Supplementary Material
Theoretical Understanding of the Information Flow on Continual Learning Performance

Joshua Andle and Salimeh Yasaei Sekeh

University of Maine, Orono ME 04469, USA
salimeh.yasaei@maine.edu

Here we first provide the full proofs of the theorems, and then validate our analysis using the Permuted MNIST dataset. These experiments demonstrate the effects that reducing the amount pruned on highly-connected layers has on the performance of a VGG16 network and the connectivities between its layers.

Let us introduce the notations used in the Supplementary Material. We are given a sequence of joint random variables $(\mathbf{X}_t, T_t)$, with realization space $\mathcal{X}_t \times \mathcal{T}_t$ where $(\mathbf{x}_t, y_t)$ is an instance of the $\mathcal{X}_t \times \mathcal{T}_t$ space. We assume that a given DNN has a total of L layers where,

- $F^{(L)}$: A function mapping the input space $\mathcal{X}$ to a set of classes $\mathcal{T}$, i.e. $F^{(L)} : \mathcal{X} \mapsto \mathcal{T}$.
- $f^{(l)}$: The l-th layer of $F^{(L)}$ with M_l as number of filters in layer l.
- $f_i^{(l)}$: i-th filter in layer l.
- $F^{(i,j)} := f^{(j)} \circ \ldots \circ f^{(i)}$: A subnetwork which is a group of consecutive layers $f^{(i)}, \ldots, f^{(j)}$.
- $F^{(j)} := F^{(1,j)} = f^{(j)} \circ \ldots \circ f^{(1)}$: First part of the network up to layer j.
- $\sigma^{(l)}$: The activation function in layer l.
- $\widetilde{f}_t^{(l)}$: Sensitive layer for task t.
- $\widetilde{F}_t^{(L)} := F_t^{(L)} / \widetilde{f}_t^{(l)}$: The network with L layers when l-th sensitive layer $\widetilde{f}^{(l)}$ is frozen while training on task t.
- $\pi(T_t)$: The prior probability of class label $T_t \in \mathcal{T}_t$.
- η_{tl}, γ_{tl}: Thresholds for sensitivity and usefulness of l-th layer $f^{(l)}$ for task t.
- $\omega^{(1:i)}$: The weight matrix of subnetwork $F^{(1,i)} := f^{(i)} \circ \ldots \circ f^{(1)}$.
- $\omega^{(l)}$: The l-layer's weight matrix
- $\widetilde{\omega}_t^{(l)} = m^{(l)} \odot \omega^{(l)}$: pruned version of the l-layer weight matrix.
- $\widetilde{w}^{(1:L)} = \left(\omega^{(1:l-1)}, \widetilde{\omega}^{(l)}, \omega^{(l+1,L)}\right)$: The weight matrix of network $F^{(L)}$ with pruned l-layer.
- ω_t^* and $\widetilde{\omega}_t^*$: The convergent or optimum parameters after training $F_t^{(L)}$ and $\widetilde{F}_t^{(L)}$ has been finished for task t, respectively.
- $\omega_t^{*(l)}$: The optimal weight set for layer l and trained on task T_t.
- $\widetilde{\omega}_t^{*(l)}$: The optimal weight set for layer l, masked and trained on task T_t.

1 Proof of Theorem 1

Recall Definitions 1 and the Pearson correlation coefficient between the i-th filter in l-layer and j-th filter in $l+1$-layer defined in (2). By conditioning over task labels, the function ρ becomes:

$$\rho(f_i^{(l)}, f_j^{(l+1)}|T_t) = \sum_{y_t \in T_t} \pi(y_t)\mathbb{E}\left[f_i^{(l)}(\mathbf{X}_t) f_j^{(l+1)}(\mathbf{X}_t)|T_t = y_t\right], \tag{1}$$

where $\pi(y_t)$ is the prior probability. Thus the function $\Delta_t(f^{(l)}, f^{(l+1)})$ is written as a proportion of the following term:

$$\begin{aligned}\Delta_t(f^{(l)}, f^{(l+1)}) &\propto \sum_{i=1}^{M_l}\sum_{j=1}^{M_{l+1}}\sum_{y_t \in T_t} \pi(y_t)\mathbb{E}\left[f_i^{(l)}(\mathbf{X}_t) f_j^{(l+1)}(\mathbf{X}_t)|T_t = y_t\right] \\ &= \sum_{i=1}^{M_l}\sum_{j=1}^{M_{l+1}}\sum_{y_t \in T_t} \pi(y_t)\mathbb{E}\left[f_i^{(l)}(\mathbf{X}_t) \cdot \sigma_j(f_i^{(l)}(\mathbf{X}_t))|T_t = y_t\right], \end{aligned} \tag{2}$$

