Adapting by Pruning: A Case Study on BERT
Supplementary Material

1 Convergence analysis

1.1 Definitions

The results presented in this appendix are more general than the convergence bound reported in the main text. To simplify the exposition, we use a slightly different notation. In particular,

- t_l and t_s are called M_{hard} and M_{small}
- $t = 1, \ldots, T$ is used as an upper index to label the SGD epochs (except for the learning parameters α_t)
- $m_{t_l} = \sigma(t_l\theta)$ is called v throughout this appendix
- $\ell(z; m_{t_l}, p)$ and $\mathcal{L}_{t_l}(\theta) = \mathcal{L}_t(\sigma(t\theta))$ of the main text are called $f(v, z)$ and $F(v)$
- Theorem 1 of the main text is Corollary 1.4
- $\mathcal{D}_{train}$ is $\mathcal{D}$
- $\nabla_v f(v, z)$ is $\nabla f(v, z)$ defined by $[\nabla f(v, z)]_i = \frac{\partial}{\partial v'_i} f(v', z)|_{v'=v}$

To avoid confusion, we list of all quantities and conventions used throughout this technical appendix:

- $d \in \mathbf{N}$: number of model parameters
- $\mathcal{Z} = \mathcal{X} \otimes \mathcal{Y}$: object-label space
- P_Z: object-label distribution
- $Z \sim P_Z$: object-label random variable
- $\mathcal{D} = \{z_n \in \mathcal{Z} | z_n \text{ is realization of } Z \sim P_Z\}_{n=1}^N$: training data set
- $f : [0, 1]^d \otimes \mathcal{Z} \to \mathbf{R}$: single-input classification error
- $F : [0, 1]^d \to \mathbf{R}$, $F(v) = \sum_{z \in \mathcal{D}} f(v, z) \approx |\mathcal{D}| \; E_{Z \sim P_Z}(f(v, Z))$: average classification error
- M_{soft} and $M_{hard} > 0$, such that $0 \leq M_{soft} \leq M_{hard}$: soft- and hard-binarization constants
- $\sigma(s) = \frac{1}{1+e^{-s}} \in [0, 1]^d$ for all $s \in \mathbf{R}^d$
- $\sigma'(s) := \sigma(s) \odot (1 - \sigma(s))$, for all $s \in \mathbf{R}^d$

- $[\nabla g(s)]_i = \frac{\partial g(s')}{\partial s'_i}|_{s'=s}$

- $\mathrm{diag}(s) \in \mathbf{R}^{d\times d}$ is such that $[\mathrm{diag}(s)]_{ii} = v_i$ and $[\mathrm{diag}(s)]_{ij} = 0$ if $i \neq j$, $i, j = 1, \ldots, d$

1.2 Assumptions and proofs

Assumption 1. *$f : [0,1]^d \otimes \mathcal{Z} \to \mathbf{R}$, is differentiable over $[0,1]^d$ for all $z \in \mathcal{Z}$ and obeys*

$$\max_{v\in[0,1]^d, z\in\mathcal{Z}} \|\nabla f(v,z)\|^2 \leq G^2, \tag{1}$$

$$f(v,z) - f(v',z) \geq \nabla f(v',z)^T(v-v'), \tag{2}$$

for all $v, v' \in [0,1]^d$ and $z \in \mathcal{Z}$.

Lemma 1.1. *Let $f : [0,1]^d \otimes \mathcal{Z} \to \mathbf{R}$ be the function defined in Assumption 1 and $F : [0,1]^d \to \mathbf{R}$ be defined by*

$$F(v) = \sum_{z\in\mathcal{D}} f(v,z) \approx |\mathcal{D}|\ E_Z(f(v,Z))$$

where $\mathcal{D} = \{z_n \in \mathcal{Z} | z_n$ is realization of $Z \sim P_Z\}_{n=1}^N$. Then

$$F(v) - F(v') \geq \nabla F(v')^T(v-v'), \tag{3}$$

for all $v, v' \in [0,1]^d$ and $z \in \mathcal{X}\times\mathcal{Y}$.

