ICLR

Composable Sparse Subnetworks via Maximum-Entropy Principle

Francesco Caso, Samuele Fonio, Simone Monaco, Nicola Saccomanno, Fabrizio Silvestri

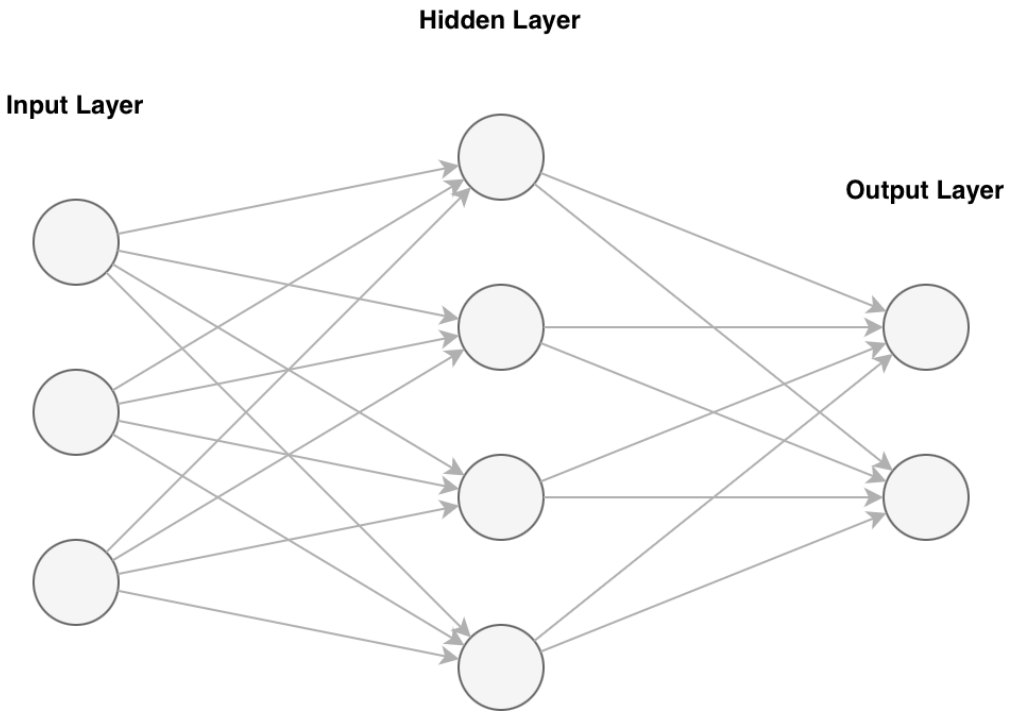
UNIVERSITÀ
DI TORINO

Politecnico
di Torino

UNIVERSITÀ
DEGLI STUDI
DI UDINE

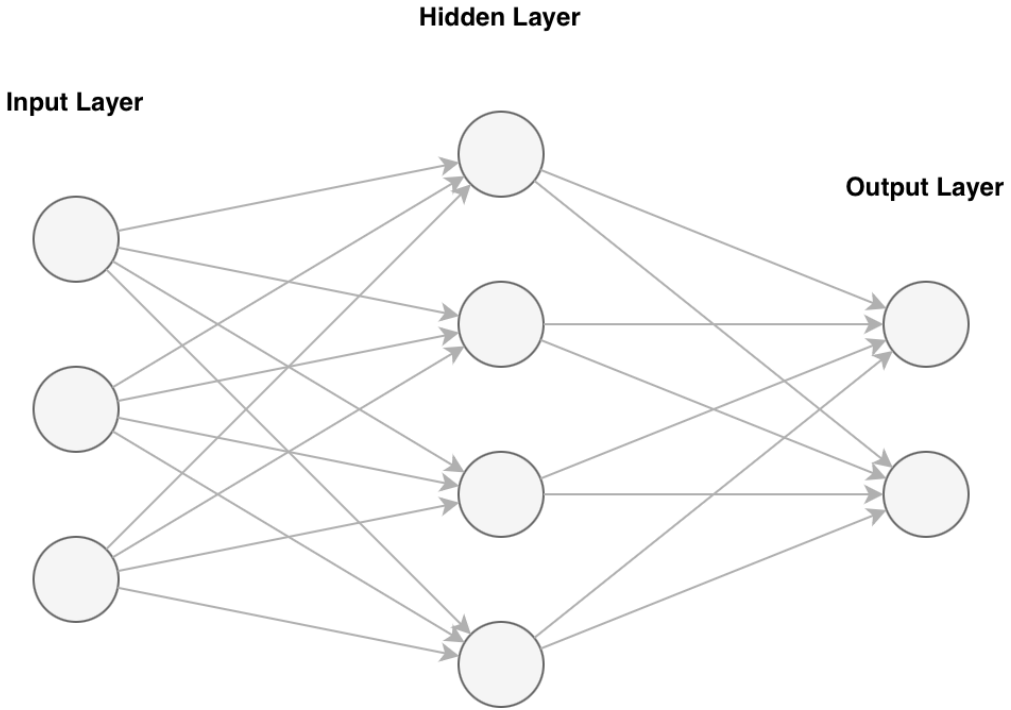
hic sunt futura

Neural networks learn class-specific functional modules



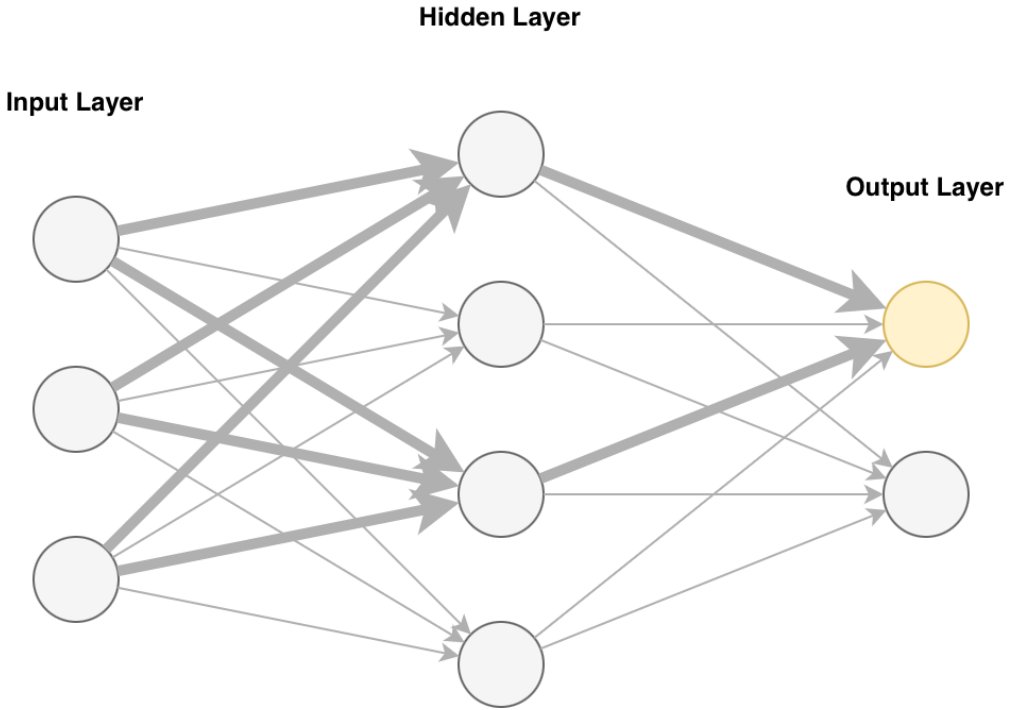
Olah et al. 2020

Neural networks learn class-specific functional modules



Olah et al. 2020

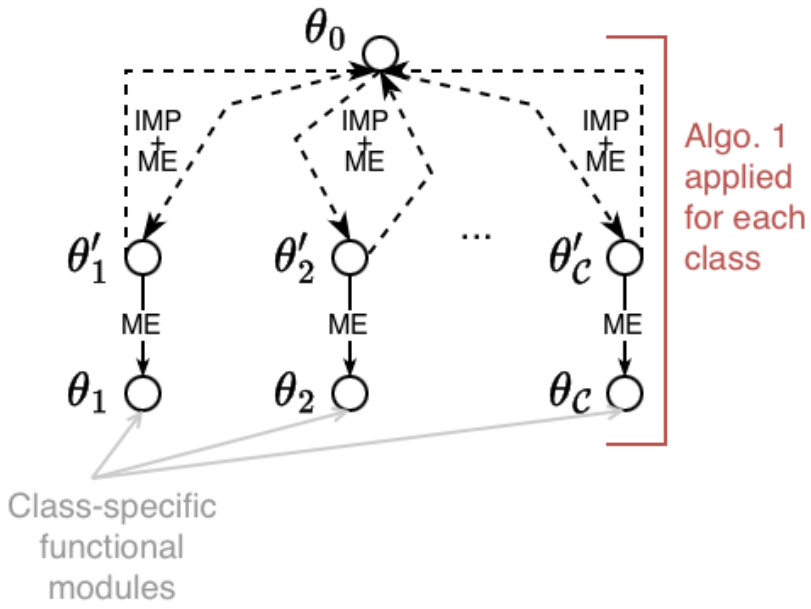
Neural networks learn class-specific functional modules



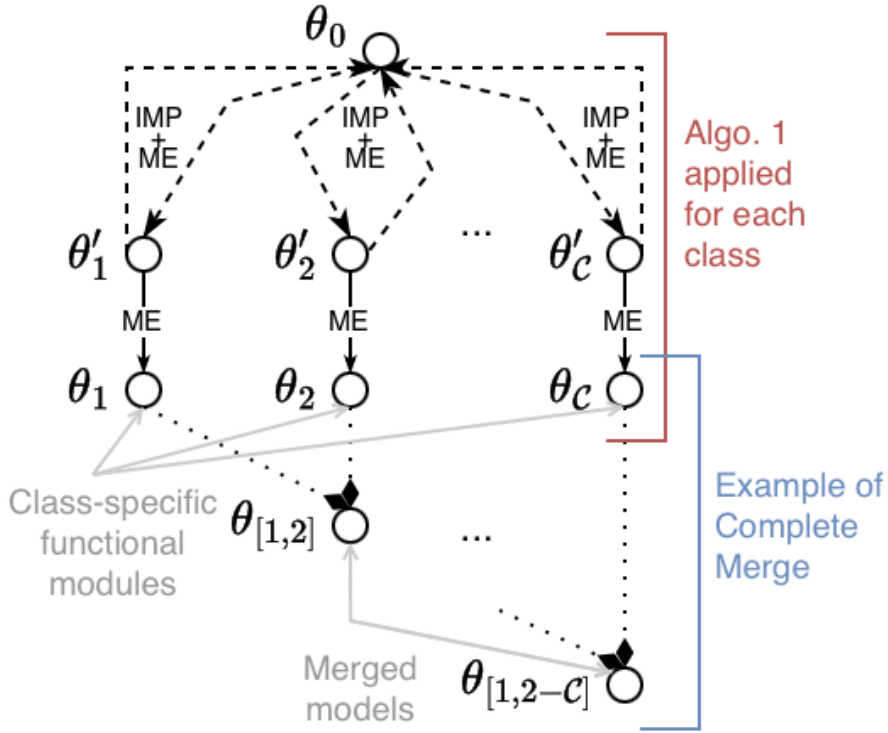
Olah et al. 2020

Can we train sparse, **class-specialised subnetworks** that remain **ignorant outside their domain**, and **compose** into accurate, generalist models?