where σ_j is the activation function and $f_j^{(l+1)}(.) = \sigma_j(f_i^{(l)}(.))$. Let σ_j be function $\overline{\sigma}_j(s) = s.\sigma_j(s)$, therefore (2) turns into:

$$\begin{aligned}&\sum_{i=1}^{M_l}\sum_{j=1}^{M_{l+1}}\sum_{y_t \in T_t} \pi(y_t)\mathbb{E}\left[\overline{\sigma}_j\left(f_i^{(l)}(\mathbf{X}_t)\right)|T_t = y_t\right] \\ &= \sum_{i=1}^{M_l}\sum_{y_t \in T_t} \pi(y_t)\mathbb{E}\left[\sum_{j=1}^{M_{l+1}} \overline{\sigma}_j\left(f_i^{(l)}(\mathbf{X}_t)\right)|T_t = y_t\right]. \end{aligned} \tag{3}$$

On the other hand recall Definitions 2, the term in (3) after conditioning on task labels with prior probabilities $\pi(y_t)$ for $y_t \in T_t$ becomes

$$\mathbb{E}_{(\mathbf{X}_t, T_t)\sim\mathcal{D}_t}\left[T_t \cdot G_l \circ f^{(l)}(K_{l-1} \circ \mathbf{X}_t)\right] = \sum_{y_t \in T_t} \pi(y_t)\mathbb{E}_{\mathbf{X}_t|y_t}\left[y_t.G_l \circ f^{(l)}(K_{l-1} \circ \mathbf{X}_t)\right]. \tag{4}$$

For brevity we use $\mathbf{X}_t$ for $K_{l-1} \circ \mathbf{X}_t$. Let G_l be a function that maps layer $f^{(l)}$ to a linear combination of filters i.e.

$$G_l : f^{(l)} \longmapsto \sum_{i=1}^{M_l} f_i^{(l)},$$

Therefore the right hand side of (4) turns into

$$\begin{aligned}&\sum_{y_t \in T_t} \pi(y_t)\mathbb{E}_{\mathbf{X}_t|y_t}\left[\sum_{i=1}^{M_l} y_t \cdot f_i^{(l)}(\mathbf{X}_t)|T_t = y_t\right] \\ &= \sum_{i=1}^{M_l}\sum_{y_t \in T_t} y_t\pi(y_t)\mathbb{E}_{\mathbf{X}_t|y_t}\left[f_i^{(l)}(\mathbf{X}_t)|T_t = y_t\right]. \end{aligned} \tag{5}$$

Set $\beta(s) := \sum_{j=1}^{M_{l+1}} \sigma_j(s)$. We know that there exists a constant C such that $C\,\mathbb{E}[s] \geq \mathbb{E}[s.\beta(s)]$. This implies that

$$C_t \sum_{i=1}^{M_l} \sum_{y_t \in T_t} y_t \pi(y_t) \mathbb{E}_{\mathbf{X}_t|y_t} \left[f_i^{(l)}(\mathbf{X}_t) | T_t = y_t \right]$$
$$\geq \sum_{i=1}^{M_l} \sum_{y_t \in T_t} \pi(y_t) \mathbb{E} \left[\sum_{j=1}^{M_{l+1}} \overline{\sigma}_j \left(f_i^{(l)}(\mathbf{X}_t) \right) | T_t = y_t \right]. \tag{6}$$

Note that here a possible example of $C_t = \sum_{j=1}^{M_{l+1}} U_j$, where U_j is an upper bound of σ_j. Notice here $y_t \geq 1$ for $y_t \in T_t$. Under the assumption that layer $f^{(l)}$ is t-task sensitive i.e. $\Delta_t(f^{(l)}, f^{(l+1)}) \geq \eta_{tl}$, the RHS of (6) is bounded by a proportion of η_{tl}. This combined with (2) implies that

$$\sum_{i=1}^{M_l} \sum_{y_t \in T_t} y_t \pi(y_t) \mathbb{E}_{\mathbf{X}_t|y_t} \left[f_i^{(l)}(\mathbf{X}_t) | T_t = y_t \right] \geq \gamma_{tl}. \tag{7}$$

where $\gamma_{tl} \propto \eta_{tl}/C_t$. This concludes that layer $f^{(l)}$ is t-task useful.