Proof of Lemma 1.1 The convexity of f implies the convexity of F as

$$F(v) - F(v') = |\mathcal{D}|^{-1} \sum_{z\in\mathcal{D}} f(v,z) - f(v',z) \tag{4}$$

$$\geq |\mathcal{D}|^{-1} \sum_{z\in\mathcal{D}} \nabla f(v',z)^T(v-v') \tag{5}$$

$$= \nabla F(v')^T(v-v'). \tag{6}$$

□

Lemma 1.2. *Let $\alpha_t = \frac{c}{\sqrt{t}}$, $c > 0$, and $\theta^t \in \mathbf{R}^d$ be defined by*

$$\theta^{t+1} = \theta^t - \alpha_t M_{soft} \nabla f(\sigma(M_{hard}\theta^t), z_t) \odot \sigma'(M_{soft}\theta^t), \tag{7}$$

where, for all $t = 1, \ldots, T$, $\sigma'(s) = \sigma(s) \odot (1 + \sigma(s))$ ($s \in \mathbf{R}^d$), and z_t is chosen randomly in $\mathcal{D}$. Then, for all $t = 1, \ldots, T$, $v^t = \sigma(M_{hard}\theta^t)$ obeys

$$v^{t+1} = v^t - \alpha_t(\nabla f(v^t, z_t) + r^t), \tag{8}$$

where

$$r^t = \nabla f(v^t, z) - M_{hard}M_{soft}\sigma'(M_{hard}\xi^t) \odot \nabla f(v^t, z) \odot \sigma'(M_{soft}\theta^t) \tag{9}$$

$$\xi^t \in [\theta^t, \theta^t - \alpha_t M_{soft} \nabla f(v^t, z) \odot \sigma'(M_{soft}\theta^t)] \tag{10}$$

for any $z_t \in \mathcal{Z}$. Furthermore, for all $t = 1, \ldots, T$, r^t, obeys

$$\|r^t\|^2 \leq G^2 C \tag{11}$$

where G is defined in (1) and

$$C = M_{hard} M_{soft} \left(\frac{1}{M_{hard} M_{soft}} - 2 g_{max}(M_{hard}) g_{max}(M_{soft}) + \frac{M_{hard} M_{soft}}{16^2} \right) \tag{12}$$

$$g_{max}(M) = \sigma(M\theta_{max})(1 - \sigma(M\theta_{max})) \tag{13}$$

$$\theta_{max} = \max_t |\theta^t| \tag{14}$$

Proof of Lemma 1.2 Let $\sigma_M(s) = \sigma(Ms)$, $\sigma'_M(s) = M\sigma(Ms)(1-\sigma(Ms))$, and $v = \sigma_M(\theta)$, for any $\theta \in \mathbf{R}^d$ and $M > 0$. Then (7) is equivalent to (8) as

$$v^{t+1} = \sigma_{M_{hard}} \left(\theta^t - \alpha_t \nabla f(\sigma_{M_{hard}}(\theta^t), z) \odot \sigma'_{M_{soft}}(\theta^t) \right) \tag{15}$$

$$= v^t - \alpha_t \sigma'_{M_{hard}}(\xi^t) \odot \nabla f(\sigma_{M_{hard}}(\theta^t), z) \odot \sigma'_{M_{soft}}(\theta^t) \tag{16}$$

$$= v^t - \alpha_t \nabla f(v^t, z) + \alpha_t r^t \tag{17}$$

$$r^t = \nabla f(v^t, z) - \operatorname{diag}\left(\sigma'_{M_{hard}}(\xi^t) \odot \sigma'_{M_{soft}}(\theta^t) \right) \cdot \nabla f(\sigma_{M_{hard}}(\theta^t), z) \tag{18}$$

$$= \nabla f(v^t, z) - \operatorname{diag}\left(\sigma'_{M_{hard}}(\xi^t) \odot \sigma'_{M_{soft}}(\theta^t) \right) \cdot \nabla f(v^t, z) \tag{19}$$

$$\xi^t \in [\theta^t, \theta^t - \alpha_t \nabla f(v^t, z) \odot \sigma'_{M_{soft}}(\theta^t)] \tag{20}$$

