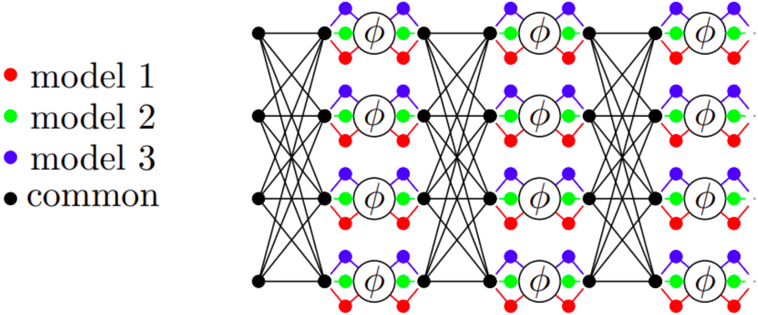
Embedded Ensembles: Infinite Width Limit and Operating Regimes

Шокоров Вячеслав Александрович

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Москва,
2022 г.

Embedded Ensembles scheme



All models in the BatchEnsemble have common fully-connected (or convolutional) weights (colored black), and a small number of pre- and post- activation modulations (colored red, green or blue) that differ for each model. For each model, only the respective family of modulations (i.e., red, green or blue) is active on a forward pass.

Standard fully-connected layer:

$$z_j^l = \frac{1}{\sqrt{N_{l-1}}} \sum_{i=1}^{N_{l-1}} x_i^{l-1} W_{ij}^l + b_j^l, \qquad x_j^l = \phi(z_j^l)$$

BatchEnsemble:

$$\begin{cases} z_{\alpha j}^l = \frac{1}{\sqrt{N_{l-1}}} \sum_{i=1}^{N_{l-1}} x_{\alpha i}^{l-1} W_{ij}^l + b_j^l, \\ x_{\alpha j}^l = u_{\alpha j}^l \ \phi(v_{\alpha j}^l z_{\alpha j}^l), \end{cases}$$

where $\mathrm{u}_\alpha = \{u_{\alpha j}^l, v_{\alpha j}^l\}$ - modulating weighs.
$\mathrm{w} = \{W_{ij}^l, b_j^l\}$ - shared weights.

Embedded Ensembles: MC dropout ensembles

BatchEnsemble:

$$\begin{cases} z^l_{\alpha j} = \frac{1}{\sqrt{N_{l-1}}} \sum_{i=1}^{N_{l-1}} x^{l-1}_{\alpha i} W^l_{ij} + b^l_j, \\ x^l_{\alpha j} = u^l_{\alpha j}\ \phi(v^l_{\alpha j} z^l_{\alpha j}), \end{cases}$$

where $\mathrm{u}_\alpha = \{u^l_{\alpha j}, v^l_{\alpha j}\}$ - modulating weighs.
$\mathrm{w} = \{W^l_{ij}, b^l_j\}$ - shared weights.
MC dropout ensemble - modification of BatchEnsemble:

$$\begin{cases} z^l_{\alpha j} = \frac{1}{\sqrt{N_{l-1}}} \sum_{i=1}^{N_{l-1}} x^{l-1}_{\alpha i} W^l_{ij} + b^l_j, \\ x^l_{\alpha j} = u^l_{\alpha j}\ \phi(z^l_{\alpha j}), \end{cases}$$

where $\mathrm{u}_{\alpha j}$ have a Bernoulli 0-1 distribution.
$\mathrm{w} = \{W^l_{ij}, b^l_j\}$ - shared weights.

Training of Embedded Ensemble

$$\Delta \mathrm{u}_\alpha = -\mu \frac{\partial L_\alpha(\mathrm{w}, \mathrm{u}_\alpha)}{\partial \mathrm{u}_\alpha}$$

$$\Delta \mathrm{w} = -\mu \mathrm{w} \frac{\gamma(M)}{M} \sum_{\alpha=1}^{M} \frac{\partial L_\alpha(\mathrm{w}, \mathrm{u}_\alpha)}{\partial \mathrm{w}}$$

μ_w, μ_u - different learning rates. Authors propose to control accumulation of gradients by means of scaling factor $\gamma(M)$. In the sequel we will argue that the natural choice for scaling factor is either $\gamma(M) = 1$ or $\gamma(M) = M$ depending on whether the ensemble is in independent or collective regime.

Theoretical Analysis

Independence of two different models can be achieved if:

- independent initialization
- have dynamic independence

Dynamic independence:

$$\Delta f_\alpha(\mathrm{x}) = \sum_{\beta=1}^{M} \frac{\partial f_\alpha(\mathrm{x})}{\partial \mathrm{w}} \Delta_\beta \mathrm{w} \propto -\sum_{\beta=1}^{M} \frac{\partial f_\alpha(\mathrm{x})}{\partial \mathrm{w}} \frac{\partial \mathcal{L}_\beta}{\partial \mathrm{w}}$$

So, dynamic independence $\Leftrightarrow \frac{\partial f_\alpha(\mathrm{x})}{\partial \mathrm{w}} \frac{\partial \mathcal{L}_\beta}{\partial \mathrm{w}} = 0$

Neural Tangent Kernel (NTK) for single network

$$\frac{df(\mathrm{x})}{dt} = -\frac{1}{B}\sum_{b=1}^{B}\Theta(\mathrm{x}, \mathrm{x}_b, \mathrm{w})\frac{\partial L(f(\mathrm{x}_b), y_b)}{\partial f(\mathrm{x}_b)},$$

$$\Theta(\mathrm{x}, \mathrm{x}', \mathrm{w}) = \frac{\partial f(\mathrm{w}, \mathrm{x})}{\partial \mathrm{w}}\frac{\partial f(\mathrm{w}, \mathrm{x}')}{\partial \mathrm{w}}.$$

Here $\Theta(\mathrm{x}, \mathrm{x}', \mathrm{w})$ is the Neural Tangent Kernel of network $f(\mathrm{x}, \mathrm{w})$.
If $N \to \infty$, the following three statements are hold:

(A1) The network output $f(\mathrm{x})$ is a draw from a Gaussian Process (GP) at initialization.