2 Proof of Theorem 2

Let ω_t^* and $\widetilde{\omega}_t^*$ be the convergent or optimum parameters after training $F_t^{(L)}$ and $\widetilde{F}_t^{(L)}$ has been finished for task t, respectively. In addition, training a classifier is performed by minimizing a loss function (via empirical risk minimization (ERM)) that decreases with the correlation between the weighted combination of the features and the label defined in (6):

$$\mathbb{E}_{(\mathbf{X}_t, T_t) \sim D_t} \left\{ L_t(F_t^{(L)}(\mathbf{X}_t), T_t) \right\} = \mathbb{E}_{(\mathbf{X}_t, T_t) \sim D_t} \left\{ T_t \cdot \ell_t(\omega) \right\}. \tag{8}$$

Set $\delta_t(\omega_t^* | \widetilde{\omega}_t^*) := \ell_t(\omega_t^*) - \ell_t(\widetilde{\omega}_t^*)$. The difference between training performance of $F^{(L)}$ and $\widetilde{F}_t^{(L)} := F_t^{(L)} / \widetilde{f}_t^{(l)} \in \mathcal{F}$, the network in which layer l is frozen while training on task t, $d(F_t^{(L)}, \widetilde{F}_t^{(L)})$, defined in (3) is given by

$$\mathbb{E}_{(\mathbf{X}_t, T_t) \sim D_t} \left[T_t \cdot \delta_t(\omega_t^* | \widetilde{\omega}_t^*) \right] = \mathbb{E}_{(\mathbf{X}_t, T_t) \sim D_t} \left[T_t \cdot \left(\ell_t(\omega_t^*) - \ell_t(\widetilde{\omega}_t^*) \right) \right]. \tag{9}$$

Using the arguments in [2] if we write the second order Taylor approximation of ℓ_t around $\widetilde{\omega}_t^*$, we get

$$\ell_t(\omega_t^*) \approx \ell_t(\widetilde{\omega}_t^*) + (\omega_t^* - \widetilde{\omega}_t^*) \nabla \ell_t(\widetilde{\omega}_t^*) + \frac{1}{2} (\omega_t^* - \widetilde{\omega}_t^*)^T \nabla^2 \ell_t(\widetilde{\omega}_t^*) (\omega_t^* - \widetilde{\omega}_t^*), \tag{10}$$

where $\nabla^2 \ell_t(\widetilde{\omega}_t^*)$ is the Hessian for loss ℓ_t at $\widetilde{\omega}_t^*$. Because the model is assumed to converge to a stationary point where the gradient's norm vanishes, $\nabla \ell_t(\widetilde{\omega}_t^*) = 0$:

$$\ell_t(\omega_t^*) - \ell_t(\widetilde{\omega}_t^*) \approx \frac{1}{2}(\omega_t^* - \widetilde{\omega}_t^*)^T \nabla^2 \ell_t(\widetilde{\omega}_t^*)(\omega_t^* - \widetilde{\omega}_t^*). \tag{11}$$

We use the property that the Hessian is positive semi-definite bound (11) by

$$\ell_t(\omega_t^*) - \ell_t(\widetilde{\omega}_t^*) \geq \widetilde{\lambda}_t^{min} \|\omega_t^* - \widetilde{\omega}_t^*\|^2, \tag{12}$$

here $\widetilde{\lambda}_t^{min}$ is the minimum eigenvalue of $\nabla^2 \ell_t(\widetilde{\omega}_t^*)$. This bounds (9) by

$$\frac{1}{2}\mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t}\left[T_t \cdot \left((\omega_t^* - \widetilde{\omega}_t^*)^T \nabla^2 \ell_t(\widetilde{\omega}_t^*)(\omega_t^* - \widetilde{\omega}_t^*)\right)\right] \tag{13}$$

$$\geq \frac{1}{2}\mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t}\left[T_t \cdot \left(\widetilde{\lambda}_t^{min} \|\omega_t^* - \widetilde{\omega}_t^*\|^2\right)\right]. \tag{14}$$

Recall that the l-th layer is defined as $f_t^{(l)} = \sigma_t^{(l)}(\omega_t \mathbf{X}_t)$. There exists a constant $C^{(l)}$ such that

$$\sigma_t^{(l)}\big((\omega_t^* - \widetilde{\omega}_t^*)\mathbf{X}_t\big) \leq C^{(l)} \big|\sigma_t^{(l)}(\omega_t^* \mathbf{X}_t) - \sigma_t^{(l)}(\widetilde{\omega}_t^* \mathbf{X}_t)\big|. \tag{15}$$

where $|.|$ is the element-wise absolute value. Next in both sides of the Ineq. (15) we map $\sigma^{(l)} \in \mathcal{L}_l$ using $G_l : \mathcal{L}_l \mapsto \mathcal{T}_t$, multiple to task T_t and take the expectation:

$$\begin{aligned} &\mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t}\left[T_t \cdot G_l \circ \sigma_t^{(l)}\big((\omega_t^* - \widetilde{\omega}_t^*)\mathbf{X}_t\big)\right] \\ &\leq C^{(l)}\, \mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t}\left[T_t \cdot G_l \circ \big|\sigma_t^{(l)}(\omega_t^* \mathbf{X}_t) - \sigma_t^{(l)}(\widetilde{\omega}_t^* \mathbf{X}_t)\big|\right]. \end{aligned} \tag{16}$$