where the first equality follows from $\sigma_M(a+b) - \sigma_M(a) = \sigma'_M(\xi)\ b$ for some $\xi \in [a, a+b]$ (mean value theorem). For any $\theta \in \mathbf{R}$ and $z \in \mathcal{Z}$, one has

$$\|r^t\|^2 = \|\nabla f(v^t, z)\|^2 + \|\operatorname{diag}\left(\sigma'_{M_{hard}}(\xi^t) \odot \sigma'_{M_{soft}}(\theta^t) \right) \cdot \nabla f(v^t, z)\|^2 \tag{21}$$

$$- 2\nabla f(v^t, z)^T \operatorname{diag}\left(\sigma'_{M_{hard}}(\xi^t) \odot \sigma'_{M_{soft}}(\theta^t) \right) \cdot \nabla f(v^t, z) \tag{22}$$

$$\leq G^2 \left(1 - 2 \min \sigma'_{M_{hard}}(\xi^t) \odot \sigma'_{M_{soft}}(\theta^t) + \left(\max \sigma'_{M_{hard}}(\xi^t) \odot \sigma'_{M_{soft}}(\theta^t) \right)^2 \right) \tag{23}$$

$$\leq G^2 \left(1 - 2\sigma'_{M_{hard}}(\theta_{max}) \sigma'_{M_{soft}}(\theta_{max}) + \left(\frac{M_{hard} M_{soft}}{16} \right)^2 \right) \tag{24}$$

$$\leq G^2 M_{hard} M_{soft} \left(\frac{1}{M_{hard} M_{soft}} - 2 g_{max}(M_{hard}) g_{max}(M_{soft}) + \frac{M_{hard} M_{soft}}{16^2} \right) \tag{25}$$

$$= G^2 C \tag{26}$$

where G is defined in Assumption 1, $\min a$ and $\max a$ are the smallest and largest entries of $a \in \mathbf{R}^d$, $\theta_{max} = \max_t |\theta^t|$,

$$g_{max}(M) = \sigma_M(\theta_{max})(1 - \sigma_M(\theta_{max})) = \sigma(M\theta_{max})(1 - \sigma(M\theta_{max})),$$

and we have used

$$\max_{s \in \mathbf{R}} \sigma'_M(s) = \sigma'_M(0) = \frac{M}{4}, \quad \min_{s \in [-a,a]} = \sigma'_M(s) = \sigma'(-a) = \sigma'(a)$$

for any $M > 0$, the Cauchy-Schwarz inequality $s^T s' \leq \|s\|\|s'\|$, and $s^T \mathrm{diag}(s') \cdot s \leq (\max s')\|s\|^2$, and defined

$$C = M_{hard}M_{soft}\left(\frac{1}{M_{hard}M_{soft}} - 2g_{max}(M_{hard})g_{max}(M_{soft}) + \frac{M_{hard}M_{soft}}{16^2}\right)$$

□

Theorem 1.3. *Let $f : [0,1]^d \otimes \mathcal{Z} \to \mathbf{R}$ and $F : [0,1]^d \to \mathbf{R}^d$ be defined as in Assumption 1 and Lemma 1.1. Let v^t and $\alpha_t = \frac{c}{\sqrt{t}}$ $(t = 1, \ldots, T,\ c > 0)$ be the sequence of weights defined in (8) and a sequence of decreasing learning rates and $v^* = \arg\min_{v \in [0,1]^d} f(v)$. Then*

$$E\left(F(v^T) - F(v^*)\right) \leq \frac{1}{c\sqrt{T}} + \frac{cG^2(1+C)(1+\log T)}{T} \tag{27}$$

where the expectation is over $Z \sim P_Z$ and G and C are defined in Lemma 1.2.

Proof of Theorem 1.3 Assumption (1) and Lemma (1.1) imply that $F : [0,1]^d \to \mathbf{R}$ is a convex function over $[0,1]^d$. The first part of Lemma 1.2 implies that the sequence of approximated $\mathbf{R}^d$-valued gradient updates (7) can be rewritten as the sequence of approximated $[0,1]^d$-valued gradient updates (8). The second part of Lemma 1.2 implies that the norm of all error terms in (8) is bounded from above. In particular, as each r^t is multiplied by the learning rate, $\alpha_t = \frac{c}{t}$, it is possible to show that (8) converges to a local optimum of $F : [0,1]^d \to \mathbf{R}$.

