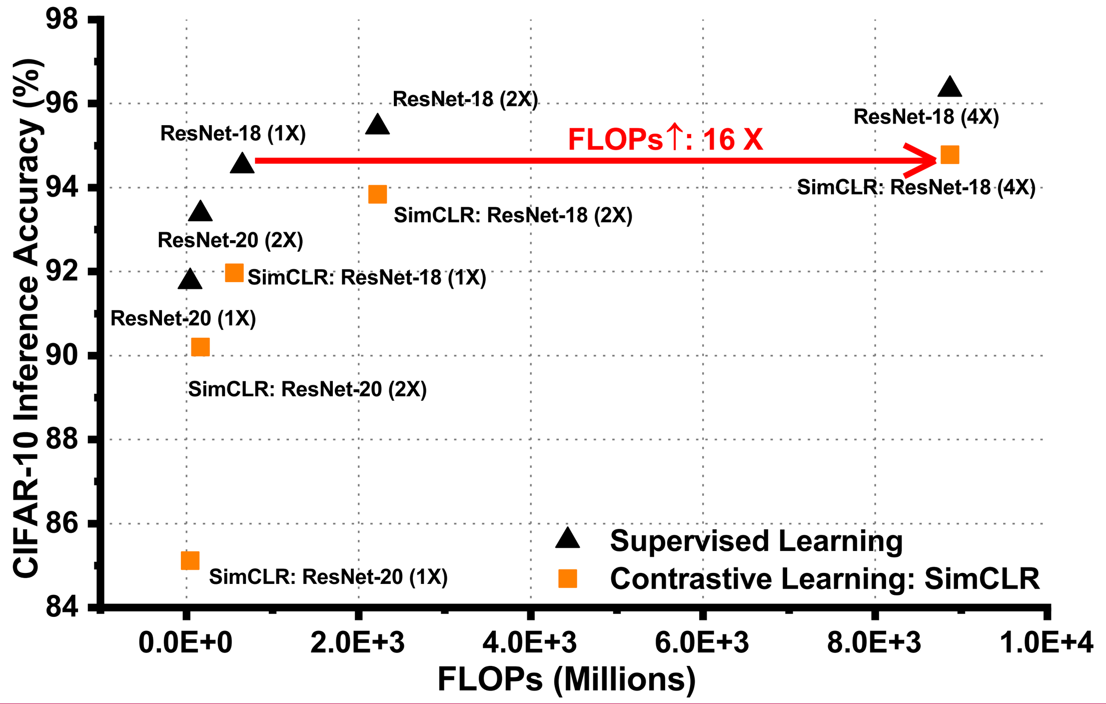
Contrastive Dual Gating: Learning Sparse Features With Contrastive Learning