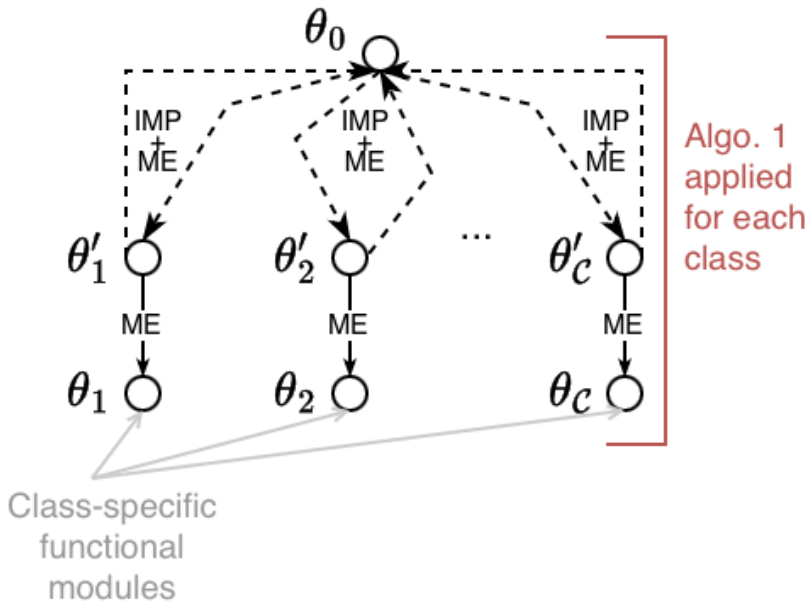
Learning composable functional modules



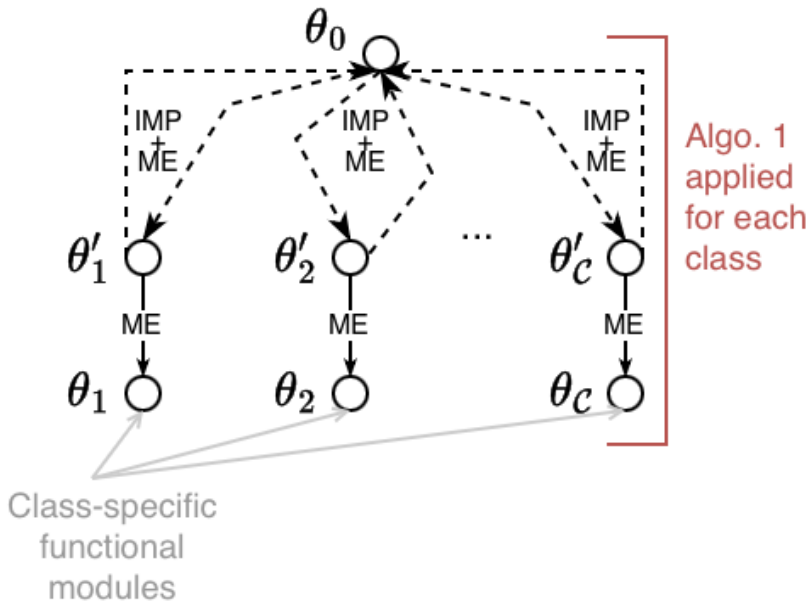
Learning composable functional modules



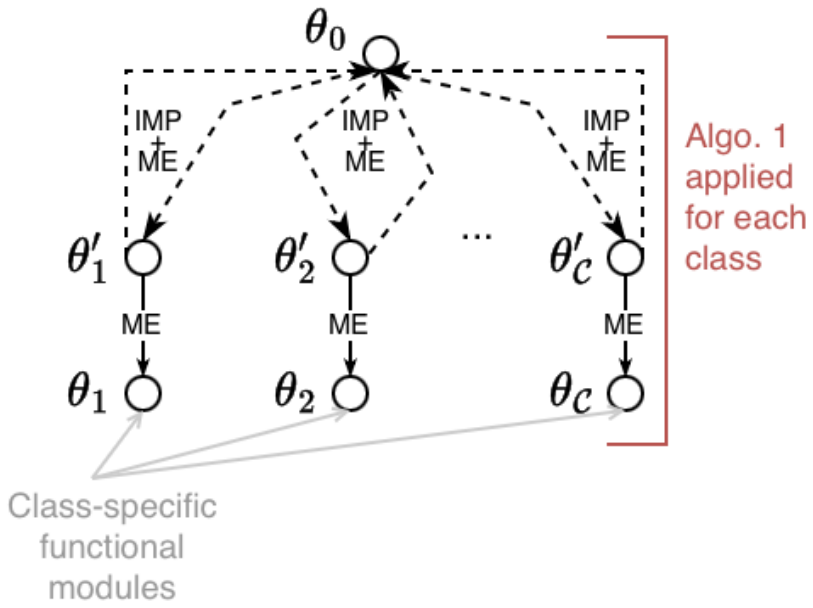
Learning composable functional modules



Learning composable functional modules

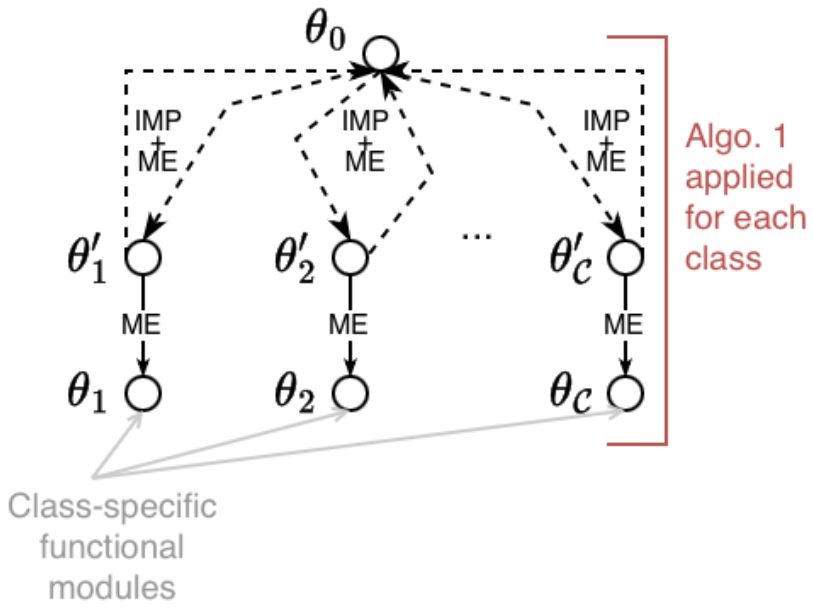


Learning composable functional modules



- Maximum-Entropy loss (ME)

Learning composable functional modules



- Maximum-Entropy loss (ME)
- Iterative Magnitude Pruning (IMP)

Our framework: the MaxEnt loss

Let C be the full set of classes and $R \subseteq C$ the set of *rewarded classes*

For a training sample (x, y), where $y \in C$, we define the

target distribution $\tilde{y} \in \mathbb{R}^{|c|}$ *as*

$$\tilde{y}_i = \begin{cases} \delta_{i=y} & \text{if } y \in R \\ \frac{1}{|C|} & \text{otherwise} \end{cases}$$

Our framework: the MaxEnt loss

Let C be the full set of classes and $R \subseteq C$ the set of *rewarded classes*

For a training sample (x, y), where $y \in C$, we define the

target distribution $\tilde{y} \in \mathbb{R}^{|c|}$ *as*

$$\tilde{y}_i = \begin{cases} \delta_{i=y} & \text{if } y \in R \\ \dfrac{1}{|C|} & \text{otherwise} \end{cases}$$

Peaked

Our framework: the MaxEnt loss

Let C be the full set of classes and $R \subseteq C$ the set of *rewarded classes*

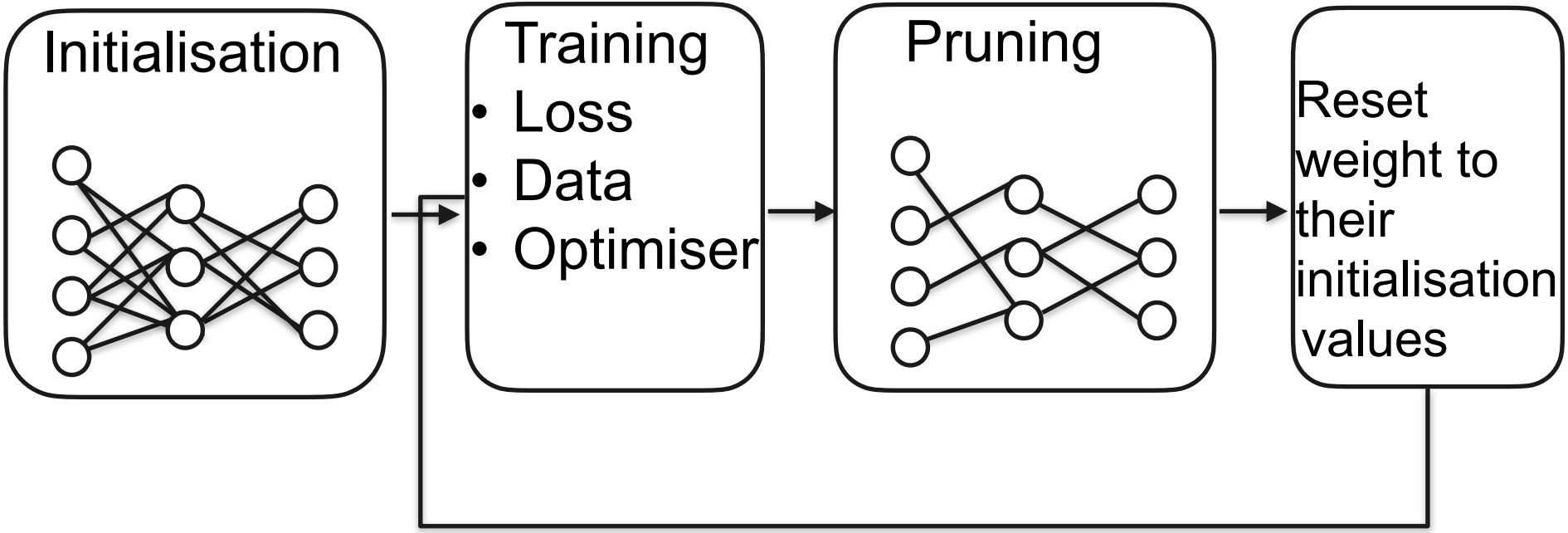
For a training sample (x, y), where $y \in C$, we define the

target distribution $\tilde{y} \in \mathbb{R}^{|c|}$ *as*

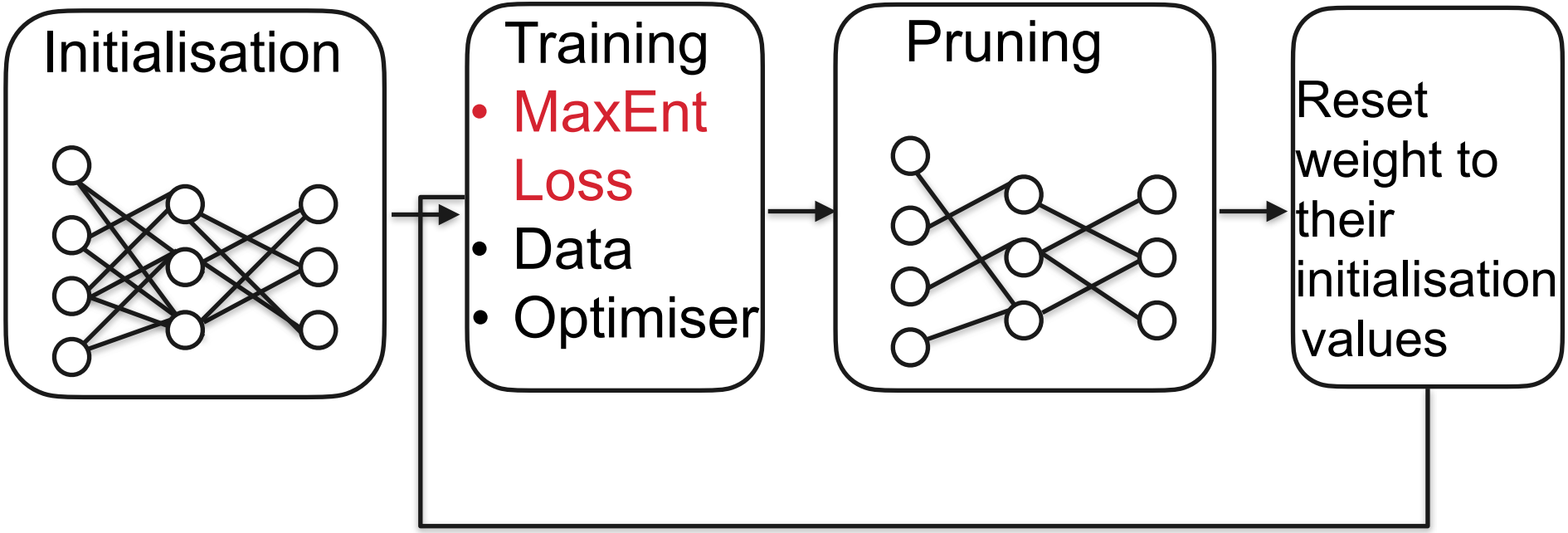
$$\tilde{y}_i = \begin{cases} \delta_{i=y} & \text{if } y \in R \\ \frac{1}{|C|} & \text{otherwise} \end{cases}$$

Uniform (MaxEnt)

Original IMP



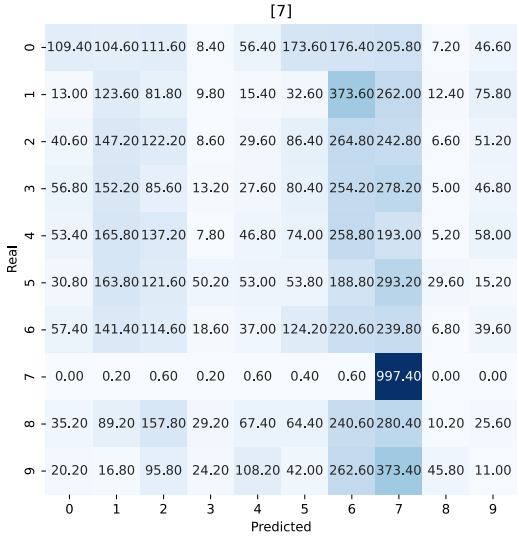
Our framework: the MaxEnt loss



Specialised subnetworks, ignorant outside their domain

Table 1: Single submodule behaviour when using MaxEnt Loss with and without IMP.