To show that (8) converges to a local optimum of $F : [0,1]^d \to \mathbf{R}$ we follows a standard technique for proving the convergence of stochastic and make the further (standard) assumption

$$E(r^t) = 0, \quad t = 1, \ldots, T. \tag{28}$$

First, we let v^t, r^t and α_t, $t = 1, \ldots, T$, be defined as in Lemma 1.2, and $z \in \mathcal{D}$ be the random sample at iteration $t+1$. Then

$$\begin{aligned}
\|v^{t+1} - v^*\|^2 &= E\left(\|v^{t+1} - v^*\|^2\right) && (29)\\
&= \|v^t - v^*\|^2 - 2\alpha_t E(\nabla f(v^t, z) - r^t)^T (v^t - v^*) && (30)\\
&\quad + \alpha_t^2 E\left(\|\nabla f(v^t, \tilde{z}) - r^t\|^2\right)\\
&\leq \|v^t - v^*\|^2 - 2\alpha_t E(\nabla f(v^t, z)^T (v^t - v^*) + \alpha_t^2\left(G^2 + E(\|r^t\|^2)\right) && (31)\\
&\leq \|v^t - v^*\|^2 + 2\alpha_t E(F(v^*) - F(v^t)) + \alpha_t^2 G^2(1+C) && (32)
\end{aligned}$$

where G^2 and C are defined in (1) and (21). Rearranging the terms one obtains

$$E(F(v^t) - F(v^*)) \leq \frac{\|v^t - v^*\|^2 - \|v^{t+1} - v^*\|^2}{2\alpha_t} + \frac{\alpha_t}{2} G^2(1+C) \tag{33}$$

Since $F(v^T) - F(v^*) = \frac{1}{T}\sum_{t=1}^{T}\left(F(v^t) - F(v^*)\right)$, the bound in (33) implies

$$E\left(F(v^T) - F(v^*)\right) = E\left(\frac{1}{T}\sum_{t=1}^{T}\left(F(v^t) - F(v^*)\right)\right) \tag{34}$$

$$\leq \frac{1}{T}\sum_{t=1}^{T} E\left(F(v^t) - F(v^*)\right) \tag{35}$$

$$\leq \frac{1}{2cT}\sum_{t=1}^{T}\left(\sqrt{t}(\|v^t - v^*\|^2 - \|v^{t+1} - v^*\|^2) + \frac{c^2G^2(1+C)}{\sqrt{t}}\right) \tag{36}$$

$$= \frac{1}{2cT}\left(\sum_{t=1}^{T}(\sqrt{t+1} - \sqrt{t})\|v^t - v^*\|^2 - T\|v^{T+1} - v^*\|^2 + c^2G^2(1+C)\sum_{t=1}^{T}\frac{1}{t}\right) \tag{37}$$

$$\leq \frac{1}{2cT}\left(\sqrt{T}\max_t\{\|v^t - v^*\|^2\}_{t=1}^{T} + c^2G^2(1+C)(1+\log T)\right) \tag{38}$$

$$\leq \frac{1}{c\sqrt{T}} + \frac{cG^2(1+C)(1+\log T)}{T} \tag{39}$$

where the second line follows from the Jensen's inequality and we have used $\sum_{t=1}^{T}(\sqrt{t+1} - \sqrt{t}) \leq \sqrt{T}$ and $\sum_{t=1}^{T}\frac{1}{t} \leq 1 + \log T$. □