[1]Jian Meng, [1]Li Yang, [2]Jinwoo Shin, [1]Deliang Fan, [1]Jae-sun Seo

[1]Arizona State University, USA [2]KAIST, South Korea

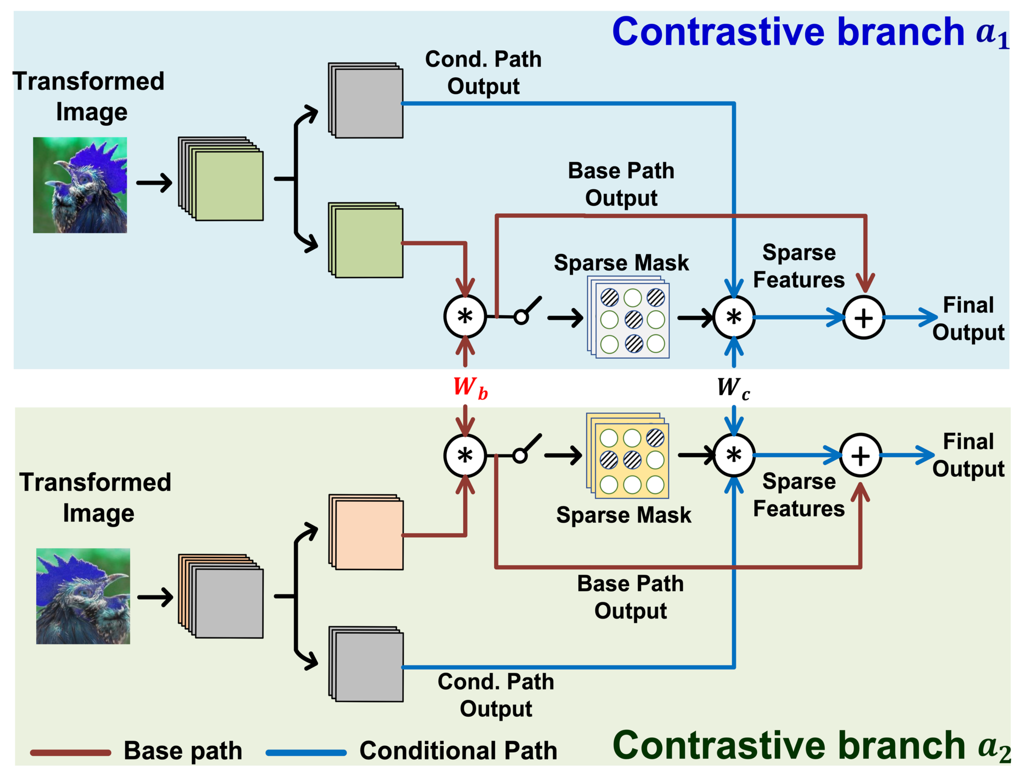
Introduction

- Recent contrastive learning-based self-supervised learning works require wide and deep models (e.g., 4X wider) to achieve comparable performance as supervised training works (1X).
- The extraordinary computation cost necessitates efficient computation reduction techniques for self-supervised learning.



Contrastive Gating

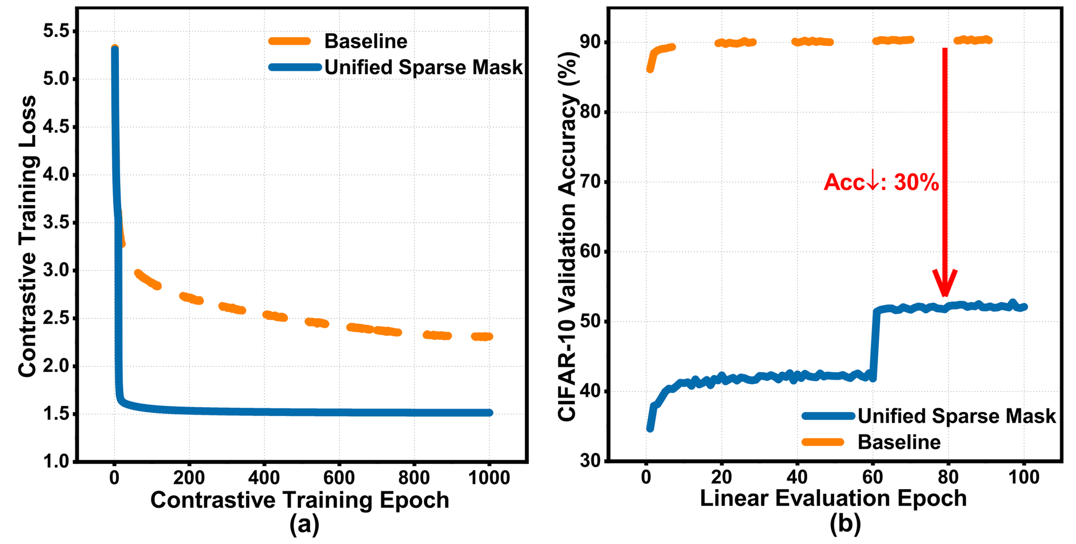
- Learning the sparse features in both contrastive branches during the unsupervised learning process.



- For each layer of the encoder, the input feature maps and weights (X, W) are divided into base $(X_b,\ W_b)$ and conditional $(X_c,\ W_c)$ paths.
- The saliency of the base path output determines the computation skipping decision $M_c \in \{0,1\}$ of the conditional path.