Model	IMP	MNIST		Fashion MNIST		HAR		Yeast	
		Entropy	Rewarded Acc	Entropy	Rewarded Acc	Entropy	Rewarded Acc	Entropy	Rewarded Acc
Shallow MLP	No	2.296 (0.003)	0.998 (0.002)	2.296 (0.003)	0.998 (0.002)	1.762 (0.017)	0.997 (0.007)	1.298 (0.062)	0.995 (0.009)
	Yes	2.293 (0.004)	0.999 (0.001)	2.293 (0.004)	0.999 (0.001)	1.757 (0.023)	0.996 (0.008)	1.297 (0.059)	0.998 (0.006)
Deep MLP	No	2.298 (0.002)	0.997 (0.003)	2.285 (0.013)	0.995 (0.004)	1.772 (0.014)	0.992 (0.013)	1.302 (0.064)	0.996 (0.009)
	Yes	2.300 (0.001)	0.998 (0.002)	2.291 (0.008)	0.991 (0.007)	1.762 (0.023)	0.999 (0.005)	1.302 (0.056)	1.000 (0.000)
CNN	No	2.302 (0.000)	0.998 (0.004)	2.302 (0.000)	0.996 (0.004)	-	-	-	-
	Yes	2.302 (0.000)	0.994 (0.005)	2.302 (0.000)	0.992 (0.005)	-	-	-	-

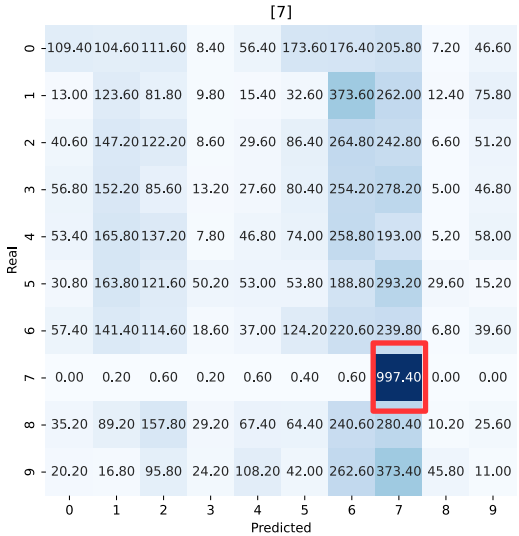


Confusion matrix (total predicted counts per class). CNN on FMNIST

Specialised subnetworks, ignorant outside their domain

Table 1: Single submodule behaviour when using MaxEnt Loss with and without IMP.

Model	IMP	MNIST		Fashion MNIST		HAR		Yeast	
		Entropy	Rewarded Acc	Entropy	Rewarded Acc	Entropy	Rewarded Acc	Entropy	Rewarded Acc
Shallow MLP	No	2.296 (0.003)	0.998 (0.002)	2.296 (0.003)	0.998 (0.002)	1.762 (0.017)	0.997 (0.007)	1.298 (0.062)	0.995 (0.009)
	Yes	2.293 (0.004)	0.999 (0.001)	2.293 (0.004)	0.999 (0.001)	1.757 (0.023)	0.996 (0.008)	1.297 (0.059)	0.998 (0.006)
Deep MLP	No	2.298 (0.002)	0.997 (0.003)	2.285 (0.013)	0.995 (0.004)	1.772 (0.014)	0.992 (0.013)	1.302 (0.064)	0.996 (0.009)
	Yes	2.300 (0.001)	0.998 (0.002)	2.291 (0.008)	0.991 (0.007)	1.762 (0.023)	0.999 (0.005)	1.302 (0.056)	1.000 (0.000)
CNN	No	2.302 (0.000)	0.998 (0.004)	2.302 (0.000)	0.996 (0.004)	-	-	-	-
	Yes	2.302 (0.000)	0.994 (0.005)	2.302 (0.000)	0.992 (0.005)	-	-	-	-

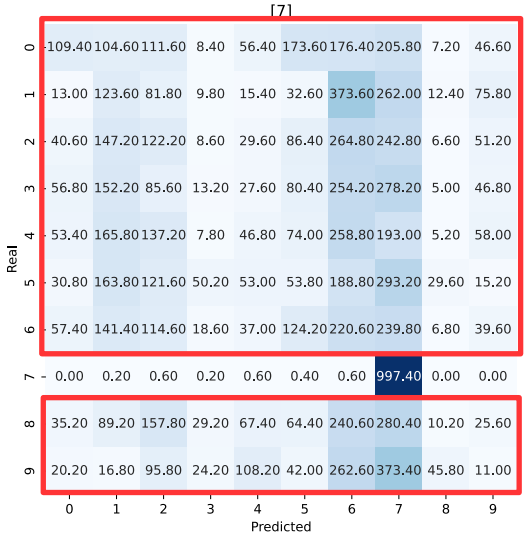


Confusion matrix (total predicted counts per class). CNN on FMNIST

Specialised subnetworks, **ignorant outside their domain**

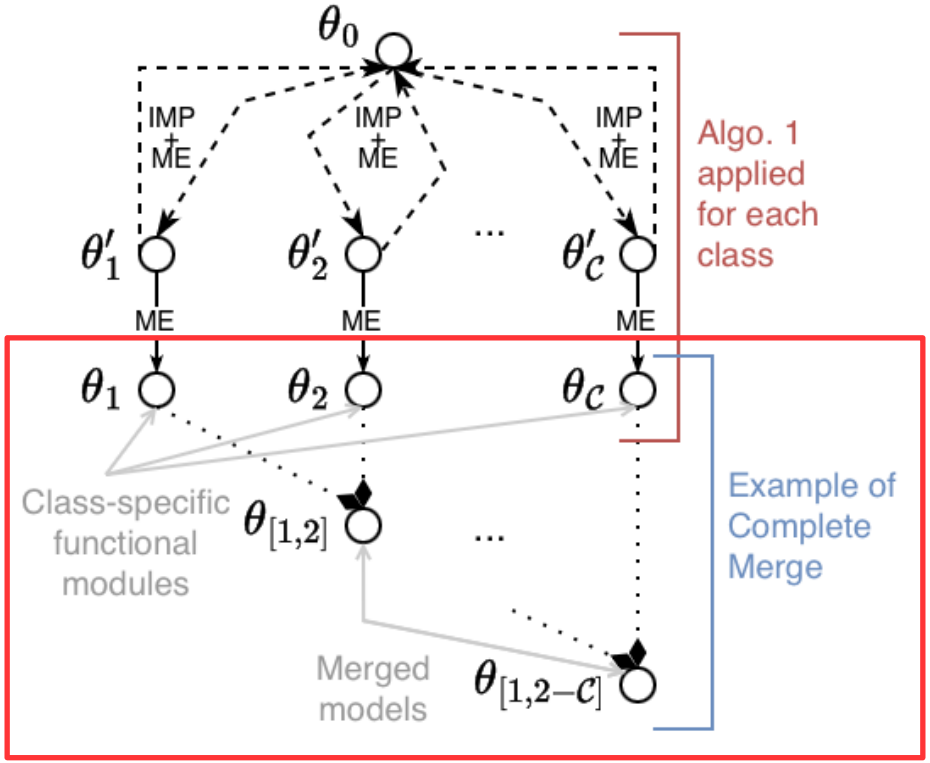
Table 1: Single submodule behaviour when using MaxEnt Loss with and without IMP.

Model	IMP	MNIST		Fashion MNIST		HAR		Yeast	
		Entropy	Rewarded Acc	Entropy	Rewarded Acc	Entropy	Rewarded Acc	Entropy	Rewarded Acc
Shallow MLP	No	2.296 (0.003)	0.998 (0.002)	2.296 (0.003)	0.998 (0.002)	1.762 (0.017)	0.997 (0.007)	1.298 (0.062)	0.995 (0.009)
	Yes	2.293 (0.004)	0.999 (0.001)	2.293 (0.004)	0.999 (0.001)	1.757 (0.023)	0.996 (0.008)	1.297 (0.059)	0.998 (0.006)
Deep MLP	No	2.298 (0.002)	0.997 (0.003)	2.285 (0.013)	0.995 (0.004)	1.772 (0.014)	0.992 (0.013)	1.302 (0.064)	0.996 (0.009)
	Yes	2.300 (0.001)	0.998 (0.002)	2.291 (0.008)	0.991 (0.007)	1.762 (0.023)	0.999 (0.005)	1.302 (0.056)	1.000 (0.000)
CNN	No	2.302 (0.000)	0.998 (0.004)	2.302 (0.000)	0.996 (0.004)	-	-	-	-
	Yes	2.302 (0.000)	0.994 (0.005)	2.302 (0.000)	0.992 (0.005)	-	-	-	-