Corollary 1.4. *Let $f : [0,1]^d \otimes \mathcal{Z} \to \mathbf{R}$ and $F : [0,1]^d \to \mathbf{R}^d$ be defined as in Assumption 1 and Lemma 1.1. Let θ^t be a sequences of stochastic weight updates defined by*

$$\theta^{t+1} = \theta^t - \alpha_t M_{soft}\nabla f(\sigma(M_{hard}\theta^t), z_t) \odot \sigma'(M_{soft}\theta^t)$$

where, for all $t = 1, \ldots, T$, $\alpha_t = \frac{c}{\sqrt{t}}$ $(c > 0)$ and z_t is chosen randomly in $\mathcal{D}$. Then

$$E(F(\sigma(M_{hard}\theta^T)) - F(\sigma(M_{hard}\theta^*)) \leq \frac{1}{c\sqrt{T}} + \frac{cG^2(1+C)(1+\log T)}{T}$$

where the expectation is over $Z \sim P_Z$, $\theta^* = \arg\min_{\theta\in\mathbf{R}^d} F(\sigma(M_{hard}\theta)$, G is defined in (1), and

$$C = M_{hard}M_{soft}\left(\frac{1}{M_{hard}M_{soft}} - 2g_{max}(M_{hard})g_{max}(M_{soft}) + \frac{M_{hard}M_{soft}}{16^2}\right) \tag{40}$$

$$g_{max}(M) = \sigma(M\theta_{max})(1 - \sigma(M\theta_{max})), \quad \theta_{max} = \max_t|\theta^t| \tag{41}$$

Proof of Corollary 1.4 The Corollary follows directly from Theorem 1.3 because σ is a strictly increasing function, which implies that the mappings $\mathbf{R}^d \to [0,1]^d$ and $[0,1]^d \to \mathbf{R}^d$ are both one-to-one. In particular, for all $t = 1, \ldots, T$ one has $\theta^t = \frac{1}{M_{hard}}\sigma^{-1}(v^t)$ and

$$\theta^* := \arg\min_{\theta\in\mathbf{R}^d} F(\sigma(M_{hard}\theta)) \tag{42}$$

$$= \frac{1}{M_{hard}}\sigma^{-1}\left(\arg\min_{v\in[0,1]^d} F(v)|_{v=\sigma(M_{hard}\theta)}\right) \tag{43}$$

$$= \frac{1}{M_{hard}}\sigma^{-1}(v_*) \tag{44}$$

Hyper-parameter	Candidate and selected values
batch size	4, **8**(all), 10
α (max learning rate for θ)	2e-4, **2e-5**(all), 2e-6
β (max learning rate for p)	2e-4, **2e-5**(all), 2e-6
γ (max regularisation)	2e-5, 2e-6, **2e-7**(all), 2e-8
num. of epoch	**2**(MNLI), **3**(SST-2), **5**(STS-B), 10
t_l	$\mathbf{10^5}$(all), 10^3, 10
t_s	10^{-2}, 10^{-1}, **1**(all)

Table 1: Hyper-parameters tested and selected for our adapt-by-pruning algorithm (Alg. 1 in the main paper). Values in boldface are the ones yielding best performance on the validation set for a certain task. For batch size, due to the hardware limit, we cannot test larger values.

with $v^* := \arg\min_{v\in[0,1]} F(v)$. It follows that $F(v^t) \to F(v*)$ implies $F(\sigma(M_{hard}\theta^t)) \to F(\sigma(M_{hard}\theta^*)))$. $\square$

2 Hyper-Parameters Selection

The hyper-parameters selection details are presented in Table 1.

3 Detailed Main Results

Table 2 compares the performance of different model adaptation and pruning methods; these data are illustrated in Fig. 2 in the main paper.

4 Topological Analysis

The component-wise and layer-wise sparsity comparison of sub-networks obtained by different pruning methods are presented in Fig. 1 (on SST-2) and Fig. 2 (on STS-B).