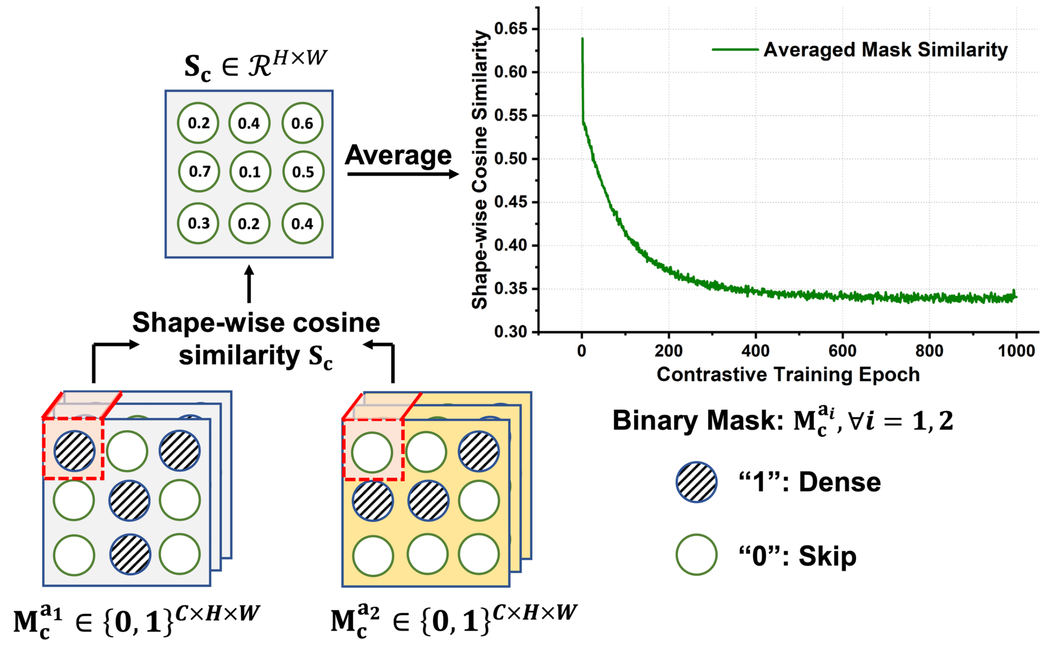
SSL-based Sparse Feature Learning

- **Non-transferability of dynamic sparse feature masks**



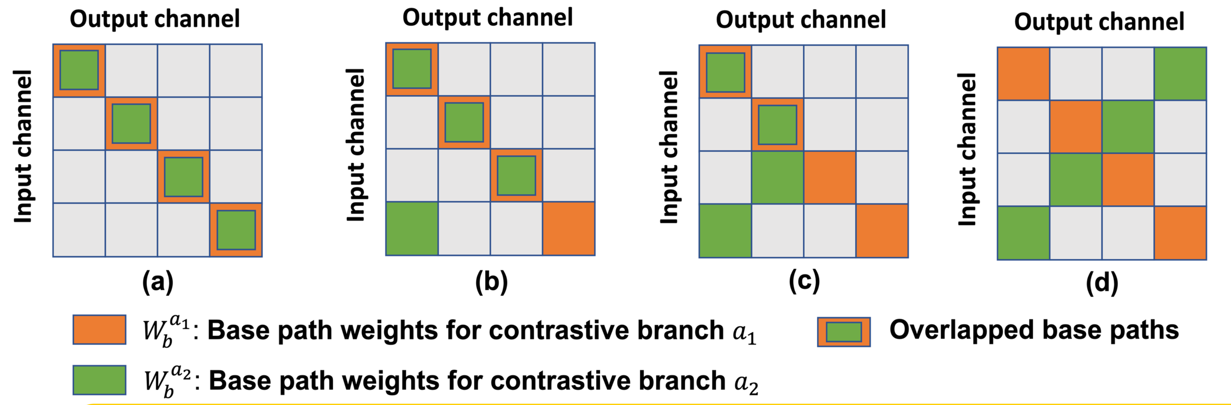
- Broadcasting the identical sparse feature mask M_c to both SimCLR [1] contrastive paths results in: (a) reduced contrastive training loss, and (b) **defective generalizability with unsuccessful supervised linear evaluation.**

- **Disjoint sparse contrastive features**



- **Given the unanimous data transformation and identical base path, contrastive training encourages the encoders to highlight different contrastive features.**

- **Unbiased contrastive grouping**



Overlap	Gating Groups	Cond. Path Sparsity (%)	Inference Acc. (%)
Baseline	-	-	89.16
100%	4	71.88	87.67
75%	4	71.02	87.59
50%	4	70.60	87.12
Disjoint (0%)	4	**72.48**	**88.59**

- **Evenly activating the disjoint channels among the different contrastive paths will enhance the sparse feature learning during contrastive training.**

Contrastive Dual Gating (CDG)

- During the forward pass of the contrastive training, CDG generates pruning masks $M_c^{a_1}$ and $M_c^{a_2}$ for both contrastive branches:

$$M_c^{a_i} = \sigma(\mathbf{normalize}(X_b^{a_i} * W_b^{a_i}) - \tau)$$

  - The contrastive branches are selected along the diagonal and inverse-diagonal of the channel groups. τ learns the gating decision during training.