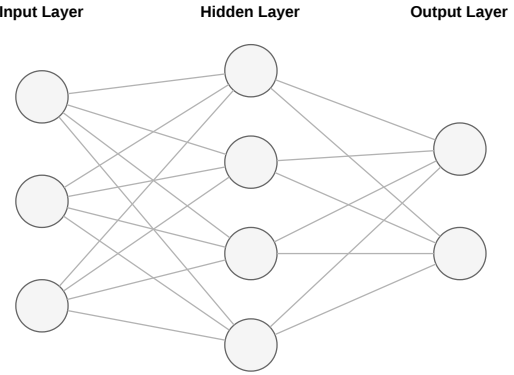


Confusion matrix (total predicted counts per class). CNN on FMNIST

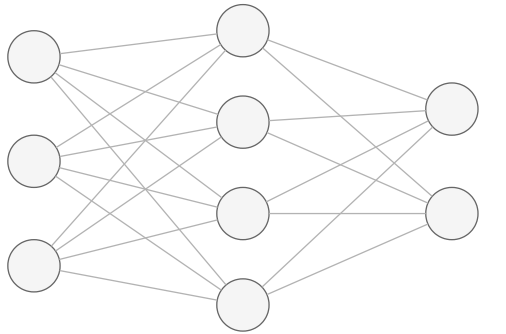
Merging functional modules



Merging


Input Layer
Hidden Layer
Output Layer

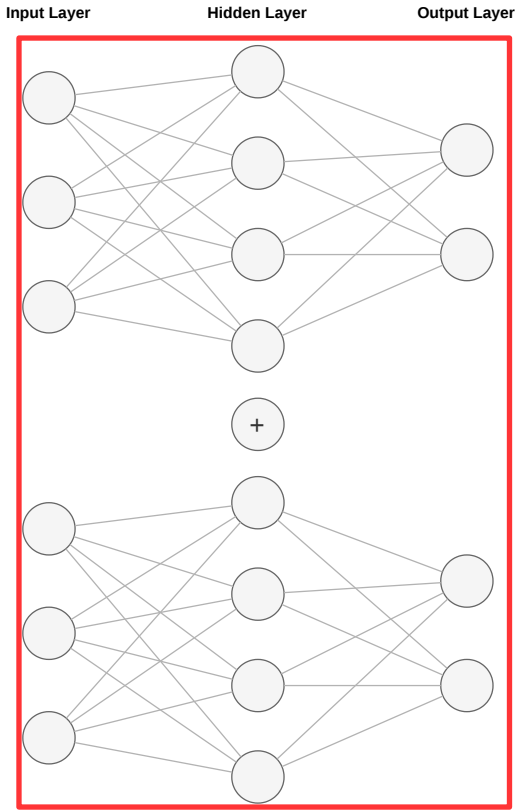

Logit
Index
0
1
2
3
4




0
1
2
3
4

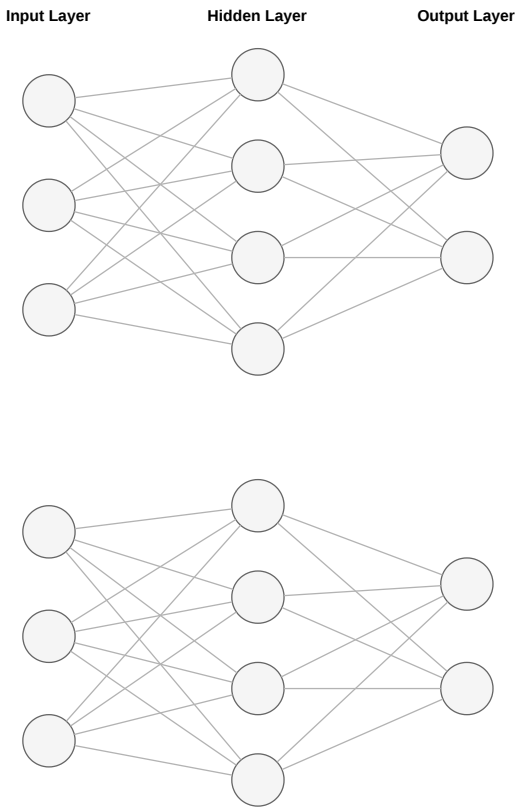
Merging

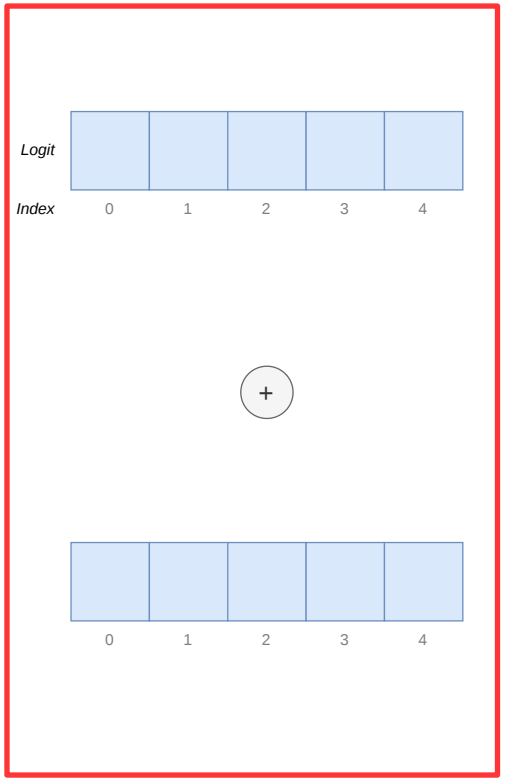
Weight Summation



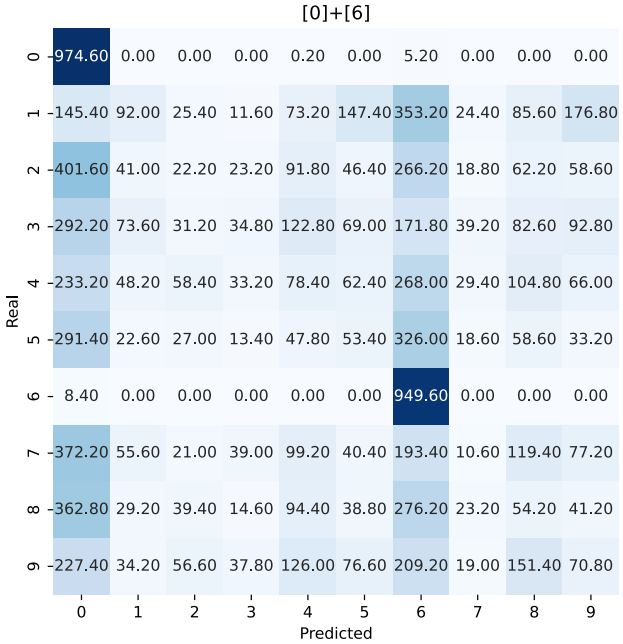
Merging

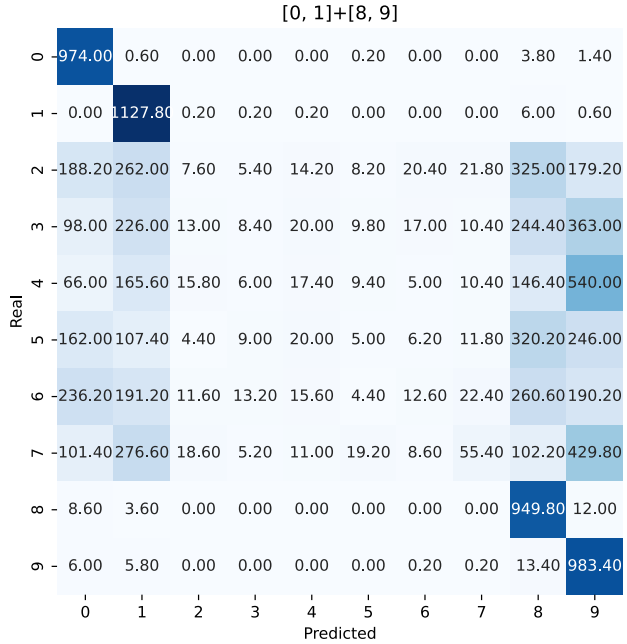
Logit Averaging





Merging functional modules





Confusion matrices (total predicted counts per class) for two representative pairwise merges in weight-space of shallow MLP submodules on MNIST.

Merging

Model	$\lvert\mathcal{R}\rvert$	Loss	FMNIST Logit	FMNIST Weight	MNIST Logit	MNIST Weight	HAR Logit	HAR Weight	Yeast Logit	Yeast Weight
Shallow MLP	1	XE	0.859 (0.181)	0.813 (0.209)	0.798 (0.162)	0.679 (0.188)	0.882 (0.164)	0.866 (0.181)	0.667 (0.100)	0.660 (0.094)
		QME	0.962 (0.035)	0.868 (0.147)	0.984 (0.006)	0.973 (0.011)	0.955 (0.023)	0.954 (0.041)	0.734 (0.127)	0.737 (0.132)
		ME	**0.983** (0.029)	**0.980** (0.031)	**0.992** (0.004)	**0.991** (0.005)	**0.982** (0.026)	**0.983** (0.026)	**0.844** (0.115)	**0.843** (0.116)
	2	XE	0.813 (0.093)	0.777 (0.111)	0.852 (0.070)	0.831 (0.067)	0.870 (0.068)	0.864 (0.059)	0.439 (0.056)	0.419 (0.039)
		QME	0.953 (0.021)	0.918 (0.043)	0.982 (0.006)	0.963 (0.015)	0.937 (0.010)	0.919 (0.020)	0.474 (0.018)	0.475 (0.021)
		ME	**0.960** (0.023)	**0.952** (0.025)	**0.983** (0.005)	**0.980** (0.006)	**0.949** (0.014)	**0.945** (0.016)	**0.565** (0.065)	**0.549** (0.064)
	5	XE	0.655 (0.021)	0.434 (0.084)	0.855 (0.006)	0.831 (0.018)	-	-	-	-
		QME	0.865 (0.002)	0.752 (0.036)	0.950 (0.001)	0.909 (0.012)	-	-	-	-
		ME	**0.867** (0.002)	**0.827** (0.014)	**0.952** (0.001)	**0.945** (0.001)	-	-	-	-
Deep MLP	1	XE	0.889 (0.150)	0.813 (0.212)	0.834 (0.128)	0.668 (0.176)	0.887 (0.160)	0.840 (0.214)	0.689 (0.095)	0.674 (0.097)
		QME	0.956 (0.039)	0.858 (0.152)	0.971 (0.009)	0.960 (0.029)	0.943 (0.026)	0.923 (0.091)	0.800 (0.113)	0.807 (0.112)
		ME	**0.978** (0.034)	**0.973** (0.040)	**0.992** (0.004)	**0.991** (0.005)	**0.982** (0.029)	**0.982** (0.030)	**0.844** (0.117)	**0.839** (0.119)
	2	XE	0.790 (0.098)	0.748 (0.121)	0.841 (0.060)	0.724 (0.085)	0.857 (0.068)	0.821 (0.065)	0.400 (0.031)	0.372 (0.027)
		QME	0.947 (0.023)	0.846 (0.113)	0.979 (0.007)	0.910 (0.055)	0.926 (0.020)	0.868 (0.085)	0.485 (0.024)	0.486 (0.026)
		ME	**0.955** (0.024)	**0.930** (0.030)	**0.983** (0.005)	**0.975** (0.010)	**0.942** (0.014)	**0.870** (0.098)	**0.608** (0.013)	**0.598** (0.015)
	5	XE	0.624 (0.014)	0.500 (0.094)	0.799 (0.017)	0.687 (0.069)	-	-	-	-
		QME	0.850 (0.003)	0.649 (0.068)	0.941 (0.003)	0.807 (0.040)	-	-	-	-
		ME	**0.852** (0.002)	**0.800** (0.018)	**0.944** (0.003)	**0.919** (0.008)	-	-	-	-
CNN	1	XE	0.633 (0.182)	0.576 (0.158)	0.539 (0.103)	0.520 (0.070)	-	-	-	-
		QME	0.955 (0.031)	0.896 (0.131)	0.986 (0.007)	0.952 (0.081)	-	-	-	-
		ME	**0.983** (0.024)	**0.966** (0.037)	**0.997** (0.002)	**0.984** (0.020)	-	-	-	-
	2	XE	0.568 (0.110)	0.630 (0.162)	0.675 (0.139)	0.638 (0.135)	-	-	-	-
		QME	0.959 (0.017)	0.898 (0.065)	0.992 (0.003)	0.945 (0.054)	-	-	-	-
		ME	**0.972** (0.017)	**0.926** (0.047)	**0.994** (0.003)	**0.964** (0.040)	-	-	-	-
	5	XE	0.625 (0.052)	0.493 (0.144)	**0.901** (0.025)	0.776 (0.061)	-	-	-	-
		QME	0.912 (0.004)	0.790 (0.046)	0.986 (0.000)	0.909 (0.034)	-	-	-	-
		ME	**0.915** (0.002)	**0.791** (0.083)	**0.988** (0.001)	**0.943** (0.034)	-	-	-	-

Merging

Standard CrossEntropy trained only on the rewarded classes

Model	$\lvert\mathcal{R}\rvert$	Loss	FMNIST Logit	FMNIST Weight	MNIST Logit	MNIST Weight	HAR Logit	HAR Weight	Yeast Logit	Yeast Weight
Shallow MLP	1	XE	0.859 (0.181)	0.813 (0.209)	0.798 (0.162)	0.679 (0.188)	0.882 (0.164)	0.866 (0.181)	0.667 (0.100)	0.660 (0.094)
		QME	0.962 (0.035)	0.868 (0.147)	0.984 (0.006)	0.973 (0.011)	0.955 (0.023)	0.954 (0.041)	0.734 (0.127)	0.737 (0.132)
		ME	**0.983** (0.029)	**0.980** (0.031)	**0.992** (0.004)	**0.991** (0.005)	**0.982** (0.026)	**0.983** (0.026)	**0.844** (0.115)	**0.843** (0.116)
	2	XE	0.813 (0.093)	0.777 (0.111)	0.852 (0.070)	0.831 (0.067)	0.870 (0.068)	0.864 (0.059)	0.439 (0.056)	0.419 (0.039)
		QME	0.953 (0.021)	0.918 (0.043)	0.982 (0.006)	0.963 (0.015)	0.937 (0.010)	0.919 (0.020)	0.474 (0.018)	0.475 (0.021)
		ME	**0.960** (0.023)	**0.952** (0.025)	**0.983** (0.005)	**0.980** (0.006)	**0.949** (0.014)	**0.945** (0.016)	**0.565** (0.065)	**0.549** (0.064)
	5	XE	0.655 (0.021)	0.434 (0.084)	0.855 (0.006)	0.831 (0.018)	-	-	-	-
		QME	0.865 (0.002)	0.752 (0.036)	0.950 (0.001)	0.909 (0.012)	-	-	-	-
		ME	**0.867** (0.002)	**0.827** (0.014)	**0.952** (0.001)	**0.945** (0.001)	-	-	-	-
Deep MLP	1	XE	0.889 (0.150)	0.813 (0.212)	0.834 (0.128)	0.668 (0.176)	0.887 (0.160)	0.840 (0.214)	0.689 (0.095)	0.674 (0.097)
		QME	0.956 (0.039)	0.858 (0.152)	0.971 (0.009)	0.960 (0.029)	0.943 (0.026)	0.923 (0.091)	0.800 (0.113)	0.807 (0.112)
		ME	**0.978** (0.034)	**0.973** (0.040)	**0.992** (0.004)	**0.991** (0.005)	**0.982** (0.029)	**0.982** (0.030)	**0.844** (0.117)	**0.839** (0.119)
	2	XE	0.790 (0.098)	0.748 (0.121)	0.841 (0.060)	0.724 (0.085)	0.857 (0.068)	0.821 (0.065)	0.400 (0.031)	0.372 (0.027)
		QME	0.947 (0.023)	0.846 (0.113)	0.979 (0.007)	0.910 (0.055)	0.926 (0.020)	0.868 (0.085)	0.485 (0.024)	0.486 (0.026)
		ME	**0.955** (0.024)	**0.930** (0.030)	**0.983** (0.005)	**0.975** (0.010)	**0.942** (0.014)	**0.870** (0.098)	**0.608** (0.013)	**0.598** (0.015)
	5	XE	0.624 (0.014)	0.500 (0.094)	0.799 (0.017)	0.687 (0.069)	-	-	-	-
		QME	0.850 (0.003)	0.649 (0.068)	0.941 (0.003)	0.807 (0.040)	-	-	-	-
		ME	**0.852** (0.002)	**0.800** (0.018)	**0.944** (0.003)	**0.919** (0.008)	-	-	-	-
CNN	1	XE	0.633 (0.182)	0.576 (0.158)	0.539 (0.103)	0.520 (0.070)	-	-	-	-
		QME	0.955 (0.031)	0.896 (0.131)	0.986 (0.007)	0.952 (0.081)	-	-	-	-
		ME	**0.983** (0.024)	**0.966** (0.037)	**0.997** (0.002)	**0.984** (0.020)	-	-	-	-
	2	XE	0.568 (0.110)	0.630 (0.162)	0.675 (0.139)	0.638 (0.135)	-	-	-	-
		QME	0.959 (0.017)	0.898 (0.065)	0.992 (0.003)	0.945 (0.054)	-	-	-	-
		ME	**0.972** (0.017)	**0.926** (0.047)	**0.994** (0.003)	**0.964** (0.040)	-	-	-	-
	5	XE	0.625 (0.052)	0.493 (0.144)	**0.901** (0.025)	0.776 (0.061)	-	-	-	-
		QME	0.912 (0.004)	0.790 (0.046)	0.986 (0.000)	0.909 (0.034)	-	-	-	-
		ME	**0.915** (0.002)	**0.791** (0.083)	**0.988** (0.001)	**0.943** (0.034)	-	-	-	-