Given distribution $\mathcal{D}_t$, assuming that the l-th layer is t-task-useful, (3), we have

$$\mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t}\left[T_t \cdot G_l \circ \sigma_t^{(l)}\big(\omega_t \mathbf{X}_t\big)\right] \geq \gamma_{tl}, \tag{17}$$

Let $\omega_t = \omega_t^* - \widetilde{\omega}_t^*$ in (17) and combined with (16), we get

$$\mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t}\left[T_t \cdot G_l \circ \big|\sigma_t^{(l)}(\omega_t^* \mathbf{X}_t) - \sigma_t^{(l)}(\widetilde{\omega}_t^* \mathbf{X}_t)\big|\right] \geq \widetilde{\gamma}_{tl}, \tag{18}$$

where $\widetilde{\gamma}_{tl} = \gamma_{tl}/C^{(l)}$. We assume that the activation $\sigma^{(l)}$ is Lipschitz continuous since it is generally true for most of the commonly used activations in neural networks such as Identity, ReLU, sigmoid, tanh, PReLU, etc. Then we know for for any $\mathbf{z}, \mathbf{s}$, there exist a constant $C_\sigma^{(l)}$ such that

$$|\sigma^{(l)}(\mathbf{z}) - \sigma^{(l)}(\mathbf{s})| \leq C_\sigma^{(l)} |\mathbf{z} - \mathbf{s}|.$$

Then it is easy to see that

$$\big|\sigma_t^{(l)}(\omega_t^* \mathbf{X}_t) - \sigma_t^{(l)}(\widetilde{\omega}_t^* \mathbf{X}_t)\big| \leq C_\sigma^{(l)} |\omega_t^* - \widetilde{\omega}_t^*||\mathbf{X}_t|, \tag{19}$$

We first apply two map functions G_l and $\overline{G}_l$ on left and right sides of the above inequality to map them to the space $\mathcal{T}_t$, second we multiply T_t, and take expectation with respect to distribution $\mathcal{D}_t$:

$$\begin{aligned}&\mathbb{E}_{(\mathbf{X}_t,T_t)\sim D_t}\left[T_t\cdot G_l\circ\left|\sigma_t^{(l)}(\omega_t^*\mathbf{X}_t)-\sigma_t^{(l)}(\widetilde{\omega}_t^*\mathbf{X}_t)\right|\right]\\&\quad\leq C_\sigma^{(l)}\,\mathbb{E}_{(\mathbf{X}_t,T_t)\sim D_t}\left[T_t\cdot\overline{G}_l\circ|\omega_t^*-\widetilde{\omega}_t^*||\mathbf{X}_t|\right].\end{aligned}\tag{20}$$

Note that ω_t^* and $\widetilde{\omega}_t^*$ are trained weight matrix from layers 1 to l with layer l included and excluded in training respectively. Combining (18), (19), and (20), we have

$$\widetilde{\gamma}_{tl}\leq C_\sigma^{(l)}\,\mathbb{E}_{(\mathbf{X}_t,T_t)\sim D_t}\left[T_t\cdot\overline{G}_l\circ|\omega_t^*-\widetilde{\omega}_t^*||\mathbf{X}_t|\right].\tag{21}$$

Since $|\mathbf{X}_t|$ is bounded, there exist a constant C_x such that $|\mathbf{X}_t|\leq C_x$. Thus, we have

$$C_\gamma\leq\mathbb{E}_{(\mathbf{X}_t,T_t)\sim D_t}\left[T_t\cdot\overline{G}_l\circ|\omega_t^*-\widetilde{\omega}_t^*|\right].\tag{22}$$

where $C_\gamma=\widetilde{\gamma}_{tl}/C_xC_\sigma^{(l)}$. Let $\overline{G}_l:|\omega_t^*-\widetilde{\omega}_t^*|\mapsto\widetilde{\lambda}_t^{min}\|\omega_t^*-\widetilde{\omega}_t^*\|^2$. This implies