Structured Contrastive Dual Gating (SCDG)

- During the forward pass of the contrastive training, SCDG generates structured pruning masks $M_c^{a_1}$ and $M_c^{a_2}$ based on the averaged salience:

$$M_{sc}^{a_i} = \sigma(\mathbf{normalize}(\mathbf{AvgPool}(X_b^{a_i} * W_b^{a_i})) - \tau)$$

Experimental Results

- The widely-used mini-neural network-based auxiliary salience predictors (e.g., FBS [2], DGC [3]) **are difficult to train from scratch**, resulting in degraded inference accuracy.

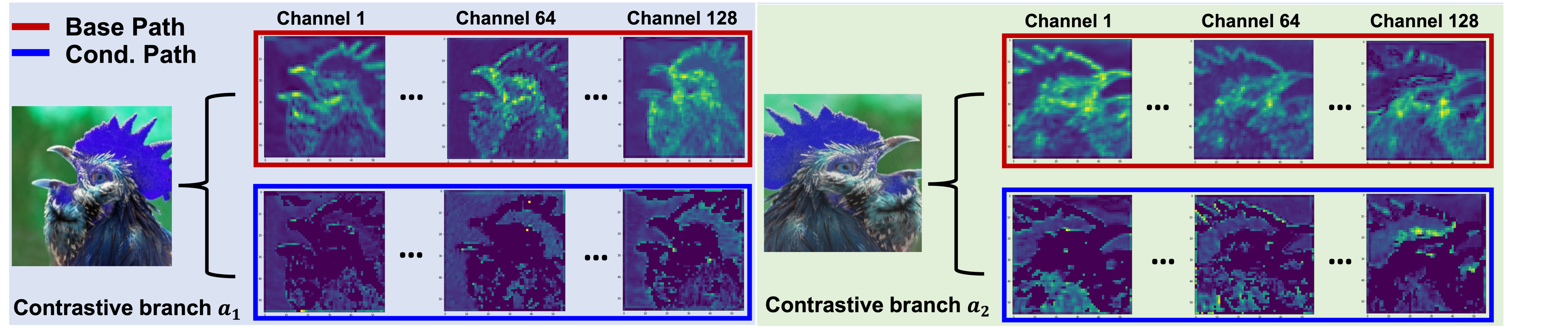
Method	# of Gating Groups	Linear Eval. Inference Accuracy (%)	Fine-tuning Inference Accuracy (%)	FLOPs Reduction
This work (CDG SimCLR)	4	**88.84**	**90.74**	**2.12×**
FBS SimCLR	-	86.91	88.89	2.00×
DGC SimCLR	4	73.10	81.77	2.11×
CGNet SimCLR	4	87.40	89.26	2.09×

- Structured-CDG (SCDG) results with spatial feature group size = 8×1×1

Model	# of Gating Groups	Dataset	Conditional Path Sparsity (%)	Inference Acc. (%)	Top-1 Acc. Drop	FLOPs Reduction	Index Reduction
ResNet-18 (1×)	4	CIFAR-10	71.64	90.37	-0.89	2.16×	8×
		CIFAR-100	66.24	65.94	-1.84	1.98×	8×
		ImageNet-100	45.52	76.63	-2.24	1.53×	8×

  - Similar accuracy as CDG with **8× sparse index reduction.**

- Base paths preserves the details with dense convolution, while the sparse conditional path only keeps the important edges.



(CDG with other datasets and frameworks are presented in main paper.)

[1] T. Chen et al., "A simple framework for contrastive learning of visual representations," ICML, 2020. [2] Z. Su et al., "Dynamic group convolution for accelerating convolutional neural networks," ECCV, 2020. [3] X. Gao et al., "Dynamic Channel Pruning: Feature Boosting and Suppression," ICLR, 2018.