1. Supplementary Material

In this appendix, we study in more detail the relationship between the choice for $\mathbf{N}$, $\mathbf{M}$, and $\mathbf{S}$ parameters of Section 3.2 that control the mask generation process and the resulting model size, mask correlation and models' diversity.

1.1. Expected models size

We first provide a formal derivation for the expected size of a generated model given the $\mathbf{N}, \mathbf{M}, \mathbf{S}$ parameters. Let us consider $\mathbf{N}$ vectors of size $\mathbf{M}\times\mathbf{S}$ filled with zeros and then in each of those vectors we randomly set $\mathbf{M}$ of these elements to be 1. Let ζ_{ij} be a random variable that represents whether or not j-th position in i-th vector is equal to one.

$$\mathbb{P}(\zeta_{ij} = 1) = \frac{\mathbf{M}}{\mathbf{M}\times\mathbf{S}} = \frac{1}{\mathbf{S}}\,. \tag{1}$$

This arises from the fact that there are $\binom{\mathbf{M}\times\mathbf{S}}{\mathbf{M}}$ ways to chose $\mathbf{M}$ positions from $\mathbf{M}\times\mathbf{S}$ places and only $\binom{\mathbf{M}\times\mathbf{S}-1}{\mathbf{M}-1}$ such configurations where j-th position is fixed to be one, therefore

$$\begin{aligned} \mathbb{P}(\zeta_{ij} = 1) &= \binom{\mathbf{M}\times\mathbf{S}-1}{\mathbf{M}-1}\cdot\binom{\mathbf{M}\times\mathbf{S}}{\mathbf{M}}^{-1} \\ &= \frac{\mathbf{M}}{\mathbf{M}\times\mathbf{S}} = \frac{1}{\mathbf{S}} \end{aligned} \tag{2}$$

Let $\epsilon_j = [\sum_{i=1}^{\mathbf{N}} \zeta_{ij} > 0]$ be the random variable representing whether at least one 1 appears in the j-th position among all generated vectors. Given that $\mathbb{P}(\zeta_{ij} = 0) = 1 - \frac{1}{\mathbf{S}}$ and $\mathbb{P}(\overline{X}) = 1 - \mathbb{P}(X)$, then probability of ϵ_j can be written as

$$\mathbb{P}(\epsilon_j = 1) = 1 - (1 - 1/\mathbf{S})^{\mathbf{N}}\,. \tag{3}$$

To compute the expected model size, we compute the expectation of ϵ_j sum, since effectively it represents expected number of features that one will acquire after a trimming procedure:

$$\begin{aligned} Size(\mathbf{N},\mathbf{M},\mathbf{S}) &= \mathbb{E}\sum_{i=1}^{\mathbf{M}\times\mathbf{S}} \epsilon_j = \sum_{i=1}^{\mathbf{M}\times\mathbf{S}} \mathbb{E}\epsilon_j \\ &= \mathbf{M}\times\mathbf{S}\left[1 - (1 - 1/\mathbf{S})^{\mathbf{N}}\right] \end{aligned} \tag{4}$$

1.2. Average IoU

We now justify our $\frac{1}{2\mathbf{S}-1}$ approximation for the average mask IoU. Let us consider two vectors produced by our mask generation algorithm. As above, let ζ_{1j} and ζ_{2j} be random variables that represent the value at the j-th position in the first and second masks, respectively. Starting from the standard definition of the IoU, we estimate the average intersection I of two such masks as the number of common ones

$$\begin{aligned} \mathbb{E}I &= \mathbb{E}\sum_{j=1}^{\mathbf{M}\times\mathbf{S}} \zeta_{1j}\cdot\zeta_{2j} = \sum_{j=1}^{\mathbf{M}\times\mathbf{S}} \mathbb{E}\zeta_{1j}\cdot\mathbb{E}\zeta_{2j} \\ &= \mathbf{M}\times\mathbf{S}\cdot\frac{1}{\mathbf{S}^2} = \frac{\mathbf{M}}{\mathbf{S}}\,. \end{aligned} \tag{5}$$

Given two masks with $\mathbf{M}$ ones each and intersection I, their IoU is $\frac{I}{2\mathbf{M}-I}$. Therefore, a simple approximation for expected IoU is

$$\begin{aligned} \mathbb{E}\left[\text{IoU}\right] &= \mathbb{E}\left[\frac{I}{2\mathbf{M}-I}\right] \approx \frac{\mathbb{E}I}{2\mathbf{M}-\mathbb{E}I} \\ &= \frac{\mathbf{M}/\mathbf{S}}{2\mathbf{M}-\mathbf{M}/\mathbf{S}} = \frac{1}{2\mathbf{S}-1} \end{aligned} \tag{6}$$

In Fig. 1, we plot this value as a function of S along with empirical values and the agreement is excellent.