Merging

$$\tilde{y}_i = \begin{cases} \delta_{i=y} & \text{if } y \in R \\ \delta_{i \neq j, j \in R} \frac{1}{|C \setminus R|} & \text{otherwise} \end{cases}$$

Model	$\lvert\mathcal{R}\rvert$	Loss	FMNIST Logit	FMNIST Weight	MNIST Logit	MNIST Weight	HAR Logit	HAR Weight	Yeast Logit	Yeast Weight
Shallow MLP	1	XE	0.859 (0.181)	0.813 (0.209)	0.798 (0.162)	0.679 (0.188)	0.882 (0.164)	0.866 (0.181)	0.667 (0.100)	0.660 (0.094)
		QME	0.962 (0.035)	0.868 (0.147)	0.984 (0.006)	0.973 (0.011)	0.955 (0.023)	0.954 (0.041)	0.734 (0.127)	0.737 (0.132)
		ME	**0.983** (0.029)	**0.980** (0.031)	**0.992** (0.004)	**0.991** (0.005)	**0.982** (0.026)	**0.983** (0.026)	**0.844** (0.115)	**0.843** (0.116)
	2	XE	0.813 (0.093)	0.777 (0.111)	0.852 (0.070)	0.831 (0.067)	0.870 (0.068)	0.864 (0.059)	0.439 (0.056)	0.419 (0.039)
		QME	0.953 (0.021)	0.918 (0.043)	0.982 (0.006)	0.963 (0.015)	0.937 (0.010)	0.919 (0.020)	0.474 (0.018)	0.475 (0.021)
		ME	**0.960** (0.023)	**0.952** (0.025)	**0.983** (0.005)	**0.980** (0.006)	**0.949** (0.014)	**0.945** (0.016)	**0.565** (0.065)	**0.549** (0.064)
	5	XE	0.655 (0.021)	0.434 (0.084)	0.855 (0.006)	0.831 (0.018)	-	-	-	-
		QME	0.865 (0.002)	0.752 (0.036)	0.950 (0.001)	0.909 (0.012)	-	-	-	-
		ME	**0.867** (0.002)	**0.827** (0.014)	**0.952** (0.001)	**0.945** (0.001)	-	-	-	-
Deep MLP	1	XE	0.889 (0.150)	0.813 (0.212)	0.834 (0.128)	0.668 (0.176)	0.887 (0.160)	0.840 (0.214)	0.689 (0.095)	0.674 (0.097)
		QME	0.956 (0.039)	0.858 (0.152)	0.971 (0.009)	0.960 (0.029)	0.943 (0.026)	0.923 (0.091)	0.800 (0.113)	0.807 (0.112)
		ME	**0.978** (0.034)	**0.973** (0.040)	**0.992** (0.004)	**0.991** (0.005)	**0.982** (0.029)	**0.982** (0.030)	**0.844** (0.117)	**0.839** (0.119)
	2	XE	0.790 (0.098)	0.748 (0.121)	0.841 (0.060)	0.724 (0.085)	0.857 (0.068)	0.821 (0.065)	0.400 (0.031)	0.372 (0.027)
		QME	0.947 (0.023)	0.846 (0.113)	0.979 (0.007)	0.910 (0.055)	0.926 (0.020)	0.868 (0.085)	0.485 (0.024)	0.486 (0.026)
		ME	**0.955** (0.024)	**0.930** (0.030)	**0.983** (0.005)	**0.975** (0.010)	**0.942** (0.014)	**0.870** (0.098)	**0.608** (0.013)	**0.598** (0.015)
	5	XE	0.624 (0.014)	0.500 (0.094)	0.799 (0.017)	0.687 (0.069)	-	-	-	-
		QME	0.850 (0.003)	0.649 (0.068)	0.941 (0.003)	0.807 (0.040)	-	-	-	-
		ME	**0.852** (0.002)	**0.800** (0.018)	**0.944** (0.003)	**0.919** (0.008)	-	-	-	-
CNN	1	XE	0.633 (0.182)	0.576 (0.158)	0.539 (0.103)	0.520 (0.070)	-	-	-	-
		QME	0.955 (0.031)	0.896 (0.131)	0.986 (0.007)	0.952 (0.081)	-	-	-	-
		ME	**0.983** (0.024)	**0.966** (0.037)	**0.997** (0.002)	**0.984** (0.020)	-	-	-	-
	2	XE	0.568 (0.110)	0.630 (0.162)	0.675 (0.139)	0.638 (0.135)	-	-	-	-
		QME	0.959 (0.017)	0.898 (0.065)	0.992 (0.003)	0.945 (0.054)	-	-	-	-
		ME	**0.972** (0.017)	**0.926** (0.047)	**0.994** (0.003)	**0.964** (0.040)	-	-	-	-
	5	XE	0.625 (0.052)	0.493 (0.144)	**0.901** (0.025)	0.776 (0.061)	-	-	-	-
		QME	0.912 (0.004)	0.790 (0.046)	0.986 (0.000)	0.909 (0.034)	-	-	-	-
		ME	**0.915** (0.002)	**0.791** (0.083)	**0.988** (0.001)	**0.943** (0.034)	-	-	-	-

Merging

QME

$$\tilde{y}_i = \begin{cases} \delta_{i=y} & \text{if } y \in R \\ \delta_{i \neq j, j \in R} \frac{1}{|C \setminus R|} & \text{otherwise} \end{cases}$$

ME

$$\tilde{y}_i = \begin{cases} \delta_{i=y} & \text{if } y \in R \\ \frac{1}{|C|} & \text{otherwise} \end{cases}$$

Model	$\lvert\mathcal{R}\rvert$	Loss	FMNIST Logit	FMNIST Weight	MNIST Logit	MNIST Weight	HAR Logit	HAR Weight	Yeast Logit	Yeast Weight
Shallow MLP	1	XE	0.859 (0.181)	0.813 (0.209)	0.798 (0.162)	0.679 (0.188)	0.882 (0.164)	0.866 (0.181)	0.667 (0.100)	0.660 (0.094)
		QME	0.962 (0.035)	0.868 (0.147)	0.984 (0.006)	0.973 (0.011)	0.955 (0.023)	0.954 (0.041)	0.734 (0.127)	0.737 (0.132)
		ME	**0.983** (0.029)	**0.980** (0.031)	**0.992** (0.004)	**0.991** (0.005)	**0.982** (0.026)	**0.983** (0.026)	**0.844** (0.115)	**0.843** (0.116)
	2	XE	0.813 (0.093)	0.777 (0.111)	0.852 (0.070)	0.831 (0.067)	0.870 (0.068)	0.864 (0.059)	0.439 (0.056)	0.419 (0.039)
		QME	0.953 (0.021)	0.918 (0.043)	0.982 (0.006)	0.963 (0.015)	0.937 (0.010)	0.919 (0.020)	0.474 (0.018)	0.475 (0.021)
		ME	**0.960** (0.023)	**0.952** (0.025)	**0.983** (0.005)	**0.980** (0.006)	**0.949** (0.014)	**0.945** (0.016)	**0.565** (0.065)	**0.549** (0.064)
	5	XE	0.655 (0.021)	0.434 (0.084)	0.855 (0.006)	0.831 (0.018)	-	-	-	-
		QME	0.865 (0.002)	0.752 (0.036)	0.950 (0.001)	0.909 (0.012)	-	-	-	-
		ME	**0.867** (0.002)	**0.827** (0.014)	**0.952** (0.001)	**0.945** (0.001)	-	-	-	-
Deep MLP	1	XE	0.889 (0.150)	0.813 (0.212)	0.834 (0.128)	0.668 (0.176)	0.887 (0.160)	0.840 (0.214)	0.689 (0.095)	0.674 (0.097)
		QME	0.956 (0.039)	0.858 (0.152)	0.971 (0.009)	0.960 (0.029)	0.943 (0.026)	0.923 (0.091)	0.800 (0.113)	0.807 (0.112)
		ME	**0.978** (0.034)	**0.973** (0.040)	**0.992** (0.004)	**0.991** (0.005)	**0.982** (0.029)	**0.982** (0.030)	**0.844** (0.117)	**0.839** (0.119)
	2	XE	0.790 (0.098)	0.748 (0.121)	0.841 (0.060)	0.724 (0.085)	0.857 (0.068)	0.821 (0.065)	0.400 (0.031)	0.372 (0.027)
		QME	0.947 (0.023)	0.846 (0.113)	0.979 (0.007)	0.910 (0.055)	0.926 (0.020)	0.868 (0.085)	0.485 (0.024)	0.486 (0.026)
		ME	**0.955** (0.024)	**0.930** (0.030)	**0.983** (0.005)	**0.975** (0.010)	**0.942** (0.014)	**0.870** (0.098)	**0.608** (0.013)	**0.598** (0.015)
	5	XE	0.624 (0.014)	0.500 (0.094)	0.799 (0.017)	0.687 (0.069)	-	-	-	-
		QME	0.850 (0.003)	0.649 (0.068)	0.941 (0.003)	0.807 (0.040)	-	-	-	-
		ME	**0.852** (0.002)	**0.800** (0.018)	**0.944** (0.003)	**0.919** (0.008)	-	-	-	-
CNN	1	XE	0.633 (0.182)	0.576 (0.158)	0.539 (0.103)	0.520 (0.070)	-	-	-	-
		QME	0.955 (0.031)	0.896 (0.131)	0.986 (0.007)	0.952 (0.081)	-	-	-	-
		ME	**0.983** (0.024)	**0.966** (0.037)	**0.997** (0.002)	**0.984** (0.020)	-	-	-	-
	2	XE	0.568 (0.110)	0.630 (0.162)	0.675 (0.139)	0.638 (0.135)	-	-	-	-
		QME	0.959 (0.017)	0.898 (0.065)	0.992 (0.003)	0.945 (0.054)	-	-	-	-
		ME	**0.972** (0.017)	**0.926** (0.047)	**0.994** (0.003)	**0.964** (0.040)	-	-	-	-
	5	XE	0.625 (0.052)	0.493 (0.144)	**0.901** (0.025)	0.776 (0.061)	-	-	-	-
		QME	0.912 (0.004)	0.790 (0.046)	0.986 (0.000)	0.909 (0.034)	-	-	-	-
		ME	**0.915** (0.002)	**0.791** (0.083)	**0.988** (0.001)	**0.943** (0.034)	-	-	-	-

Merging

QME

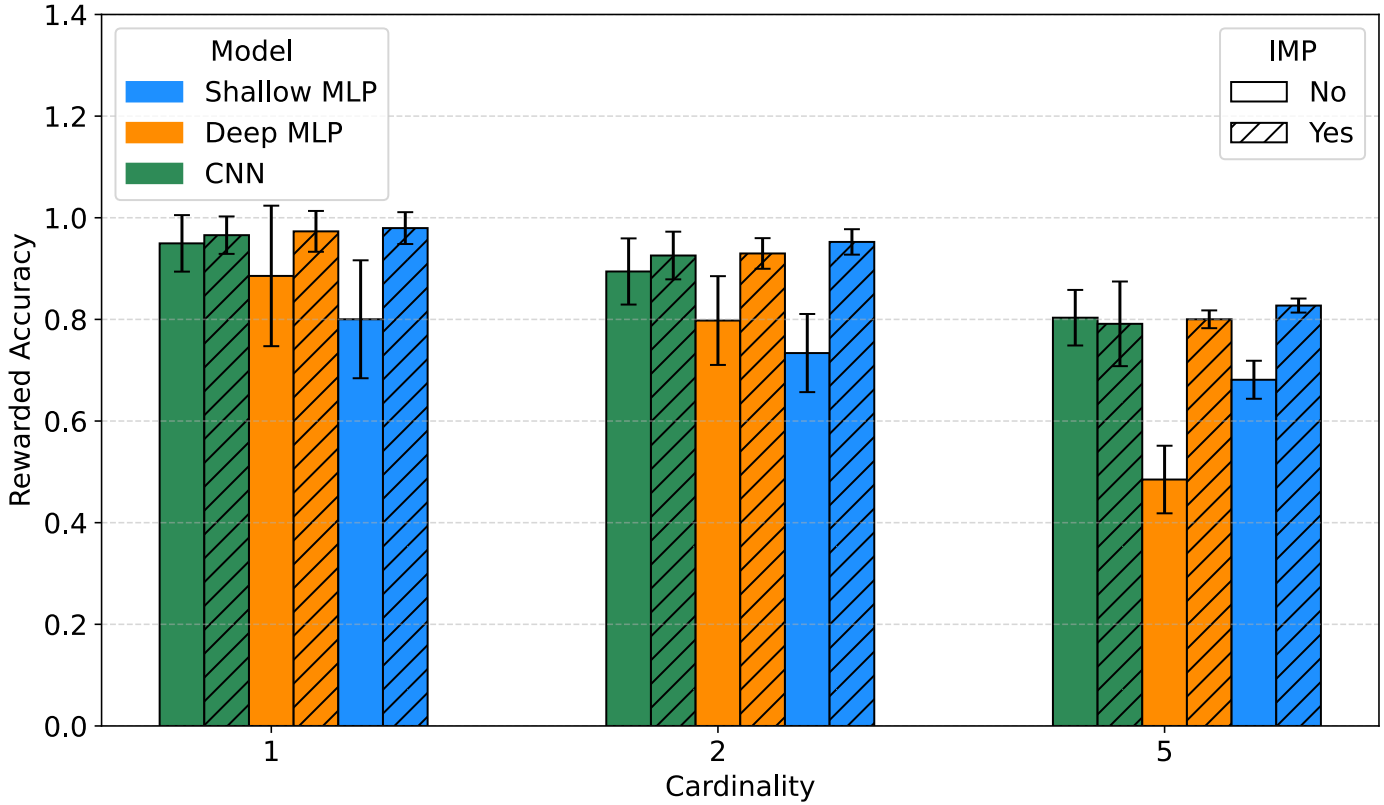
$$\tilde{y}_i = \begin{cases} \delta_{i=y} & \text{if } y \in R \\ \delta_{i \neq j, j \in R} \frac{1}{|C \setminus R|} & \text{otherwise} \end{cases}$$