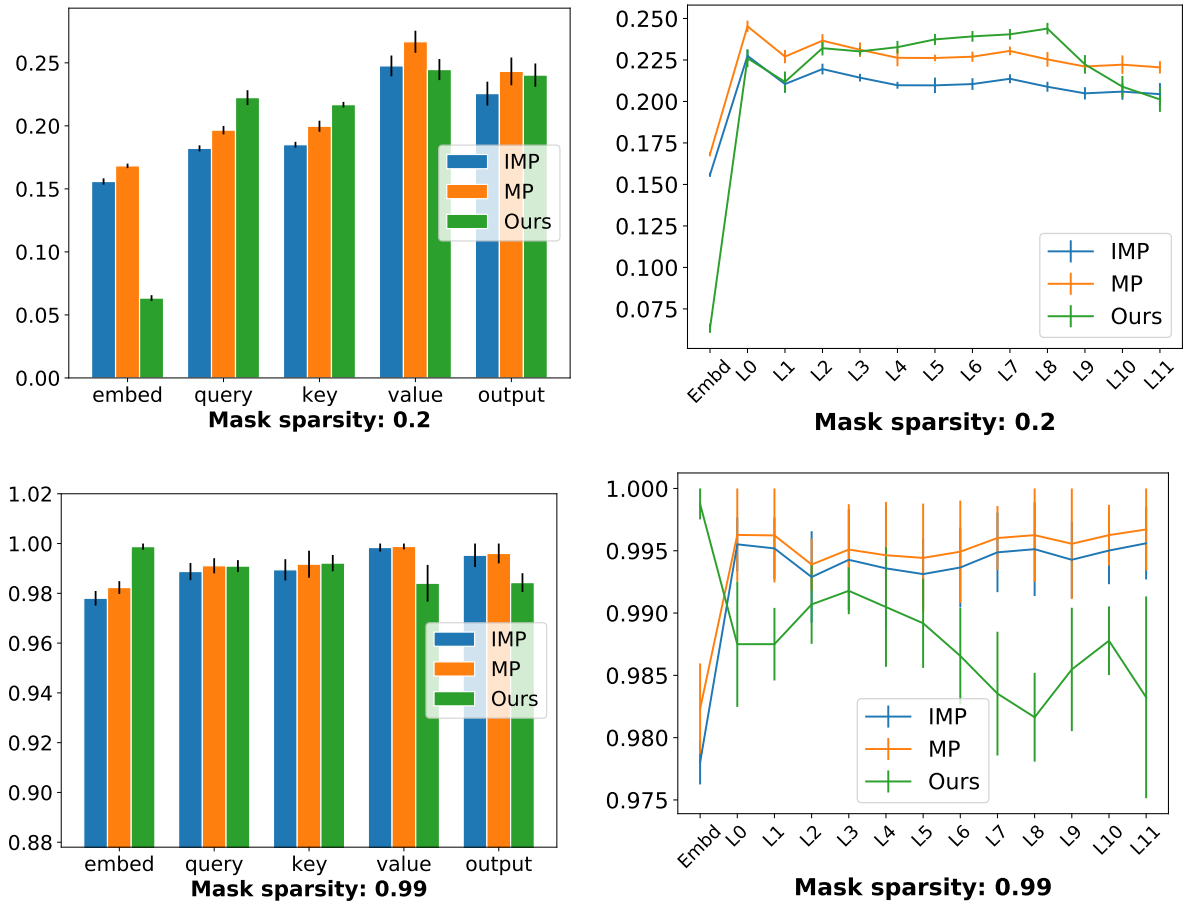

Figure 1: Component- (left) and layer-wise (right) sparsity of the sub-networks obtained by our method and IMP/MP, on the SST-2 dataset.

