Overcoming Multi-model Forgetting Supplementary Material

A. Proofs

Lemma 1. *Given a dataset $\mathcal{D}$ and two architectures with shared parameters $\boldsymbol{\theta}_s$ and private parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, and provided that $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_s, \mathcal{D}) = p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_s, \mathcal{D})p(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_s, \mathcal{D})$, we have*

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_s \mid \mathcal{D}) \propto \frac{p(\mathcal{D} \mid \boldsymbol{\theta}_2, \boldsymbol{\theta}_s)p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s \mid \mathcal{D})p(\boldsymbol{\theta}_2, \boldsymbol{\theta}_s)}{\int p(\mathcal{D} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_s)p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s)d\boldsymbol{\theta}_1}. \quad (1)$$

Proof. Using Bayes' theorem and ignoring constants, we have

$$\begin{aligned}
p(\boldsymbol{\theta} \mid \mathcal{D}) &= \frac{p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_s, \mathcal{D})}{p(\mathcal{D})} \\
&\propto p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2, \boldsymbol{\theta}_s, \mathcal{D})p(\boldsymbol{\theta}_2, \boldsymbol{\theta}_s, \mathcal{D}) \\
&= p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_s, \mathcal{D})p(\mathcal{D} \mid \boldsymbol{\theta}_2, \boldsymbol{\theta}_s)p(\boldsymbol{\theta}_2, \boldsymbol{\theta}_s) \\
&\propto \frac{p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s, \mathcal{D})p(\mathcal{D} \mid \boldsymbol{\theta}_2, \boldsymbol{\theta}_s)p(\boldsymbol{\theta}_2, \boldsymbol{\theta}_s)}{p(\mathcal{D}, \boldsymbol{\theta}_s)} \\
&\propto \frac{p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s, \mathcal{D})p(\mathcal{D} \mid \boldsymbol{\theta}_2, \boldsymbol{\theta}_s)p(\boldsymbol{\theta}_2, \boldsymbol{\theta}_s)}{\int p(\mathcal{D} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_s)p(\boldsymbol{\theta}_s, \boldsymbol{\theta}_1)d\boldsymbol{\theta}_1} \\
&\propto \frac{p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s \mid \mathcal{D})p(\mathcal{D} \mid \boldsymbol{\theta}_2, \boldsymbol{\theta}_s)p(\boldsymbol{\theta}_2, \boldsymbol{\theta}_s)}{\int p(\mathcal{D} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_s)p(\boldsymbol{\theta}_s, \boldsymbol{\theta}_1)d\boldsymbol{\theta}_1},
\end{aligned}$$

where we used the conditional independence assumption $p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2, \boldsymbol{\theta}_s, \mathcal{D}) = p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_s, \mathcal{D})$ in the third line. □

We now derive a closed-form expression for the denominator of equation (1).

Lemma 2. *Suppose we have the maximum likelihood estimate $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s)$ for the first model, write $\mathrm{Card}(\boldsymbol{\theta}_1) + \mathrm{Card}(\boldsymbol{\theta}_s) = p_1 + p_s = p$, and let the negative Hessian $\boldsymbol{H}_p(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s)$ of the log posterior probability distribution $\log p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s \mid \mathcal{D})$ evaluated at $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s)$ be partitioned into four blocks corresponding to $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s)$ as*

$$\boldsymbol{H}_p(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s) = \left[\begin{array}{c|c} \boldsymbol{H}_{11} & \boldsymbol{H}_{1s} \\ \hline \boldsymbol{H}_{s1} & \boldsymbol{H}_{ss} \end{array}\right].$$

If the parameters of each model follow Normal distributions, i.e., $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s) \sim \mathcal{N}_p(\mathbf{0}, \sigma^2 \boldsymbol{I}_p)$, with $\boldsymbol{I}_p$ the p-dimensional identity matrix, then the denominator of equation (1), $A = \int p(\mathcal{D} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_s)p(\boldsymbol{\theta}_s, \boldsymbol{\theta}_1)d\boldsymbol{\theta}_1$ can be written as

$$A = \exp\{l_p(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s) - \frac{1}{2}\boldsymbol{v}^\top \boldsymbol{\Omega} \boldsymbol{v}\} \times (2\pi)^{p_1/2} |\det(\boldsymbol{H}_{11}^{-1})|^{1/2}, \quad (2)$$

where $\boldsymbol{v} = \boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}}_s$ and $\boldsymbol{\Omega} = \boldsymbol{H}_{ss} - \boldsymbol{H}_{1s}^\top \boldsymbol{H}_{11}^{-1} \boldsymbol{H}_{1s}$.