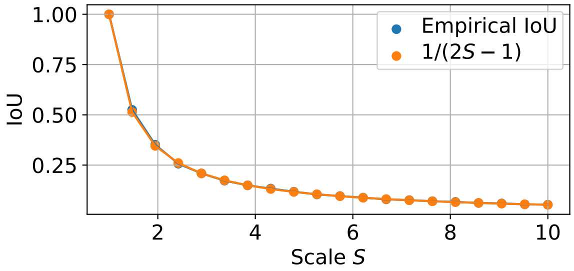

Figure 1. **Empirical and Analytical IoU**. The plot represents how close are real IoU of generated masks and analytical approximation for it. In wide range of $\mathbf{S}$ values we demonstrate a close match of considered quantities.

1.3. Diversity analysis

It is a known that less correlated ensembles of models deliver better performance, produce more accurate predictions [3, 5, 2], and demonstrate lower calibration error [4]. Hence, a strength of Masksembles is that it provides the ability to control how correlated models within Masksembles are by adjusting the $\mathbf{N}$, $\mathbf{M}$, and $\mathbf{S}$ parameters.

We now perform a diversity analysis of Masksembles models using the metric of [1]. It involves comparing two different models trained on the same data in terms of how different their predictions are. Measuring fraction of the test data points on which predictions of models disagree, the *diversity*, and normalizing it by the models error rate, we write

$$\text{Diversity} = \frac{\text{fraction of disagreed labels}}{1.0 - \text{accuracy}} \tag{7}$$

Plotting this diversity against accuracy allows us to look at our models from a bias-variance trade-off perspective.

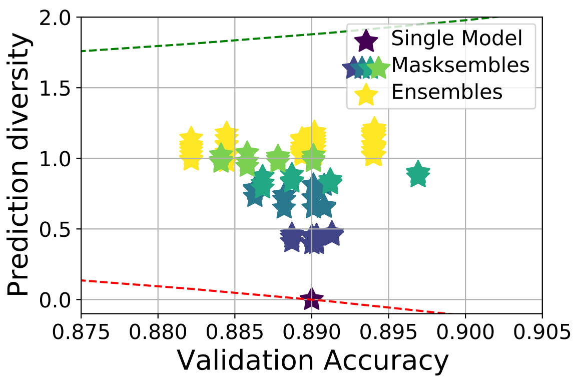

Figure 2. **Diversity vs Accuracy trade-off for Cifar10**. Every point is associated with a pair of compared models. Larger diversity values correspond to less correlated models. **Green** and **Red** dashed lines represent the worst and the best theoretically possible diversity for a fixed accuracy.

Since we want our models to be less correlated—larger *diversity*—and at the same time to be accurate, the upper-right corner of Fig. 2 is where the best models should be.

For this experiments we used trained single model; several Masksembles models with **N** $= 4$, fixed **M** to have the same capacity as the single one and varying **S** in $[2, 3, 4, 5]$; ensembles model with 4 members. Fig. 2 depicts the results on CIFAR10 dataset. Ensembles have the largest diversity but Masksembles gives us the ability to achieve very similar results by controlling its parameters. The single model has 0 diversity by the definition.

1.4. Calibration Plots

Since *Expected Calibration Error* (ECE) provides only limited and aggregated information about model's calibration therefore in this section we present full calibration plots for experiments performed on CIFAR and ImageNet datasets in experiments sections.

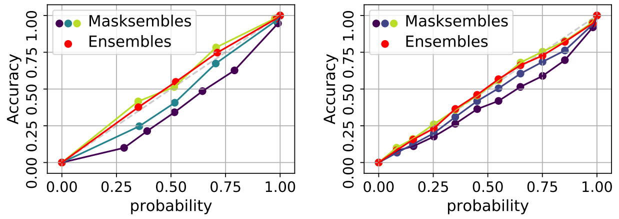

Figure 3. **Calibration plots.** Calibration results for **CIFAR10** (left) and **ImageNet** (right) test sets. Perfect calibration corresponds to $y = x$ curve. Masksembles exhibit a more ensembles-like behavior for lower values of masks overlap.

For both datasets, the calibration plots support our claim that lower model correlation yields better calibration. Furthermore, reducing masks overlap for Masksembles enables us to match Ensembles behavior very closely.

References

[1] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. arXiv preprint arXiv:1912.02757, 2019.

[2] L. K. Hansen and P. Salamon. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(10):993–1001, 1990.

[3] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in neural information processing systems, pages 6402–6413, 2017.

[4] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In Advances in Neural Information Processing Systems, pages 13991–14002, 2019.

[5] Michael P Perrone and Leon N Cooper. When networks disagree: Ensemble methods for hybrid neural networks. Technical report, BROWN UNIV PROVIDENCE RI INST FOR BRAIN AND NEURAL SYSTEMS, 1992.