$$K(\gamma_{tl})\leq\mathbb{E}_{(\mathbf{X}_t,T_t)\sim D_t}\left[T_t\cdot\left(\widetilde{\lambda}_t^{min}\|\omega_t^*-\widetilde{\omega}_t^*\|^2\right)\right],\tag{23}$$

is lower bounded by $K(\gamma_{tl})\propto\gamma_{tl}/(C^{(l)}C_xC_\sigma^{(l)}$ which is an increasing function of γ_{tl}. This concludes the proof and shows that in (9),

$$d(F_t^{(L)},\widetilde{F}_t^{(L)})=\mathbb{E}_{(\mathbf{X}_t,T_t)\sim D_t}\left[T_t\cdot\delta_t(\omega_t^*|\widetilde{\omega}_t^*)\right]\geq K(\gamma_{tl}).$$

3 Proof of Theorem 4

Let $\widetilde{\omega}_{t+1}^*$ be the optimal weight after training $\widetilde{F}_{t+1}^{(L)}$ on task $t+1$. Following the arguments and notations in the proof of Theorem 2:

$$\begin{aligned}&\mathbb{E}_{(\mathbf{X}_t,T_t)\sim D_t}\left\{L_t(\widetilde{F}_{t+1}^{(L)}(\mathbf{X}_t),T_t)-L_t(F_t^{(L)}(\mathbf{X}_t),T_t)\right\}\\&\quad=\mathbb{E}_{(\mathbf{X}_t,T_t)\sim D_t}\left\{T_t\cdot\left(\ell_t(\widetilde{\omega}_{t+1}^*)-\ell_t(\omega_t^*)\right)\right\}.\end{aligned}\tag{24}$$

Subtract and add the term $\ell_t(\widetilde{\omega}_t^*)$ in (3):

$$\begin{aligned}&\mathbb{E}_{(\mathbf{X}_t,T_t)\sim D_t}\left\{T_t\cdot\left(\ell_t(\widetilde{\omega}_{t+1}^*)-\ell_t(\widetilde{\omega}_t^*)\right)+\left(\ell_t(\widetilde{\omega}_t^*)-\ell_t(\omega_t^*)\right)\right\}\\&=\mathbb{E}_{(\mathbf{X}_t,T_t)\sim D_t}\left\{T_t\cdot\left(\ell_t(\widetilde{\omega}_{t+1}^*)-\ell_t(\widetilde{\omega}_t^*)\right)\right\}+\mathbb{E}_{(\mathbf{X}_t,T_t)\sim T_t}\left\{\left(\ell_t(\widetilde{\omega}_t^*)-\ell_t(\omega_t^*)\right)\right\}\end{aligned}\tag{25}$$

Using the arguments in [2] we know that

$$\mathbb{E}_{(\mathbf{X}_t,T_t)\sim D_t}\left\{T_t\cdot\left(\ell_t(\widetilde{\omega}_{t+1}^*)-\ell_t(\widetilde{\omega}_t^*)\right)\right\}\leq\frac{1}{2}\mathbb{E}_{(\mathbf{X}_t,T_t)\sim D_t}\left\{T_t\cdot\widetilde{\lambda}_t^{max}\|\widetilde{\omega}_{t+1}^*-\widetilde{\omega}_t^*\|^2\right\},\tag{26}$$

where $\widetilde{\lambda}_t^{max}$ is maximum eigenvalue of $\nabla^2 \ell_t(\widetilde{\omega}_t^*)$. Let $\widetilde{w}_t'$ be the convergent or (near-)optimum parameters after training $\widetilde{F}_t^{(L)}$ has been finished for the first task. Then

$$\|\widetilde{\omega}_{t+1}^* - \widetilde{\omega}_t^*\| \leq \|\widetilde{\omega}_{t+1}^* - \widetilde{\omega}_t'\| + \|\widetilde{\omega}_t' - \widetilde{\omega}_t^*\|. \tag{27}$$

Since $\|\widetilde{\omega}_{t+1}^* - \widetilde{\omega}_t'\|$ is a constant, say C, we only need to bound $\|\widetilde{\omega}_t' - \widetilde{\omega}_t^*\|$. Consider two different convergence criterion:

- $\ell_t(\widetilde{\omega}_t') - \ell_t(\widetilde{\omega}_t^*) \leq \epsilon$: We write the second order Taylor approximation of ℓ_t around $\widetilde{\omega}_t^*$:

$$\ell_t(\widetilde{\omega}_t') - \ell_t(\widetilde{\omega}_t^*) \approx \frac{1}{2}(\widetilde{\omega}_t' - \widetilde{\omega}_t^*)^T \nabla^2 \ell_t(\widetilde{\omega}_t^*)(\widetilde{\omega}_t' - \widetilde{\omega}_t^*) \leq \frac{1}{2}\widetilde{\lambda}_t^{max} \|\widetilde{\omega}_t' - \widetilde{\omega}_t^*\|^2, \tag{28}$$

  where $\widetilde{\lambda}_t^{max}$ is the maximum eigenvalue of $\nabla^2 \ell_t(\widetilde{\omega}_t^*)$. Hence the convergence criterion can be written as $\frac{1}{2}\widetilde{\lambda}_t^{max} \|\widetilde{\omega}_t' - \widetilde{\omega}_t^*\|^2 \leq \epsilon$, equivalently

$$\|\widetilde{\omega}_t' - \widetilde{\omega}_t^*\|^2 \leq \frac{2\sqrt{\epsilon}}{\widetilde{\lambda}_t^{max}} \tag{29}$$

- $\nabla^2 \ell_t(\widetilde{w}_t') \leq \epsilon$: Write the first order Taylor approximation of $\nabla \ell_t$ around $\widetilde{\omega}_t^*$:

$$\nabla \ell_t(\widetilde{\omega}_t') - \nabla \ell_t(\widetilde{\omega}_t^*) \approx \nabla^2 \ell_t(\widetilde{\omega}_t^*)(\widetilde{\omega}_t' - \widetilde{\omega}_t^*) \leq \widetilde{\lambda}_t^{max} \|\widetilde{\omega}_t' - \widetilde{\omega}_t^*\|. \tag{30}$$

  Hence the convergence criterion can be written as $\widetilde{\lambda}_t^{max} \|\widetilde{\omega}_t' - \widetilde{\omega}_t^*\| \leq \epsilon$, equivalently

$$\|\widetilde{\omega}_t' - \widetilde{\omega}_t^*\| \leq \frac{\epsilon}{\widetilde{\lambda}_t^{max}}. \tag{31}$$

Denote $C_\epsilon = \max\{\epsilon, 2\sqrt{\epsilon}\}$. Combining (29) and (31) we get the LHS of (26) bounded by

$$\frac{1}{2}\mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t} \left\{ T_t \cdot \widetilde{\lambda}_t^{max} \left(C + \frac{C_\epsilon}{\widetilde{\lambda}_t^{max}} \right)^2 \right\}. \tag{32}$$

Recalling Theorem 1, we obtain

$$\mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t} \left\{ T_t \cdot (\ell_t(\widetilde{\omega}_t^*) - \ell_t(\omega_t^*)) \right\} \leq -K(\gamma_{tl}), \tag{33}$$

where $K(\gamma_{tl})$ is an increasing function of γ_{tl}. Combining (32) and (33) in (3), we have

$$\begin{aligned} &\mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t} \left\{ T_t \cdot \left(\ell_t(\widetilde{\omega}_{t+1}^*) - \ell_t(\omega_t^*) \right) \right\} \\ &\leq \frac{1}{2}\mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t} \left\{ T_t \cdot \widetilde{\lambda}_t^{max} \left(C + \frac{C_\epsilon}{\widetilde{\lambda}_t^{max}} \right)^2 \right\} - K(\gamma_{tl}), \end{aligned} \tag{34}$$

By setting $\epsilon(\widetilde{\lambda}_t^{max}, \gamma_{tl})$ the RHS of (3), the function ϵ is a decreasing function of γ_{tl} for given $\widetilde{\lambda}_t^{max}$ and the theorem is proved.

4 Proof of Theorem 5

Recall

- $\omega^{(1:i)}$: The weight matrix of subnetwork $F^{(1,i)} := f^{(i)} \circ \ldots \circ f^{(1)}$.
- $\omega^{(l)}$: The l-layer's weight matrix
- $\widetilde{\omega}_t^{(l)} = m^{(l)} \odot \omega^{(l)}$: pruned version of the l-layer weight matrix.
- $\widetilde{w}^{(1:L)} = \left(\omega^{(1:l-1)}, \widetilde{\omega}^{(l)}, \omega^{(l+1,L)}\right)$: The weight matrix of network $F^{(L)}$ with pruned l-layer.