Proof. We have

$$\begin{aligned}
p(\mathcal{D} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_s)p(\boldsymbol{\theta}_s, \boldsymbol{\theta}_1) &\propto e^{l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s) - (\boldsymbol{\theta}_1, \boldsymbol{\theta}_s)^T(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s)/2\sigma^2} \\
&\propto e^{l_p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s)},
\end{aligned}$$

where $l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s) = \log p(\mathcal{D} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_s)$, and $l_p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s) = l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s) - (\boldsymbol{\theta}_1, \boldsymbol{\theta}_s)^T(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s)/2\sigma^2$.

Let $\boldsymbol{H}_p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s) = \boldsymbol{H}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s) + \sigma^{-2}\boldsymbol{I}_p$ be the negative Hessian of $l_p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s)$, with $\boldsymbol{I}_p$ the p-dimensional identity matrix and $\boldsymbol{H}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s)$ the negative Hessian of $l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s)$.

Using the second-order Taylor expansion of $l_p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s)$ around its maximum likelihood estimate $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s)$, we have

$$l_p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s) = l_p(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s) - \frac{1}{2}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_s')^T \boldsymbol{H}_p(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s)(\boldsymbol{\theta}_1', \boldsymbol{\theta}_s'); \quad (3)$$

where $(\boldsymbol{\theta}_1', \boldsymbol{\theta}_s') = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_s) - (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s)$. The first derivative is zero since it is evaluated at the maximum likelihood estimate. We now partition our negative Hessian matrix as

$$\boldsymbol{H}_p(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s) = \left[\begin{array}{c|c} \boldsymbol{H}_{11} & \boldsymbol{H}_{1s} \\ \hline \boldsymbol{H}_{s1} & \boldsymbol{H}_{ss} \end{array}\right],$$

which gives

$$\begin{aligned}
\boldsymbol{B} =& [(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s) - (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s)]^T \boldsymbol{H}_p(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s)[(\boldsymbol{\theta}_1, \boldsymbol{\theta}_s) - (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_s)] \\
=& (\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1)^T \boldsymbol{H}_{11}(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1) + (\boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}}_s)^T \boldsymbol{H}_{ss}(\boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}}_s) \\
&+ (\boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}}_s)^T \boldsymbol{H}_{s1}(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1) \\
&+ (\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1)^T \boldsymbol{H}_{1s}(\boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}}_s) \\
=& (\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1)^T \boldsymbol{H}_{11}(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1) + (\boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}}_s)^T \boldsymbol{H}_{ss}(\boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}}_s) \\
&+ (\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1)^T (\boldsymbol{H}_{1s} + \boldsymbol{H}_{s1}^T)(\boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}}_s).
\end{aligned}$$

Let us define $\boldsymbol{u} = \boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1$, $\boldsymbol{v} = \boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}}_s$ and $\boldsymbol{w} =$

$\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{1s}\boldsymbol{v}$. We then have,

$$\begin{aligned}
\boldsymbol{C} =&(\boldsymbol{u}+\boldsymbol{w})^T\boldsymbol{H}_{11}(\boldsymbol{u}+\boldsymbol{w})\\
=&\boldsymbol{u}^T\boldsymbol{H}_{11}\boldsymbol{u}+\boldsymbol{u}^T\boldsymbol{H}_{11}\boldsymbol{w}+\boldsymbol{w}^T\boldsymbol{H}_{11}\boldsymbol{w}+\boldsymbol{w}^T\boldsymbol{H}_{11}\boldsymbol{u}\\
=&(\boldsymbol{\theta}_1-\hat{\boldsymbol{\theta}}_1)^T\boldsymbol{H}_{11}(\boldsymbol{\theta}_1-\hat{\boldsymbol{\theta}}_1)\\
&+(\boldsymbol{\theta}_1-\hat{\boldsymbol{\theta}}_1)^T\boldsymbol{H}_{11}\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{1s}(\boldsymbol{\theta}_s-\hat{\boldsymbol{\theta}}_s)\\
&+\boldsymbol{v}^T\boldsymbol{H}_{1s}^T\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{11}\boldsymbol{H}_{1s}\boldsymbol{v}\\
&+\boldsymbol{v}^T\boldsymbol{H}_{1s}^T\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{11}(\boldsymbol{\theta}_1-\hat{\boldsymbol{\theta}}_1)\\
=&\boldsymbol{B}-\boldsymbol{v}^T\boldsymbol{H}_{ss}\boldsymbol{v}+\boldsymbol{v}^T\boldsymbol{H}_{1s}^T H_{11}^{-1}\boldsymbol{H}_{1s}\boldsymbol{v}\\
=&\boldsymbol{B}-\boldsymbol{v}^T(\boldsymbol{H}_{ss}-\boldsymbol{H}_{1s}^T\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{1s})v\\
=&\boldsymbol{B}-\boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v},
\end{aligned}$$

with $\boldsymbol{\Omega}=\boldsymbol{H}_{ss}-\boldsymbol{H}_{1s}^T\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{1s}$.

