# SCOP: Scientific Control for Reliable Neural Network Pruning (Supplementary Material)

**Yehui Tang**[1,2], **Yunhe Wang**[2], **Yixing Xu**[2], **Dacheng Tao**[3],
**Chunjing Xu**[2], **Chao Xu**[1], **Chang Xu**[3]
[1]Key Lab of Machine Perception (MOE), Dept. of Machine Intelligence, Peking University.
[2]Noah's Ark Lab, Huawei Technologies.
[3]School of Computer Science, Faculty of Engineering, University of Sydney.
yhtang@pku.edu.cn, {yunhe.wang, yixing.xu, xuchunjing}@huawei.com,
{dacheng.tao,c.xu}@sydney.edu.au, chaoxu@cis.pku.edu.cn.

## S1 Proof for Lemma 2

**Lemma 2** *Suppose $\tilde{\mathbf{v}}^l$ is the second-order knockoff counterpart of input feature $\mathbf{v}^l$, the corresponding output feature $\tilde{\boldsymbol{v}}^{l+1} = t(\tilde{\boldsymbol{v}}^l) + \tilde{\boldsymbol{b}}^l$ is still the second-order knockoff of $\boldsymbol{v}^{l+1} = t(\boldsymbol{v}^l) + \boldsymbol{b}^l$, where bias $\boldsymbol{b}^l$ and $\tilde{\boldsymbol{b}}^l$ are random variables with zero means, and the covariance matrix of the joint distribution $[\boldsymbol{b}^l, \tilde{\boldsymbol{b}}^l]$ satisfies:*

$$\mathrm{cov}[\boldsymbol{b}^l, \tilde{\boldsymbol{b}}^l] = \begin{bmatrix} \Sigma_{\boldsymbol{b}^l} & \Sigma_{\boldsymbol{b}^l} + W^T \mathrm{diag}\{s^l\} W - diag\{s^{l+1}\} \\ \Sigma_{\boldsymbol{b}^l} + W^T diag\{s^l\} W - \mathrm{diag}\{s^{l+1}\} & \Sigma_{\boldsymbol{b}^l} \end{bmatrix}. \tag{S.1}$$

*where $\boldsymbol{b}^l$, $\tilde{\boldsymbol{b}}^l$ have the same covariance $\Sigma_{\boldsymbol{b}^l}$ and diag$\{s^l\}$, diag$\{s^{l+1}\}$ are diagonal matrices.*

**Proof** *Recall that $\tilde{\boldsymbol{v}}^l$ is the second-order knockoff of $\boldsymbol{v}^l$ meaning the expectation & covariance of $[\boldsymbol{v}^l, \tilde{\boldsymbol{v}}^l]$ and $[\boldsymbol{v}^l, \tilde{\boldsymbol{v}}^l]_{swap(\tilde{S})}$ are the same for arbitrary subset $\tilde{S}$, and $\tilde{\boldsymbol{v}}^l \perp\!\!\!\perp Y | \boldsymbol{v}^l$. The input $\tilde{\boldsymbol{v}}^l$ contains no targets and thus the output $\tilde{\boldsymbol{v}}^{l+1} = t(\tilde{\boldsymbol{v}}^l) + \tilde{\boldsymbol{b}}^l$ also contains no information about labels, which satisfies the independence condition (Eq. (4)). Their means are equal due to batch normalization [3], and hence we focus on checking whether the covariance matrices of $[\boldsymbol{v}^l, \tilde{\boldsymbol{v}}^l]$ and $[\boldsymbol{v}^l, \tilde{\boldsymbol{v}}^l]_{swap(\tilde{S})}$ match in the following. Actually, $[\boldsymbol{v}^l, \tilde{\boldsymbol{v}}^l]$ and $[\boldsymbol{v}^l, \tilde{\boldsymbol{v}}^l]_{swap(\tilde{S})}$ having the same covariance is equivalent to that the covariance of $[\boldsymbol{v}^l, \tilde{\boldsymbol{v}}^l]$ satisfies the following form,* i.e.*,*

$$\mathrm{cov}[\boldsymbol{v}^l, \tilde{\boldsymbol{v}}^l] = \begin{bmatrix} \Sigma_{\boldsymbol{v}^l} & \Sigma_{\boldsymbol{v}^l} - \mathrm{diag}\{s^l\} \\ \Sigma_{\boldsymbol{v}^l} - \mathrm{diag}\{s^l\} & \Sigma_{\boldsymbol{v}^l} \end{bmatrix}, \tag{S.2}$$