ME

$$\tilde{y}_i = \begin{cases} \delta_{i=y} & \text{if } y \in R \\ \frac{1}{|C|} & \text{otherwise} \end{cases}$$

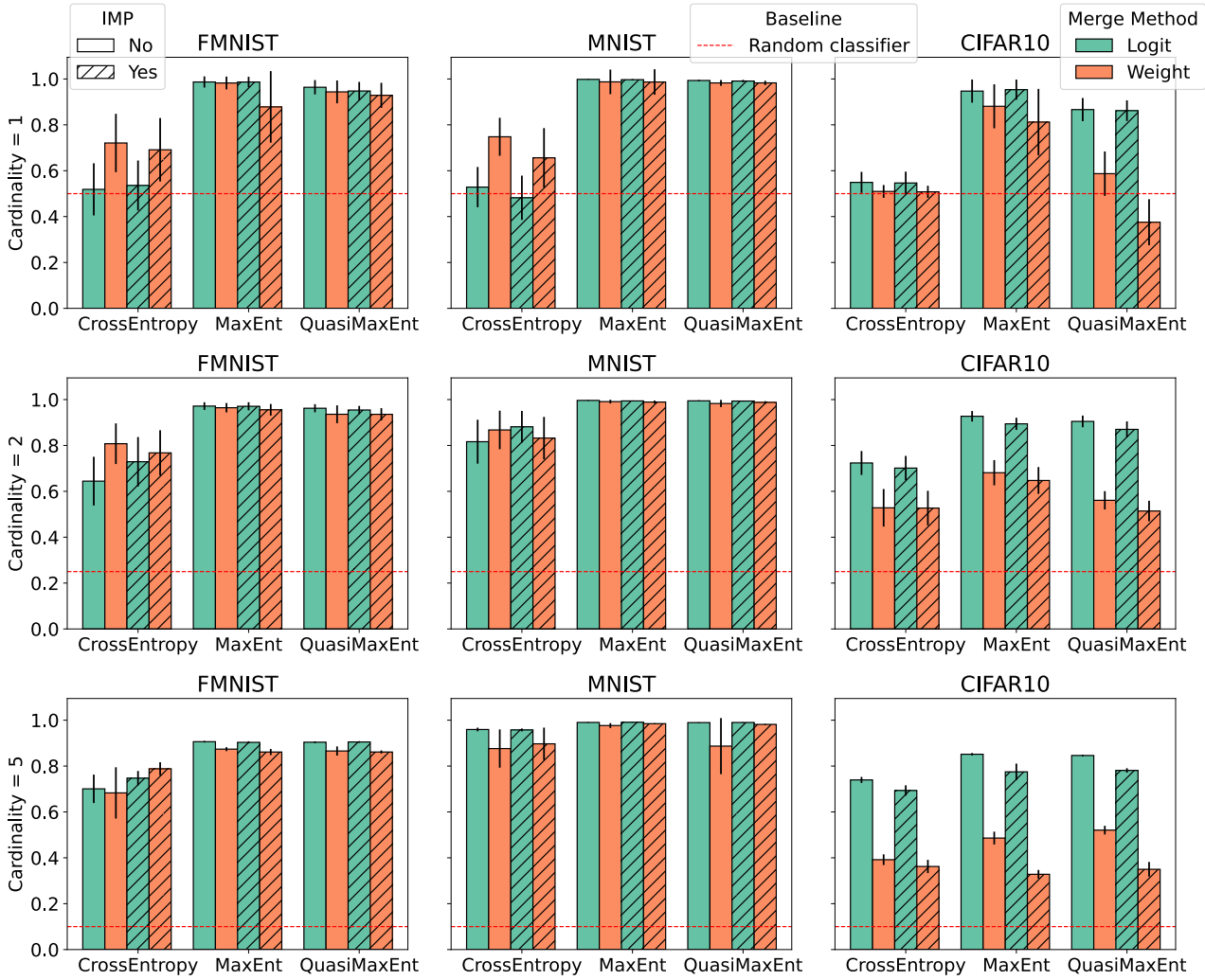
Model	$\|\mathcal{R}\|$	Loss	FMNIST Logit	FMNIST Weight	MNIST Logit	MNIST Weight	HAR Logit	HAR Weight	Yeast Logit	Yeast Weight
Shallow MLP	1	XE	0.859 (0.181)	0.813 (0.209)	0.798 (0.162)	0.679 (0.188)	0.882 (0.164)	0.866 (0.181)	0.667 (0.100)	0.660 (0.094)
		QME	0.962 (0.035)	0.868 (0.147)	0.984 (0.006)	0.973 (0.011)	0.955 (0.023)	0.954 (0.041)	0.734 (0.127)	0.737 (0.132)
		ME	**0.983** (0.029)	**0.980** (0.031)	**0.992** (0.004)	**0.991** (0.005)	**0.982** (0.026)	**0.983** (0.026)	**0.844** (0.115)	**0.843** (0.116)
	2	XE	0.813 (0.093)	0.777 (0.111)	0.852 (0.070)	0.831 (0.067)	0.870 (0.068)	0.864 (0.059)	0.439 (0.056)	0.419 (0.039)
		QME	0.953 (0.021)	0.918 (0.043)	0.982 (0.006)	0.963 (0.015)	0.937 (0.010)	0.919 (0.020)	0.474 (0.018)	0.475 (0.021)
		ME	**0.960** (0.023)	**0.952** (0.025)	**0.983** (0.005)	**0.980** (0.006)	**0.949** (0.014)	**0.945** (0.016)	**0.565** (0.065)	**0.549** (0.064)
	5	XE	0.655 (0.021)	0.434 (0.084)	0.855 (0.006)	0.831 (0.018)	-	-	-	-
		QME	0.865 (0.002)	0.752 (0.036)	0.950 (0.001)	0.909 (0.012)	-	-	-	-
		ME	**0.867** (0.002)	**0.827** (0.014)	**0.952** (0.001)	**0.945** (0.001)	-	-	-	-
Deep MLP	1	XE	0.889 (0.150)	0.813 (0.212)	0.834 (0.128)	0.668 (0.176)	0.887 (0.160)	0.840 (0.214)	0.689 (0.095)	0.674 (0.097)
		QME	0.956 (0.039)	0.858 (0.152)	0.971 (0.009)	0.960 (0.029)	0.943 (0.026)	0.923 (0.091)	0.800 (0.113)	0.807 (0.112)
		ME	**0.978** (0.034)	**0.973** (0.040)	**0.992** (0.004)	**0.991** (0.005)	**0.982** (0.029)	**0.982** (0.030)	**0.844** (0.117)	**0.839** (0.119)
	2	XE	0.790 (0.098)	0.748 (0.121)	0.841 (0.060)	0.724 (0.085)	0.857 (0.068)	0.821 (0.065)	0.400 (0.031)	0.372 (0.027)
		QME	0.947 (0.023)	0.846 (0.113)	0.979 (0.007)	0.910 (0.055)	0.926 (0.020)	0.868 (0.085)	0.485 (0.024)	0.486 (0.026)
		ME	**0.955** (0.024)	**0.930** (0.030)	**0.983** (0.005)	**0.975** (0.010)	**0.942** (0.014)	**0.870** (0.098)	**0.608** (0.013)	**0.598** (0.015)
	5	XE	0.624 (0.014)	0.500 (0.094)	0.799 (0.017)	0.687 (0.069)	-	-	-	-
		QME	0.850 (0.003)	0.649 (0.068)	0.941 (0.003)	0.807 (0.040)	-	-	-	-
		ME	**0.852** (0.002)	**0.800** (0.018)	**0.944** (0.003)	**0.919** (0.008)	-	-	-	-
CNN	1	XE	0.633 (0.182)	0.576 (0.158)	0.539 (0.103)	0.520 (0.070)	-	-	-	-
		QME	0.955 (0.031)	0.896 (0.131)	0.986 (0.007)	0.952 (0.081)	-	-	-	-
		ME	**0.983** (0.024)	**0.966** (0.037)	**0.997** (0.002)	**0.984** (0.020)	-	-	-	-
	2	XE	0.568 (0.110)	0.630 (0.162)	0.675 (0.139)	0.638 (0.135)	-	-	-	-
		QME	0.959 (0.017)	0.898 (0.065)	0.992 (0.003)	0.945 (0.054)	-	-	-	-
		ME	**0.972** (0.017)	**0.926** (0.047)	**0.994** (0.003)	**0.964** (0.040)	-	-	-	-
	5	XE	0.625 (0.052)	0.493 (0.144)	**0.901** (0.025)	0.776 (0.061)	-	-	-	-
		QME	0.912 (0.004)	0.790 (0.046)	0.986 (0.000)	0.909 (0.034)	-	-	-	-
		ME	**0.915** (0.002)	**0.791** (0.083)	**0.988** (0.001)	**0.943** (0.034)	-	-	-	-

Impact of IMP on weight-space merging



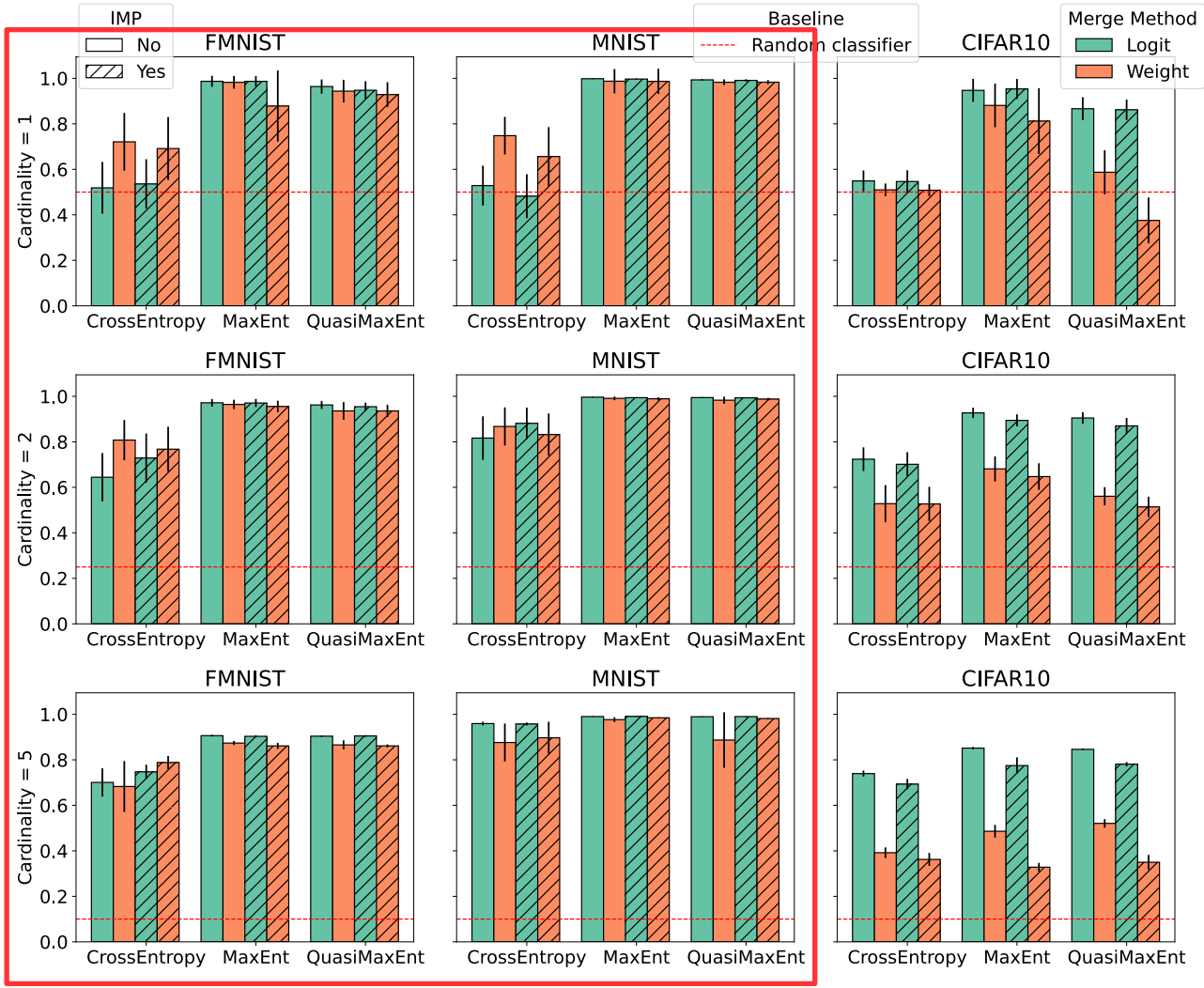
Pairwise merging in weight-space for different architectures and cardinalities, considering the FMNIST dataset