The optimal weight matrix $\widetilde{\omega}_{t+1}^*$ with mask $m^*{}_{t+1}$:

$$\widetilde{EO}_t = \mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t}\left\{|T_t \cdot \left(\ell_t(\widetilde{\omega}_{t+1}^*) - \ell_t(\omega_t^*)|\right)\right\}. \tag{35}$$

By adding and subtracting term $\ell_t(\omega_{t+1}^*)$ in (35), we bound EO_t by

$$\begin{aligned}\widetilde{EO}_t \leq\ & \mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t}\left\{T_t \cdot |\left(\ell_t(\widetilde{\omega}_{t+1}^*) - \ell_t(\omega_{t+1}^*)|\right)\right\} \\ & + \mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t}\left\{T_t \cdot |\left(\ell_t(\omega_{t+1}^*) - \ell_t(\omega_t^*)|\right)\right\}. \end{aligned} \tag{36}$$

Once we assume that only one connection is frozen in the training process, we can use the following upper bound of the model [1]:

$$|\ell_t(\widetilde{\omega}_{t+1}^*) - \ell_t(\omega_{t+1}^*| \leq \frac{\|\omega^{*(l)}_{t+1} - \widetilde{\omega}^{*(l)}_{t+1})\|_F}{\|\omega^{*(l)}_{t+1}\|_F} \prod_{j=1}^{L} \|\omega^{*(l)}_{t+1}\|_F, \tag{37}$$

where $\|.\|_F$ is Frobenius norm. Here $\widetilde{\omega}_{t+1}^{*(l)}$ is the optimal weight set for layer l, masked and trained on task T_{t+1}, $\widetilde{\omega}^{*(l)}_{t+1} = m^{*(l)}_{t+1} \odot \omega^{*(l)}_{t+1}$, therefore

$$\|\omega^{*(l)}_{t+1} - \widetilde{\omega}^{*(l)}_{t+1}\|_F = \|\omega^{*(l)}_{t+1} - m^{*(l)}_{t+1} \odot \omega^{*(l)}_{t+1}\|_F = \|\omega^{*(l)}_{t+1}\left(\mathbb{1} - m^{*(l)}_{t+1}\right)\|_F.$$

The assumption $P^{*(l)}_m = \frac{\|m^{*(l)}_{t+1}\|_0}{|\omega^{*(l)}{}_{t+1}|} \to 1$ (a.s.) is equivalent to $\left(\mathbb{1} - m^{*(l)}_{t+1}\right) \to 0$ (a.s.). Therefore,

$$|\ell_t(\widetilde{\omega}_{t+1}^*) - \ell_t(\omega_{t+1}^*)| \longrightarrow 0 \quad (a.s.) \tag{38}$$

This implies that the term in the RHS of (36) convergences to zero. Now from (7) in the proof of Theorem 1, if the l-th layer is fully sensitive i.e. $\eta_{tl} \to 1$ then $\gamma_{tl} \propto 1/C_t$, where C_t is constant. Next using analogous arguments in (26)-(32) in the proof of Theorem 2, we have

$$\begin{aligned}&\mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t}\left\{T_t \cdot |\left(\ell_t(\omega_{t+1}^*) - \ell_t(\omega_t^*)|\right)\right\} \\ &\leq \frac{1}{2}\mathbb{E}_{(\mathbf{X}_t, T_t)\sim D_t}\left\{T_t \cdot \lambda_t^{max}\left(C + \frac{C_\epsilon}{\lambda_t^{max}}\right)^2\right\}, \end{aligned} \tag{39}$$

where λ_t^{max} is the maximum eigenvalue of $\nabla^2 \ell_t(\omega_t^*)$. Here C and C_ϵ are constants.

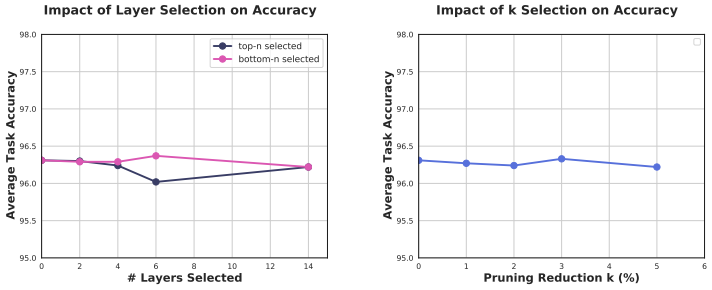

Fig. 1: The average accuracy across tasks of Permuted MNIST is reported for varying values of n when $k = 2\%$ (left) and k when $n = 4$ (right), where n is the number of layers selected for reduced pruning and k is the hyper-parameter dictating how much the pruning on selected layers is reduced by. We compare the performance when the n layers are selected as the most connected layers (top-n) or least connected.

5 Further Experiments

5.1 Experimental Setup

For the experiments outlined here we use a VGG16 model on the Permuted MNIST dataset to determine how the characteristics of information flow differ from what we observed on split CIFAR-10/100. After training on a given task T_t, and prior to pruning, we calculate $\Delta_t(f^{(l)}, f^{(l+1)})$ between each adjacent pair of convolutional or linear layers as in Definition 1, (1). As a baseline we prune 80% of the unfrozen weights in each layer (freezing the remaining 20%), pruning the lowest-magnitude weights. We run a single trial for each experiment.