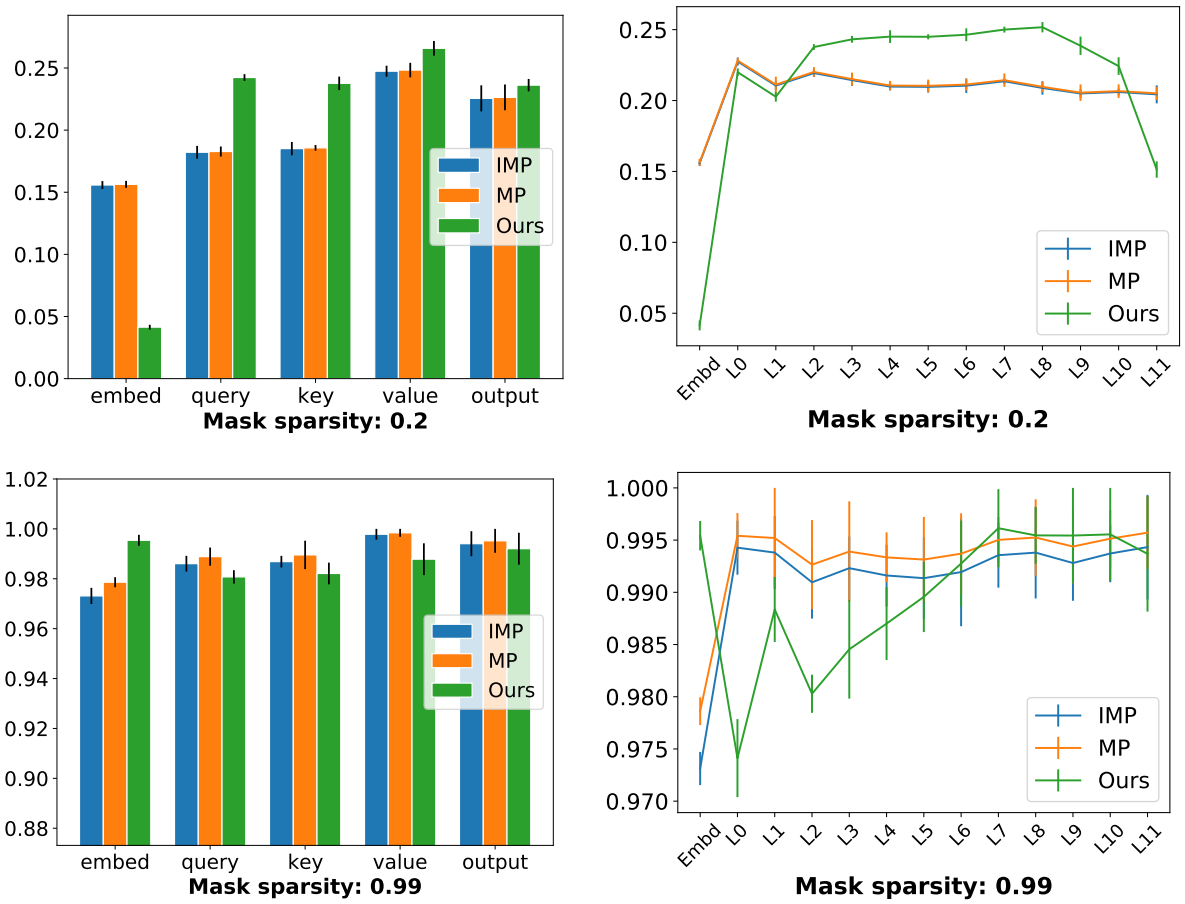

Figure 2: Component- (left) and layer-wise (right) sparsity of the sub-networks obtained by our method and IMP/MP, on the STS-B dataset.

Sparsity	STS-B (rho)				SST-2 (acc)				MNLI (acc)			
	Rnd	IMP	MP	Ours	Rnd	IMP	MP	Ours	Rnd	IMP	MP	Ours
Feat. Ext.	.761 ± .003				.867 ± .004				.585 ± .001			
Fine-Tune	.855 ± .013				.926 ± .017				.843 ± .014			
0.2	.193	.743	.739	**.794**	.645	.862	.863	**.908**	.372	.674	.675	**.769**
0.5	.126	.707	.706	**.837**	.558	.825	.819	**.856**	.341	.641	.648	**.709**
0.7	.016	.383	.607	**.720**	.509	.806	.793	**.843**	.331	.551	.564	**.696**
0.9	.009	.203	.147	**.624**	.509	.717	.722	**.836**	.322	.458	.436	**.679**
0.95	.002	.077	.095	**.600**	.509	.691	.509	**.831**	.333	.357	.343	**.671**
0.99	.006	.067	-.022	**.547**	.506	.509	.509	**.810**	.337	.348	.341	**.628**
0.999	NA	.028	.023	**.301**	NA	.509	.509	**.772**	NA	.341	.341	**.446**
0.9999	NA	NA	NA	NA	NA	.509	.509	.509	NA	.341	.341	.341
Rnd bsl	.003 ± .026				.509 ± .018				.341 ± .003			

Table 2: Performance of subnetworks obtained by different pruning algorithms. Performance (mean± stddev) of feature extraction, fine-tuning and random baselines are also presented. Results are averaged over 100 (Rnd and random baseline) or five runs (the others) with different random seeds. NA: no results produced as some layers in the network are completely pruned. Boldface: significantly better results.