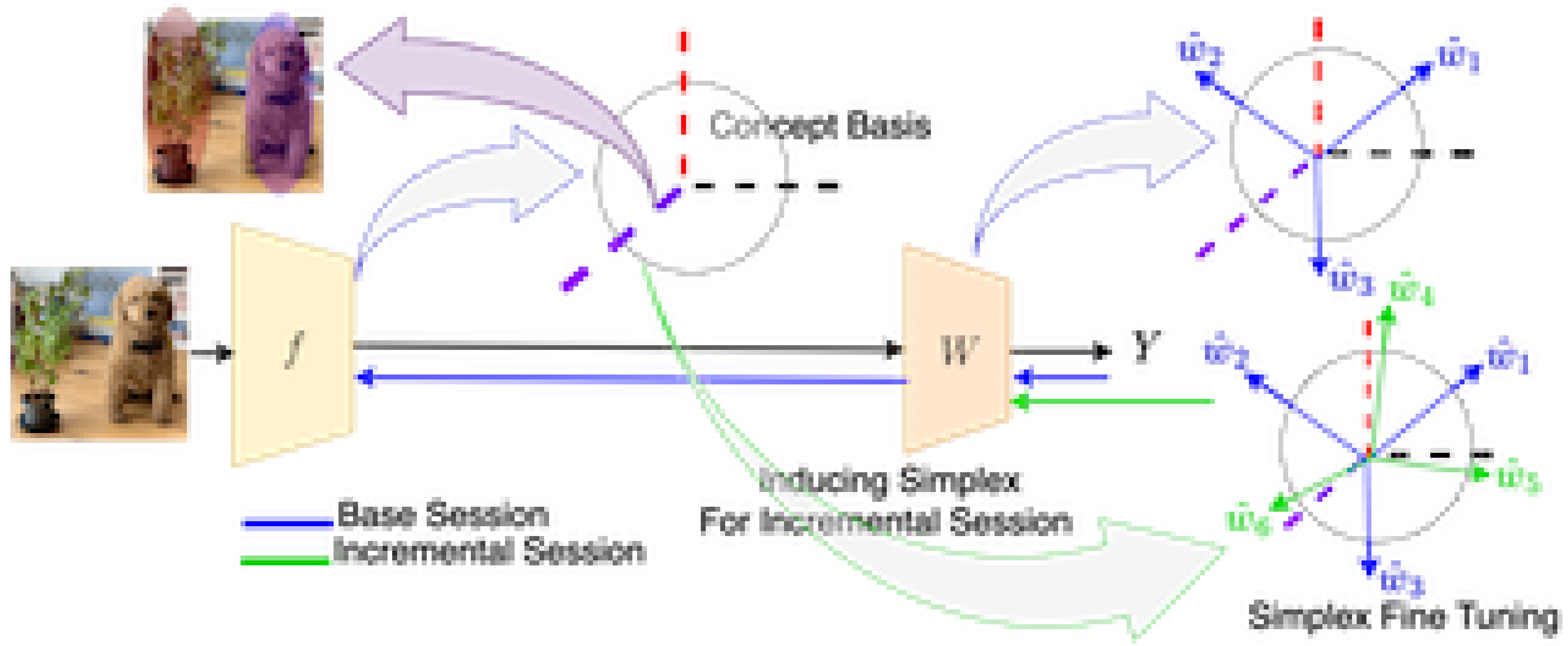


Figure: Architecture of MASIL

Classifier Simplex Representation

Normalized simplex representation on a unit hyper sphere of each class then:

$$\mathbf{w}_{y_i}^T s_{y_i} = 1 \;\forall\; y_i \in \cup_{j=0}^{j=t} \mathcal{C}_j \tag{1}$$

The optimization problem as minimization of features obtained from feature extractor to the vertices of simplex as:

$$\min_{\mathbf{h}_i \in \mathbf{H}^T n} \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{h}_i - s_{y_i}||^2 \tag{2}$$

where s_{y_i} is vertex of simplex and is treated as the class center for class y_i. The best approximation of collapsed feature representation $\mathbf{H} \approx \mathbf{P}\mathbf{Q}^T$. For any input (x_i, y_i) belongs to $\mathcal{D}_j,\ j > 0$

$$\mathbf{h}_i = \mathbf{P}(x_i)\mathbf{Q}^T \tag{3}$$

Feature representation of all instances is denoted as $\mathbf{H}_{y_i} \in \mathbf{H}$, ($\mathcal{NC}1$) implies covariance $\sum_{\mathbf{H}_{y_i}} \rightarrow 0$. i.e. the features collapse to their corresponding class means i.e. $\mathbf{h}^*_{y_i} = \sum_{i=1}^{n_{y_i}} \mathbf{h}_i$, where n_{y_i} is the number of instances for class y_i, this is minimum when $s_{y_i} = \mathbf{h}^*_{y_i}$

$$\hat{\mathbf{w}}_{y_i} = \frac{1}{|\mathcal{D}_j|} \left(\sum_{(x_i, y_i) \in \mathcal{D}_j} \mathbf{P}(x_i) \right) \mathbf{Q}^T \;\forall j > 0 \tag{4}$$

where the coefficients $\mathbf{P}(x_i)$ for each instance of class y_i are calculated using NNLS.