Thus

$$\boldsymbol{B}=(\boldsymbol{u}+\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{1s}v)^T\boldsymbol{H}_{11}(\boldsymbol{u}+\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{1s}\boldsymbol{v})+\boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v}. \tag{4}$$

Given equation (4), we are now able to prove Lemma 2, as integral

$$\begin{aligned}
D =& \int e^{l_p(\boldsymbol{\theta}_1,\boldsymbol{\theta}_s)}d\boldsymbol{\theta}_1=\int e^{l_p(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_s)-\frac{1}{2}\boldsymbol{B}}d\boldsymbol{\theta}_1\\
=&\int e^{l_p(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_s)}e^{-\frac{1}{2}\boldsymbol{B}}d\boldsymbol{\theta}_1=e^{l_p(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_s)}\int e^{-\frac{1}{2}\boldsymbol{B}}d\boldsymbol{\theta}_1\\
=&\int e^{-\frac{1}{2}((\boldsymbol{u}+\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{1s}\boldsymbol{v})^T\boldsymbol{H}_{11}(\boldsymbol{u}+\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{1s}\boldsymbol{v})+\boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v})}d\boldsymbol{\theta}_1\\
&\times e^{l_p(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_s)}\\
=&\int e^{-\frac{1}{2}((\boldsymbol{u}+\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{1s}\boldsymbol{v})^T\boldsymbol{H}_{11}(\boldsymbol{u}+\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{1s}\boldsymbol{v}))}e^{-\frac{1}{2}\boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v}}d\boldsymbol{\theta}_1\\
&\times e^{l_p(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_s)}\\
=&e^{l_p(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_s)-\frac{1}{2}\boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v}}\int e^{-\frac{1}{2}(\boldsymbol{\theta}_1-\boldsymbol{z})^T\boldsymbol{H}_{11}(\boldsymbol{\theta}_1-\boldsymbol{z})}d\boldsymbol{\theta}_1\\
=&e^{l_p(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_s)-\frac{1}{2}\boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v}}(2\pi)^{\frac{p_1}{2}}|\det(\boldsymbol{H}_{11}^{-1})|^{\frac{1}{2}}\\
&\times(2\pi)^{-\frac{p_1}{2}}|\det(\boldsymbol{H}_{11}^{-1})|^{-\frac{1}{2}}\\
&\times\int e^{-\frac{1}{2}(\boldsymbol{\theta}_1-\boldsymbol{z})^T\boldsymbol{H}_{11}(\boldsymbol{\theta}_1-\boldsymbol{z})}d\boldsymbol{\theta}_1\\
=&e^{l_p(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_s)-\frac{1}{2}\boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v}}(2\pi)^{\frac{p_1}{2}}|\det(\boldsymbol{H}_{11}^{-1})|^{\frac{1}{2}},
\end{aligned}$$

where we re-arranged the terms so that the integral is over a normal distribution with mean $\boldsymbol{z}=\hat{\boldsymbol{\theta}}_1-\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{1s}(\boldsymbol{\theta}_s-\hat{\boldsymbol{\theta}}_s)$ and covariance matrix $\boldsymbol{H}_{11}^{-1}$, which can be computed in closed form. □

From Lemma 1 and Lemma 2, we can obtain equation 3 by replacing the denominator with the closed form above and taking the log on both size of equation (1). This yields

$$\begin{aligned}
\log p(\boldsymbol{\theta}|\mathcal{D}) \propto& \log p(\mathcal{D}\mid\boldsymbol{\theta}_2,\boldsymbol{\theta}_s)+\log p(\boldsymbol{\theta}_1,\boldsymbol{\theta}_s)\\
&+\log p(\boldsymbol{\theta}_2,\boldsymbol{\theta}_s)\\
&-\log\{\int p(\mathcal{D}\mid\boldsymbol{\theta}_1,\boldsymbol{\theta}_s)p(\boldsymbol{\theta}_1,\boldsymbol{\theta}_s)d\boldsymbol{\theta}_1\}\\
=&\log p(\mathcal{D}\mid\boldsymbol{\theta}_2,\boldsymbol{\theta}_s)+\log p(\boldsymbol{\theta}_1,\boldsymbol{\theta}_s)\\
&+\log p(\boldsymbol{\theta}_2,\boldsymbol{\theta}_s)-l_p(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_s)+\frac{1}{2}\boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v}\\
&+\log\{(2\pi)^{\frac{p_1}{2}}|\det(\boldsymbol{H}_{11}^{-1})|^{\frac{1}{2}}\}\\
\propto&\log p(\mathcal{D}\mid\boldsymbol{\theta}_2,\boldsymbol{\theta}_s)+\log p(\boldsymbol{\theta}_2,\boldsymbol{\theta}_s)\\
&+\log p(\boldsymbol{\theta}_1,\boldsymbol{\theta}_s\mid\mathcal{D})+\frac{1}{2}\boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v}\ .
\end{aligned}$$