5.2 Permuted MNIST Experiment

In this section we provide the results of implementing our experiments on the Permuted MNIST dataset. For these experiments we vary the hyper-parameters n and k, where n is the number of layers selected for reduced pruning and k is the percent of pruning reduction provided to the selected layers. By varying n we can see that unlike our observations with CIFAR-10/100, the overall performance remains roughly the same as the baseline (Fig. 1) and selecting the bottom-n layers performed as well as or better than selecting the top-n. To provide an empirical observation of information downstream in the network's layers, we report the differences in connectivities for these experiments as well in Fig. 2. Similarly to our findings with CIFAR-10/100, the changes to n don't appear to substantially alter the patterns in connectivity between layers that we observe throughout the Permuted MNIST. As with CIFAR-10/100, we can see the trends in subsequent tasks where the early layers' connectivities increase throughout the tasks and the later layers decrease.

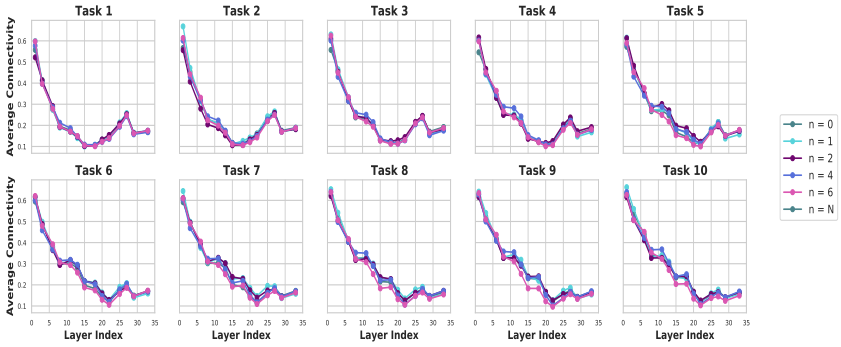

Fig. 2: The average connectivities across layers with the subsequent layer is reported. The scores are plotted for each task in Permuted MNIST, when the n most-connected layers are selected to have their pruning percent reduced by $k = 2\%$.

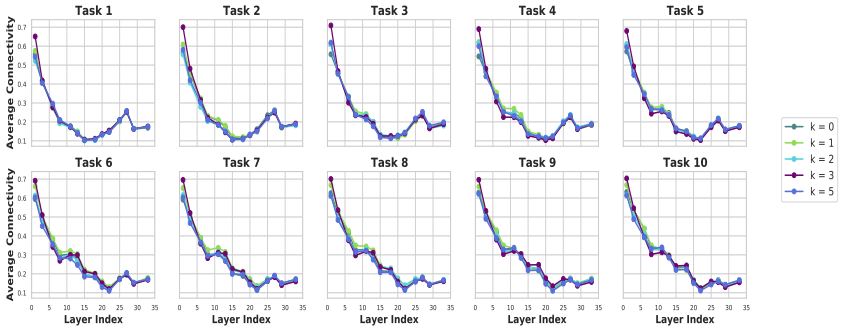

Fig. 3: For each layer the average connectivity value with the subsequent layer is reported. The connectivities are plotted for each task in Permuted MNIST, for various k when $n = 4$ most connected layers are selected for reduced pruning.

We demonstrate the effects of altering k on performance with Permuted MNIST in 1. Once again, unlike with CIFAR-10/100 we observe similar or slightly lower performance compared to the baseline which decreases as the percent continues to increase. The effects of varying k on connectivity can be seen in Fig. 3. The results we observe with Permuted MNIST closely match those seen from CIFAR-10/100 for connectivities but not performances, which indicates that although we observe similar patterns across experiments and tasks for the two datasets, additional steps or a more systematic freezing approach may need to be established to optimally apply our knowledge of information flow to

different models. In addition, this observation provides an experimental evidence that hyperparameters γ_{tl} and η_{tl} are data dependent.

References

1. Lee, J., Park, S., Mo, S., Ahn, S., Shin, J.: Layer-adaptive sparsity for the magnitude-based pruning. In: International Conference on Learning Representations (2020)
2. Mirzadeh, S.I., Farajtabar, M., Pascanu, R., Ghasemzadeh, H.: Understanding the role of training regimes in continual learning. Advances in Neural Information Processing Systems **33**, 7308–7320 (2020)