*where $\Sigma_{\boldsymbol{v}^l}$ is the covariance of $\boldsymbol{v}^l$ and diag$\{s^l\}$ is any diagonal matrix making the covariance matrix positive semi-definite [1]. As $\tilde{\boldsymbol{v}}^l$ is the second-order knockoff counterpart of $\boldsymbol{v}^l$, Eq.(S.2) satisfies for $[\boldsymbol{v}^l, \tilde{\boldsymbol{v}}^l]$. Via simple calculation, the covariance matrix of output $[\boldsymbol{v}^{l+1}, \tilde{\boldsymbol{v}}^{l+1}]$ is:*

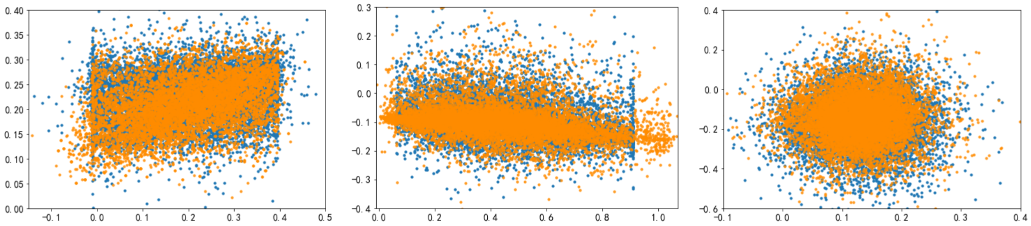
$$\mathrm{cov}[\boldsymbol{v}^{l+1}, \tilde{\boldsymbol{v}}^{l+1}] = \begin{bmatrix} W^T \Sigma_{\boldsymbol{v}^l} W + \Sigma_{\boldsymbol{b}^l} & W^T \Sigma_{\boldsymbol{v}^l} W + \Sigma_{\boldsymbol{b}^l} - diag\{s^{l+1}\} \\ W^T \Sigma_{\boldsymbol{v}^l} W + \Sigma_{\boldsymbol{b}^l} - diag\{s^{l+1}\} & W^T \Sigma_{\boldsymbol{v}^l} W + \Sigma_{\boldsymbol{b}^l} \end{bmatrix}, \tag{S.3}$$

*Denoting $W^T \Sigma_{\boldsymbol{v}^l} W + \Sigma_{\boldsymbol{b}^l}$ as $\Sigma_{\boldsymbol{v}^{l+1}}$, Eq. (S.3) is reformulated as:*

$$\mathrm{cov}[\boldsymbol{v}^{l+1}, \tilde{\boldsymbol{v}}^{l+1}] = \begin{bmatrix} \Sigma_{\boldsymbol{v}^{l+1}} & \Sigma_{\boldsymbol{v}^{l+1}} - diag\{s^{l+1}\} \\ \Sigma_{\boldsymbol{v}^{l+1}} - diag\{s^{l+1}\} & \Sigma_{\boldsymbol{v}^{l+1}} \end{bmatrix}, \tag{S.4}$$

*which satisfies the form of Eq. (S.2), indicating that $[\mathbf{v}^{l+1}, \tilde{\boldsymbol{v}}^{l+1}]$ and $[\mathbf{v}^{l+1}, \tilde{\boldsymbol{v}}^{l+1}]_{swap(\tilde{S})}$ also have the same covariance. Thus, $\tilde{\boldsymbol{v}}^{l+1}$ is still the second-order knockoff of $\mathbf{v}^{l+1}$.*

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.



(a) Input data. (b) Features in shallow layers. (c) Features in deep layers.

Figure S1: Visualization of the distribution of features *w.r.t.* samples on ImageNet.