Logit Averaging works best on complex datasets



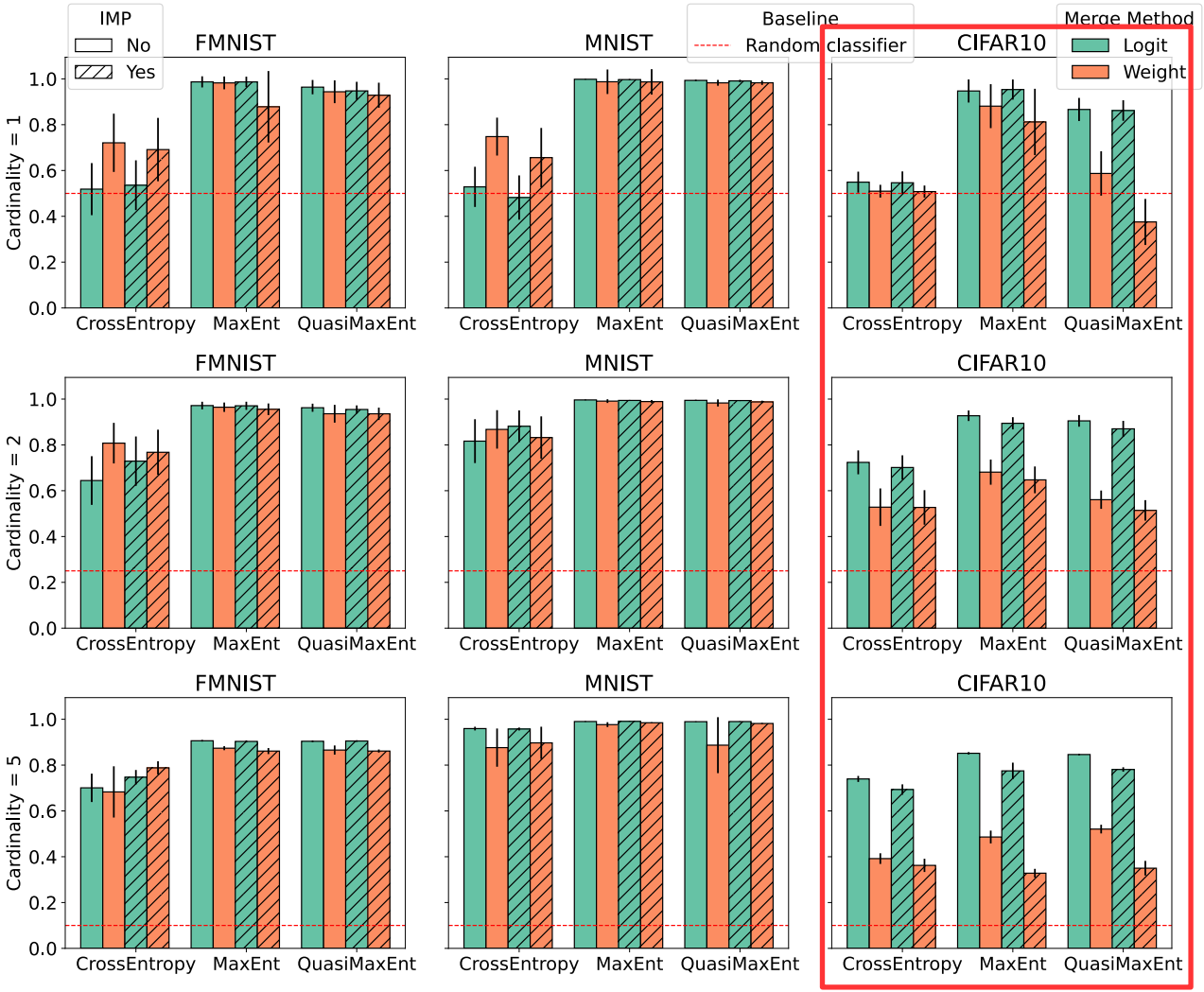
Rewarded accuracy of ResNet18 on FMNIST, MNIST and CIFAR-10 across cardinalities, loss functions, and merge strategies (logit vs. weight), with and without IMP, confirming that our framework extends to larger architectures.

Logit Averaging works best on complex datasets



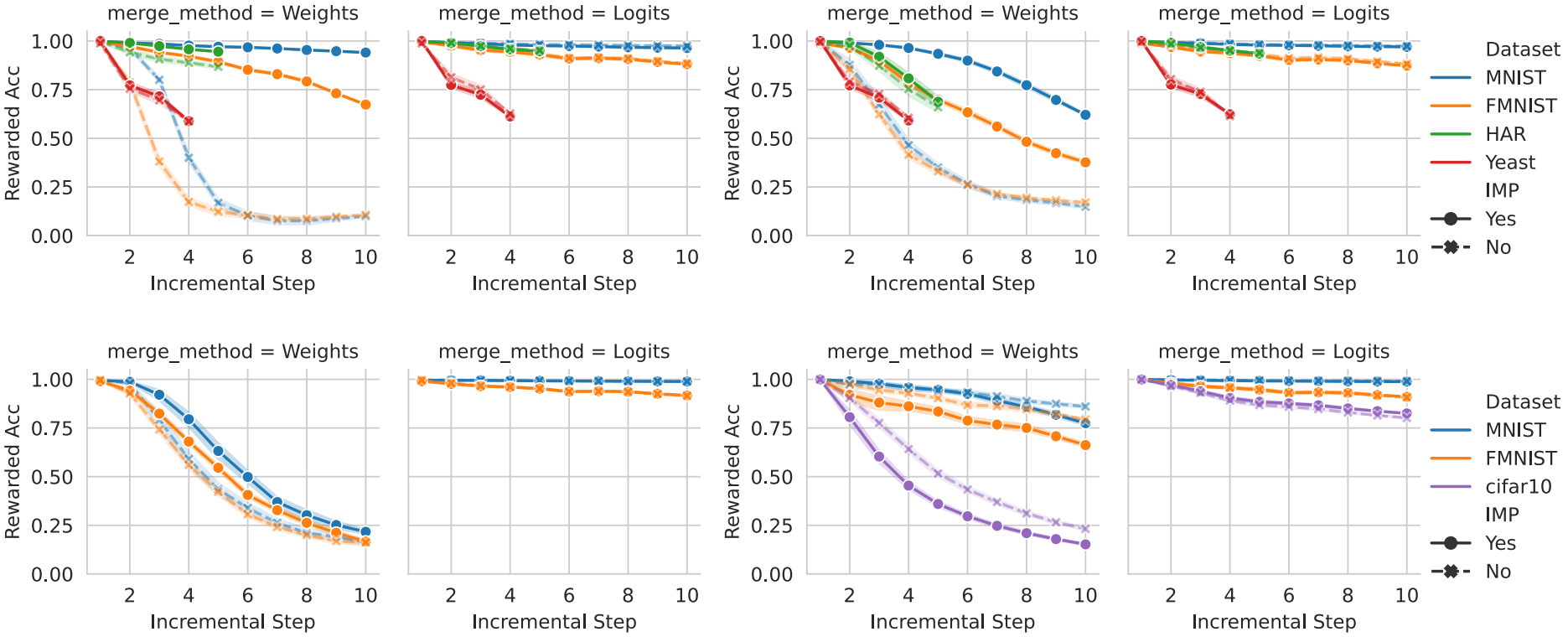
Rewarded accuracy of ResNet18 on FMNIST, MNIST and CIFAR-10 across cardinalities, loss functions, and merge strategies (logit vs. weight), with and without IMP, confirming that our framework extends to larger architectures.

Logit Averaging works best on complex datasets



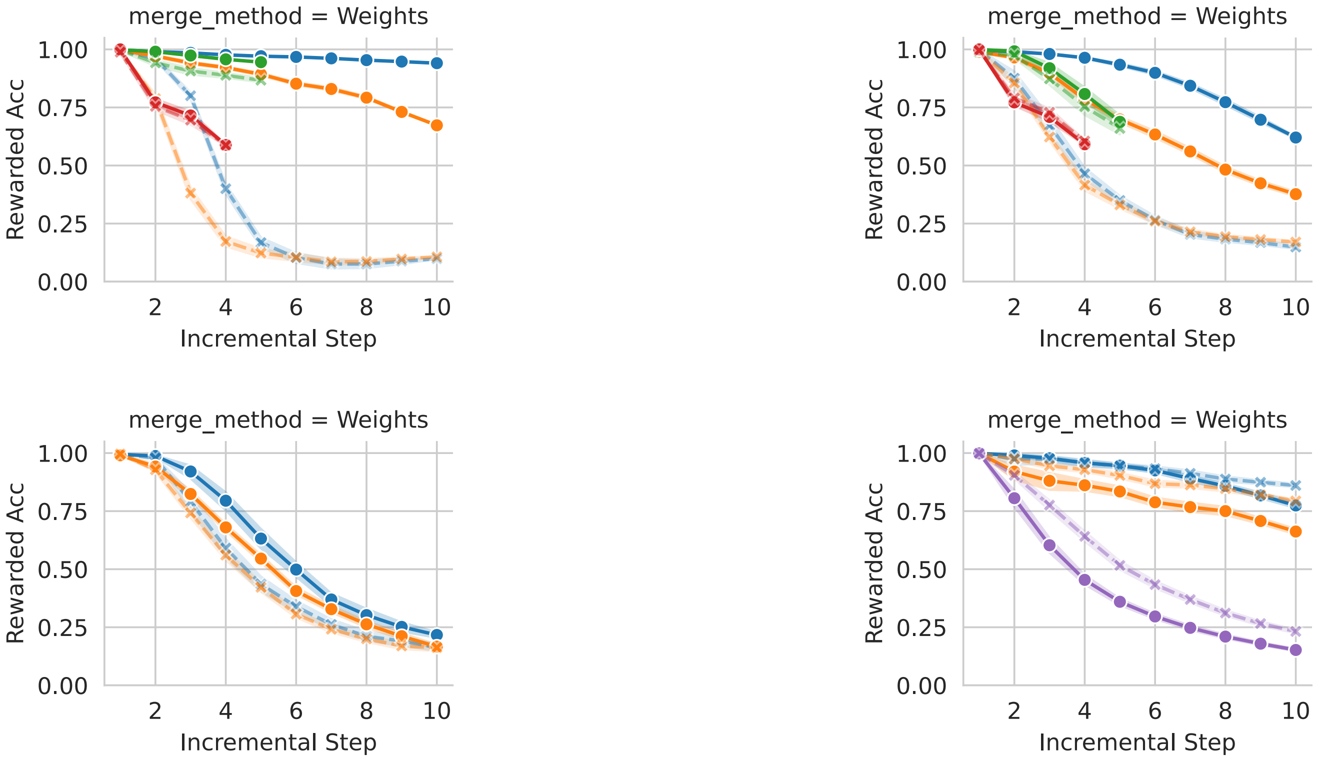
Rewarded accuracy of ResNet18 on FMNIST, MNIST and CIFAR-10 across cardinalities, loss functions, and merge strategies (logit vs. weight), with and without IMP, confirming that our framework extends to larger architectures.