B. Plots for CNN Search

In our CNN search experiment, we search for a“micro” cell as in (Pham et al., 2018). We employ the hyper-parameters available in the released ENAS code. The plots depicting error difference as a function of training epochs as provided in Figure 1 *(a)*, *(b)*and *(c)*. Note that here again the original ENAS is subject to multi-model forgetting, and our WPL helps reducing it. In Figure 1 *(d)*, we show the mean reward as training progresses. While the shape of the reward curve is different from the RNN case, because of a different formulation of the reward function, the general trend is the same; Our approach initially produces lower rewards, but is better at maintaining good models until the end of the search, as indicated by higher rewards in the second half of training.

C. Best architectures found by the search

In Figure 2, we show the best architectures found by our neural architecture search for the RNN and CNN cases.

References

Pham, H., Guan, M. Y., Zoph, B., Le, Q. V., and Dean, J. Efficient Neural Architecture Search via Parameter Sharing. *International Conference on Machine Learning (ICML)*, 2018.

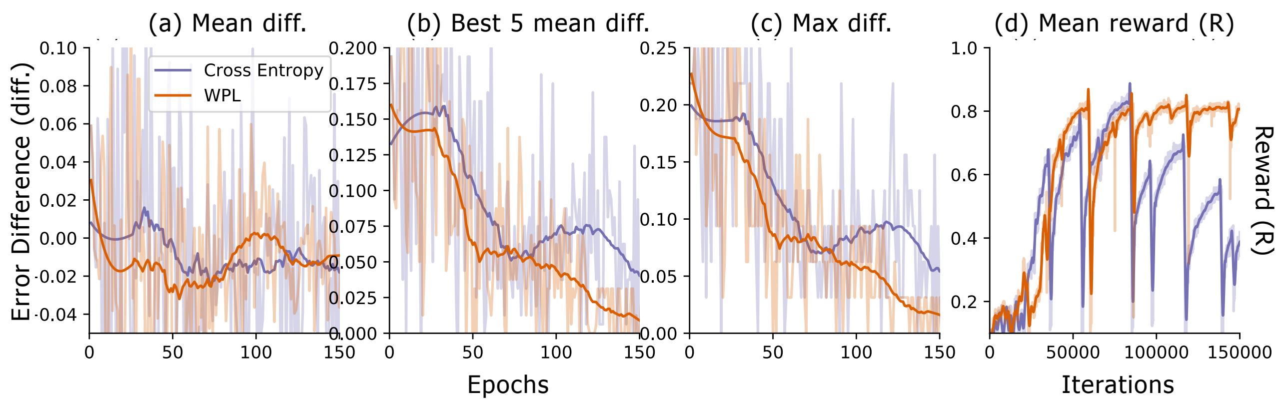
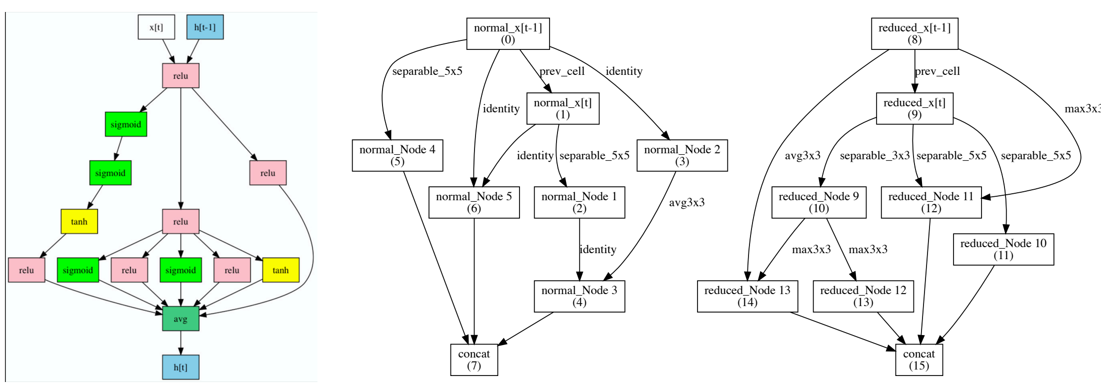


Figure 1. **Error differences when searching for CNN architectures.** Quantitatively, the multi-model forgetting effect is reduced by up to 99% for *(a)*, 96% for *(b)*, and 98% for *(c)*.



(a) Searched RNN cell (b) Searched Micro-CNN normal cell (c) Searched Micro-CNN reduction cell

Figure 2. **Best architectures found for RNN and CNN.** We display the best architecture found by ENAS+WPL, in *(a)* for the RNN cell, and in *(b)* and *(c)* for the CNN normal and reduction cells.