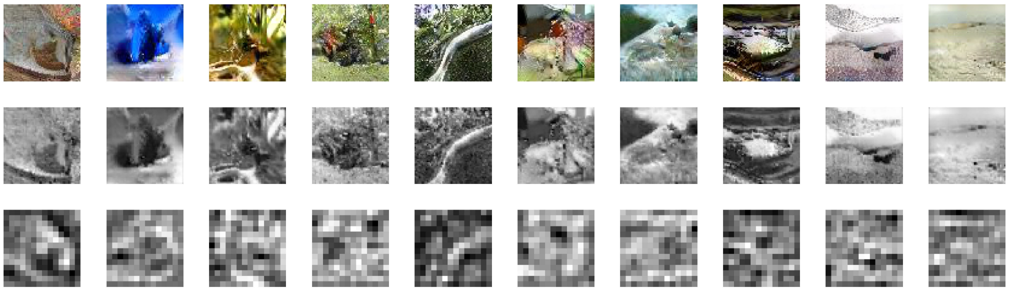


Figure S2: Knockoff data and features on ImageNet. From top to bottom are input images, features in shallow layers and features in the deep layers.

To ensure that biases $[\boldsymbol{b}^l, \tilde{\boldsymbol{b}}^l]$ exist, the covariance matrix in Eq. (S.1) needs to be positive semi-definite. Through standard Schur complement calculation, the semi-definite condition can be derived, *i.e.*, $\text{cov}[\boldsymbol{b}^l, \tilde{\boldsymbol{b}}^l] \succeq 0$ if and only if Eq. (S.5) holds,

$$\begin{aligned} &\text{diag}\{s^{l+1}\} \succeq W^T \text{diag}\{s^l\} W, \\ &2\Sigma_{\boldsymbol{b}^l} \succeq \text{diag}\{s^{l+1}\} - W^T \text{diag}\{s^l\} W. \end{aligned} \tag{S.5}$$

With $\boldsymbol{s}^l$ given by the previous layer, it is easy to satisfy Eq. (S.5) by appointing $\text{diag}\{\boldsymbol{s}^{l+1}\}$ and $\Sigma_{\boldsymbol{b}^l}$. For example, the diagonal elements in $\text{diag}\{s^{l+1}\}$ can be set as the maximum eigenvalue of matrix $W^T \text{diag}\{s^l\} W$ added by a small positive constant $\gamma$ and then $\Sigma_{\boldsymbol{b}^l}$ is set as $(\text{diag}\{s^{l+1}\} - W^T \text{diag}\{s^l\} W)/2 + \gamma I$, where $I$ is the identify matrix. The diagonal matrix $\text{diag}\{s^0\}$ in the input layer is calculated with the real and knockoff data. With the given mean and covariance, biases $[\boldsymbol{b}^l, \tilde{\boldsymbol{b}}^l]$ can be directly generate to modify the outputs of convolutional layers in the procedure of excavating redundant filters.

## S2 Details of generating knockoff data

We train a generative adversarial network [2] to construct knockoff data. The knockoff data are generated by the generator and then sent to the discriminator to verify whether the knockoff condition (Definition 1) holds. The generator and discriminator are optimized alternately and the loss function provided by [4] is adopted. Considering that images are high dimension data, we take multiple pixels as a whole when exchanging elements between real data and knockoff data. To reduce the training cost on ImageNet, knockoff data with size 64 × 64 are generated first and then upsampled to 224 × 224. For the network architectures, multiple convolutional (deconvolutional) layers are stacked to construct the generator and discriminator on CIFAR-10, and ReNet-like architectures are adopted on ImageNet. All the models are optimized with Adam [5] optimizer. On CIFAR-10, the learning rate, batchsize and number of iterations are set to 0.001,128 and 20K, which are set to 0.001, 512 and 30K on ImageNet.

## S3 More Visualization Results

The distribution of features *w.r.t.* samples are shown in Figure S1, and 10K samples are sampled from ImagNet dataset. The blue points denote the concatenation of real features and knockoff features, *i.e.*, $[\mathcal{A}^l, \tilde{\mathcal{A}}^l]$. The orange points are the features after swapping, *i.e.*, $[\mathcal{A}^l, \tilde{\mathcal{A}}^l]_{swap(\tilde{S})}$, which are obtained by swapping half of the elements in $\mathcal{A}^l$ and $\tilde{\mathcal{A}}^l$. Both the two features are mapped to the 2D space for intuitive visualization and each point denotes a sample. It shows that swapping elements between $\mathcal{A}^l$ and $\tilde{\mathcal{A}}^l$ does not have much affect of the distribution, which is accordant with the exchangeability property (Eq. (3)) of knockoffs. More knockoff data and features are shown in Figure S2, which intuitively show that no targets are included in the knockoff data and features.

## References

[1] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[4] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks. In *International Conference on Learning Representations*, 2019.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.