Complete merge



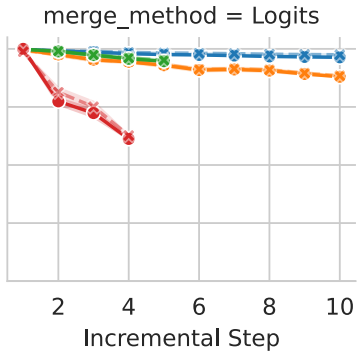
Average rewarded accuracy across incremental submodule merging across architectures

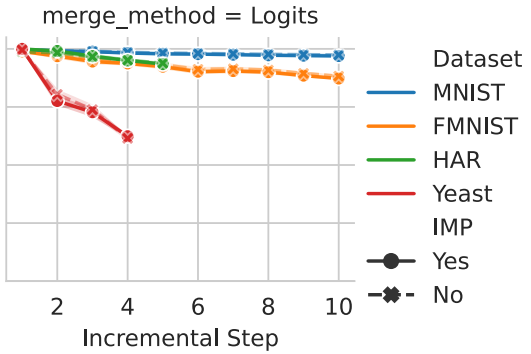
Complete merge

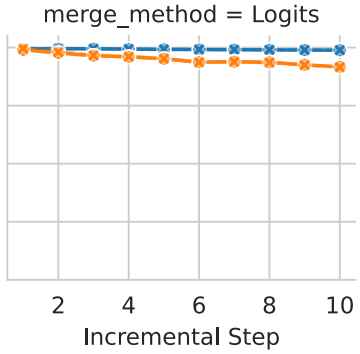


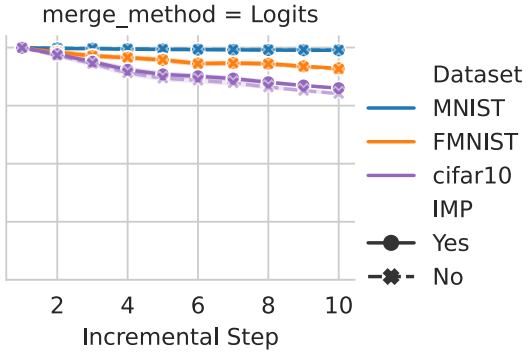
Average rewarded accuracy across incremental submodule merging across architectures

Complete merge



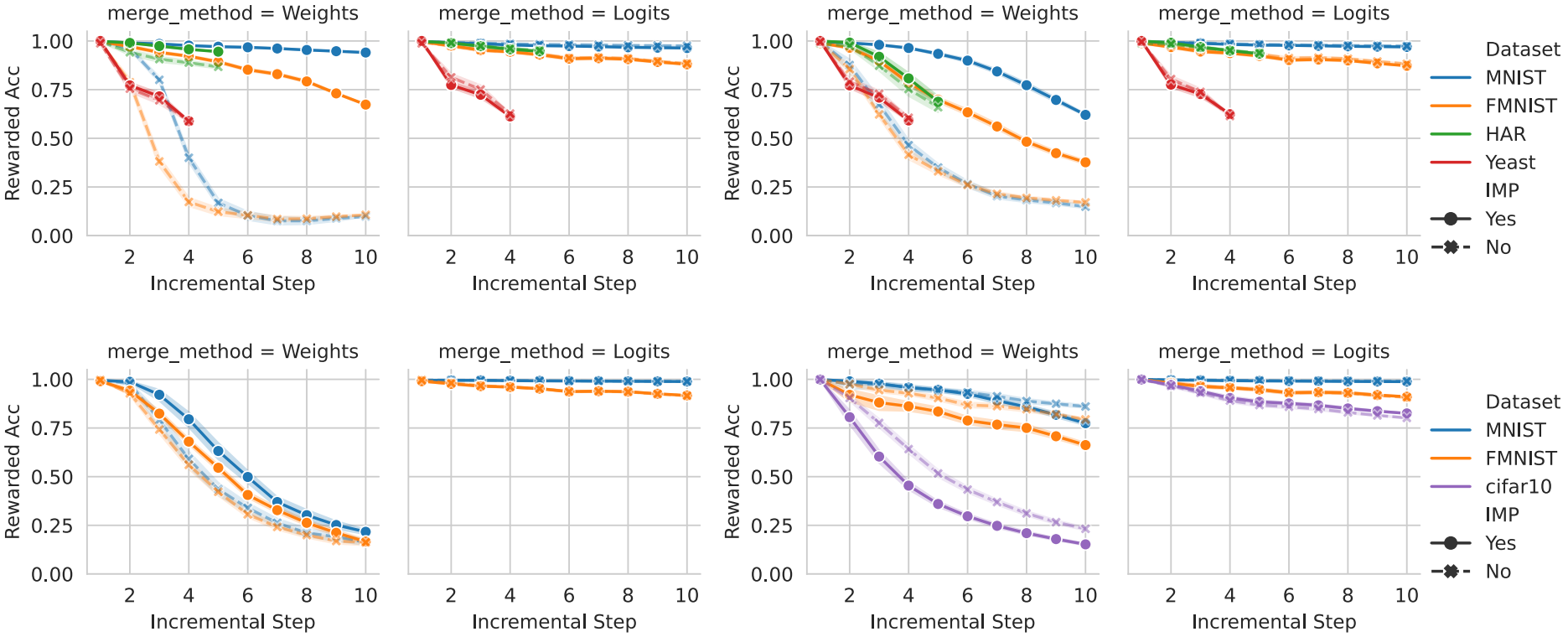






Average rewarded accuracy across incremental submodule merging across architectures

Complete merge



Average rewarded accuracy across incremental submodule merging across architectures

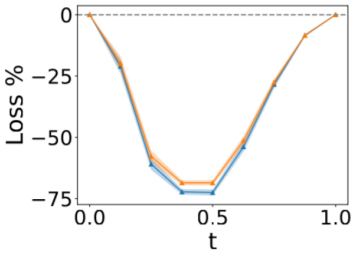
Mode connectivity

Following Frankle et al. and Lubana et al., we say that θ_1 and θ_2 are *mode connected* along a path $\gamma(t)$ if:

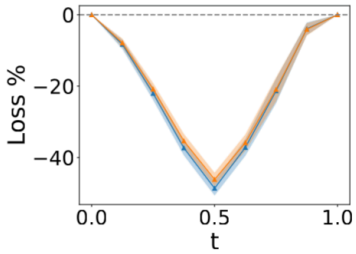
$$\forall t \in [0,1], \quad \mathscr{L}(f_{\gamma(t)}(\mathscr{D})) \leq (1-t)\mathscr{L}(f_{\theta_1}(\mathscr{D})) + t\mathscr{L}(f_{\theta_2}(\mathscr{D})) + \epsilon$$

$$\gamma_{\theta_1 \to \theta_2}(t) = \begin{cases} \theta_1 + 2t \cdot \theta_2 & \text{if } t \leq 0.5 \\ 2(1-t) \cdot \theta_1 + \theta_2 & \text{if } t > 0.5 \end{cases}$$

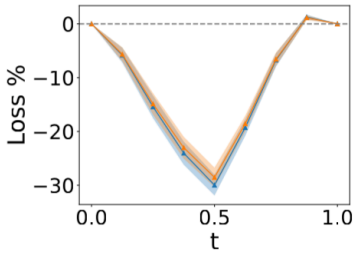
Mode connectivity of complete merge



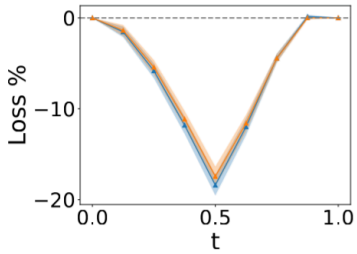
$\theta_1 = [0]$
$\theta_2 = [1]$



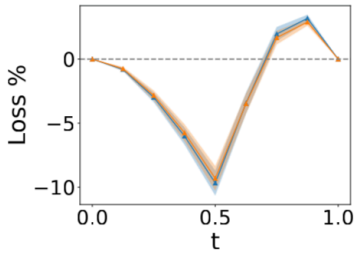
$\theta_1 = [0]+[1]+[2]$
$\theta_2 = [3]$



$\theta_1 = [0]+\ldots+[4]$
$\theta_2 = [5]$



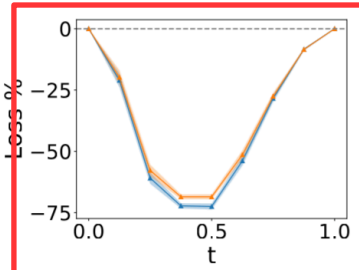
$\theta_1 = [0]+\ldots+[6]$
$\theta_2 = [7]$



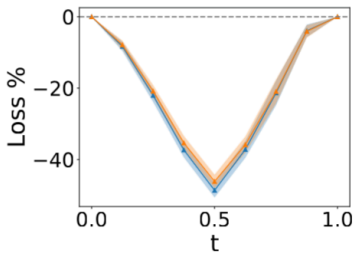
$\theta_1 = [0]+\ldots+[8]$
$\theta_2 = [9]$

Loss barrier along the linear interpolation path γθ1→θ2 (t) considering multiple steps of the complete merge starting from 0 to 9 for the FMNIST dataset (train and test errors, in blue and orange respectively, mean and std across 5 runs). Loss values at each t are relativized with respect to the corresponding interpolated errors of θ1 and θ2 and rescaled as percentages. Values close to or lower than 0 indicate mode connectivity. Barriers remain near zero throughout most of the merge sequence, supporting the hypothesis that ME training induces weight-space composability, with only mild interference emerging in the later steps

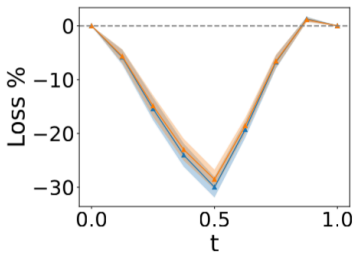
Mode connectivity of complete merge



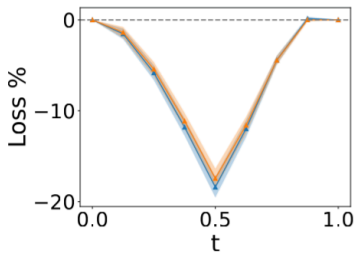
$\theta_1 = [0]$
$\theta_2 = [1]$



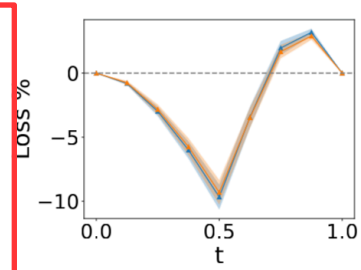
$\theta_1 = [0]+[1]+[2]$
$\theta_2 = [3]$



$\theta_1 = [0]+\ldots+[4]$
$\theta_2 = [5]$



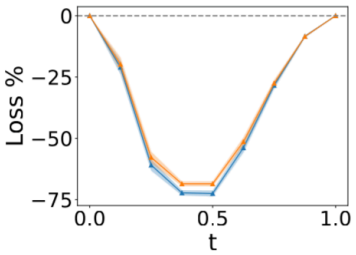
$\theta_1 = [0]+\ldots+[6]$
$\theta_2 = [7]$



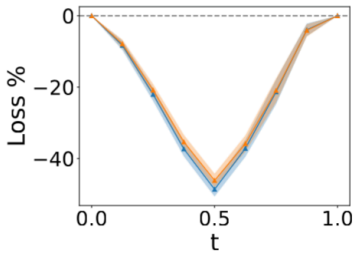
$\theta_1 = [0]+\ldots+[8]$
$\theta_2 = [9]$

Loss barrier along the linear interpolation path γθ1→θ2 (t) considering multiple steps of the complete merge starting from 0 to 9 for the FMNIST dataset (train and test errors, in blue and orange respectively, mean and std across 5 runs). Loss values at each t are relativized with respect to the corresponding interpolated errors of θ1 and θ2 and rescaled as percentages. Values close to or lower than 0 indicate mode connectivity. Barriers remain near zero throughout most of the merge sequence, supporting the hypothesis that ME training induces weight-space composability, with only mild interference emerging in the later steps

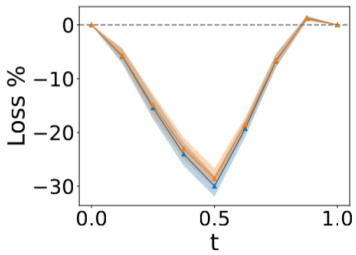
Mode connectivity of complete merge



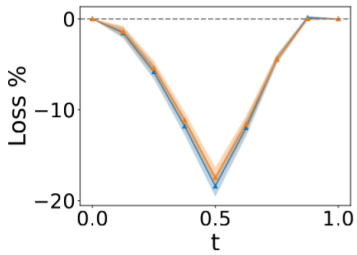
$\theta_1 = [0]$
$\theta_2 = [1]$



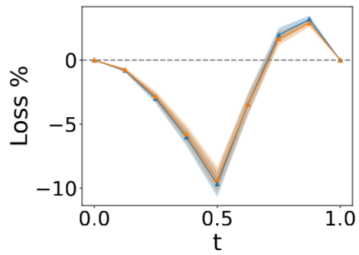
$\theta_1 = [0]+[1]+[2]$
$\theta_2 = [3]$



$\theta_1 = [0]+\ldots+[4]$
$\theta_2 = [5]$



$\theta_1 = [0]+\ldots+[6]$
$\theta_2 = [7]$



$\theta_1 = [0]+\ldots+[8]$
$\theta_2 = [9]$

Loss barrier along the linear interpolation path γθ1→θ2 (t) considering multiple steps of the complete merge starting from 0 to 9 for the FMNIST dataset (train and test errors, in blue and orange respectively, mean and std across 5 runs). Loss values at each t are relativized with respect to the corresponding interpolated errors of θ1 and θ2 and rescaled as percentages. Values close to or lower than 0 indicate mode connectivity. Barriers remain near zero throughout most of the merge sequence, supporting the hypothesis that ME training induces weight-space composability, with only mild interference emerging in the later steps

Conclusions

- Can we train functional modules? **Yes, through the MaxEnt principle.**

- Can we compose them? **Yes. For simple datasets, through Weight Summation; for complex datasets, through Logit Averaging, producing products of experts.**

Conclusions

- Can we train functional modules? **Yes, through the MaxEnt principle.**

- Can we compose them? **Yes. For simple datasets, through Weight Summation; for complex datasets, through Logit Averaging, producing products of experts.**

Conclusions

- Can we train functional modules? **Yes, through the MaxEnt principle.**

- Can we compose them? **Yes. For simple datasets, through Weight Summation; for complex datasets, through Logit Averaging, producing products of experts.**

Paper

Code