(A2) The NTK $\Theta(\mathrm{x}, \mathrm{x}')$ converges to a non-random deterministic value at initialization.

(A3) The NTK $\Theta(\mathrm{x}, \mathrm{x}')$ stays constant during training.

[Jacot et al., 2018]

Neural Tangent Kernel (NTK) for Embedded Ensemble

$$\frac{df_\alpha(\mathrm{x})}{dt} = -\frac{1}{B}\sum_{b,\beta}\Theta_{\alpha\beta}(\mathrm{x},\mathrm{x}_b)\frac{\partial L(f_\beta(\mathrm{x}_b), y_b)}{\partial f_\beta(\mathrm{x}_b)},$$

$$\Theta_{\alpha\beta}(\mathrm{x},\mathrm{x}') = \frac{\gamma(M)}{M}\Theta^{\mathrm{com}}_{\alpha\beta}(\mathrm{x},\mathrm{x}') + \delta_\alpha\Theta^{\mathrm{ind}}_\alpha(\mathrm{x},\mathrm{x}')$$

$$\Theta^{\mathrm{com}}_{\alpha\beta}(\mathrm{x},\mathrm{x}') = \frac{\partial f(\mathrm{w},\mathrm{u}_\alpha,\mathrm{x})}{\partial \mathrm{w}}\frac{\partial f(\mathrm{w},\mathrm{w}_\beta\mathrm{x}')}{\partial \mathrm{w}}.$$

$$\Theta^{\mathrm{ind}}_{\alpha}(\mathrm{x},\mathrm{x}') = \frac{\partial f(\mathrm{w},\mathrm{u}_\alpha,\mathrm{x})}{\partial \mathrm{u}_\alpha}\frac{\partial f(\mathrm{w},\mathrm{w}_\alpha\mathrm{x}')}{\partial \mathrm{u}_\alpha}.$$

Theorem 1 (Outputs f_α).

1. **(Gaussianity)** Consider a BatchEnsemble or MC dropout ensemble at initialization. Then in the sequential infinite-width limit $N_l \to \infty$ the collection of ensemble model outputs $f_\alpha(x)$ converge (in law) to a zero mean Gaussian Process (GP).
2. **(Independence)** If $U_1^L \equiv \mathbb{E}[u_{\alpha j}^L] = 0$, then the GP covariance $\mathbb{E}[f_\alpha(\mathrm{x}) f_\beta(\mathrm{x}')] = 0$ for all x, x' and different ensemble models $\alpha \neq \beta$.
3. **(Breakdown of independence)** Let the activation function $\phi \in \mathcal{S}$ and $U_1^L \equiv \mathbb{E}[u_{\alpha j}^L] \neq 0$. Then $\mathbb{E}[f_\alpha(\mathrm{x}) f_\beta(\mathrm{x}')] > 0$ for all $\alpha \neq \beta$ and all (resp., all linearly independent) pairs of non-zero inputs x, x' for network depths $\mathrm{L} > 1$.

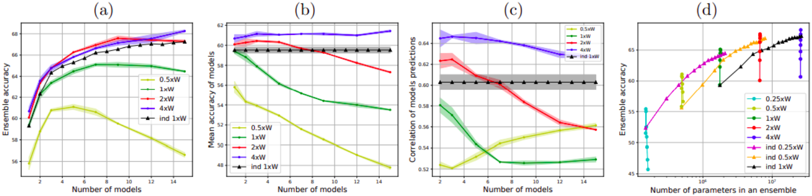
$\mathcal{S}$ is family of non-negative non-decreasing locally Lipschitz functions.

Neural Tangent Kernel (NTK) for Embedded Ensemble

Theorem 2 (NTK $\Theta_{\alpha\beta}$).

1. **(Determinacy)** Consider a BatchEnsemble or MC dropout ensemble at initialization. Then in the sequential limit $N_l \to \infty$ the ensemble NTK $\Theta_{\alpha\beta}(\mathrm{x}, \mathrm{x}')$ converges to a deterministic value.
2. **(Dynamic independence)** If $U_1^L = 0$, then the ensemble NTK $\Theta_{\alpha\beta}(\mathrm{x}, \mathrm{x}') = 0$ for all x, x' and different ensemble models $\alpha \neq \beta$.
3. **(Breakdown of dynamic independence)** If $U_1^L \neq 0$ and the activation function $\phi \in \mathcal{S}$, then $\Theta_{\alpha\beta}(\mathrm{x}, \mathrm{x}') > 0$ for all $\alpha \neq \beta$ and all pairs of inputs x, x' with scalar product $\mathrm{x}^T \mathrm{x}' > 0$.

Expirements



Performance of BatchEnsembles in independent regime and usual independent ensembles for different model widths and numbers of models. Each curve corresponds to a particular model width. The respective value (0.25x – 4x) is the number of neurons in each layer relative to the baseline network. **(a)**, **(b)**, **(c)**: From left to right: Accuracy of ensemble predictions on test set, mean accuracy of individual EE models on test set, mean accuracy of individual EE models on train set, . **(d)**: Scatter plot comparing various ensembles with respect to accuracy and the absolute number of parameters.