Simplex Finetuning

Due to inherent irreducible error to NNLS, we approach the optimal representation of simplex for few shot class by further fine-tuning the classifier weights.
The mean representation of instance features for each class) in $\mathcal{M}$ given as:

$$\mathcal{M}_{y_i} = \frac{1}{n_{y_i}} \sum_{i=1}^{n_{y_i}} \mathbf{h}_i, \forall \mathcal{M}_{y_i} \in \mathcal{M} \tag{5}$$

where, n_{y_i} is the number of instances of class y_i. The updated loss function during fine tuning stage include the base session classes and few shot class is given as:

$$\min_{\mathbf{w}_{y_i}} \mathcal{L}(\mathbf{w}_{y_i}) = \frac{1}{|\mathcal{D}_j|} \sum_{(x_i, y_i) \in \mathcal{D}_j} ||\mathbf{w}_{y_i}^T \mathbf{h}_i - 1||_F^2 + \frac{1}{|\mathcal{M}|} \sum_{(\mathcal{M}_{y_i}, y_i) \in \mathcal{M}} ||\mathbf{w}_{y_i}^T \mathcal{M}_{y_i} - 1||_F^2 + \alpha ||\mathbf{w}_{y_i} - \hat{\mathbf{w}}_{y_i}||_F^2, \alpha \in [0, 1] \tag{6}$$

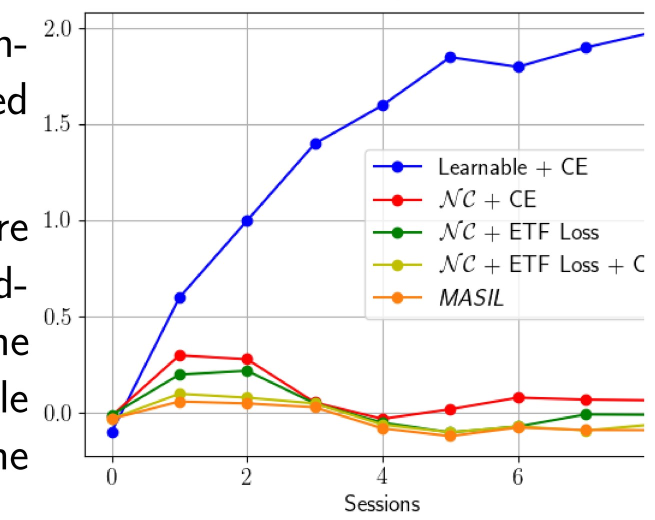
Benchmark Evaluation and Ablation Study

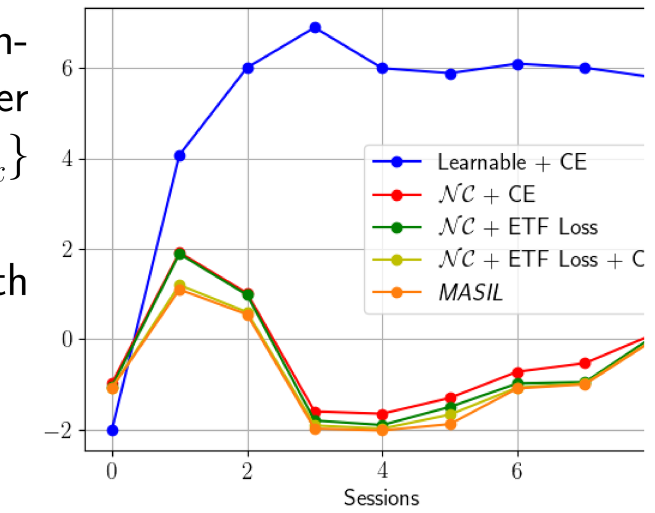
Models	miniImageNet		CIFAR-100		CUB-200	
	Final (↑)	Average (↑)	Final (↑)	Average (↑)	Final (↑)	Average (↑)
Learnable + CE	50.04	61.30	52.13	62.68	50.38	59.58
$\mathcal{NC}$ + CE	56.66	68.23	54.42	64.00	56.83	65.51
$\mathcal{NC}$ + ETF Loss	58.31	67.82	56.11	67.50	59.44	67.28
$\mathcal{NC}$ + ETF Loss + CF	58.72	68.04	56.13	67.51	59.72	67.45
MASIL*(Ours)*	58.86	68.11	57.23	67.84	60.24	67.54

Ablation Studies on three datasets investigating the effects of Simplex based loss, Concept Factorization and Simplex Fine Tuning.

Analysing Classifier Weights

Average cosine similarities Train (Up) and Test (Bottom) on miniImageNet.





- We further analysed the classifier weights alignment with respect to the mean feature (collapsed feature) of each class.
- We used the classifier weights and mean feature from each of the models described in ablation studies to validate the effect of *MASIL* in learning the maximal separable classifier, where the separable property between classes is measured by cosine similarity.
- Specifically, we plotted the average cosine similarities between mean feature and the classifier weights of different classes i.e. $\text{Avg}_{k \neq k'}\{h_k \cdot w'_k\}$ for both train and test datasets.
- Plots confirming the maximum separability with *MASIL*.

Conclusions

- Induce the simplex from concept factorization helps in few shot cases
- Base and novel classes can be learnt together during incremental session to further mitigate forgetting and overfitting
- *MASIL* as an step towards learning the maximum separable classifier in a competitive setting of continual learning i.e. FSCIL