PHYDI: Initializing Parameterized Hypercomplex Neural Networks as Identity Functions

Matteo Mancanelli, Eleonora Grassucci, Aurelio Uncini, and Danilo Comminiello
DIET Dept., Sapienza University of Rome

Abstract

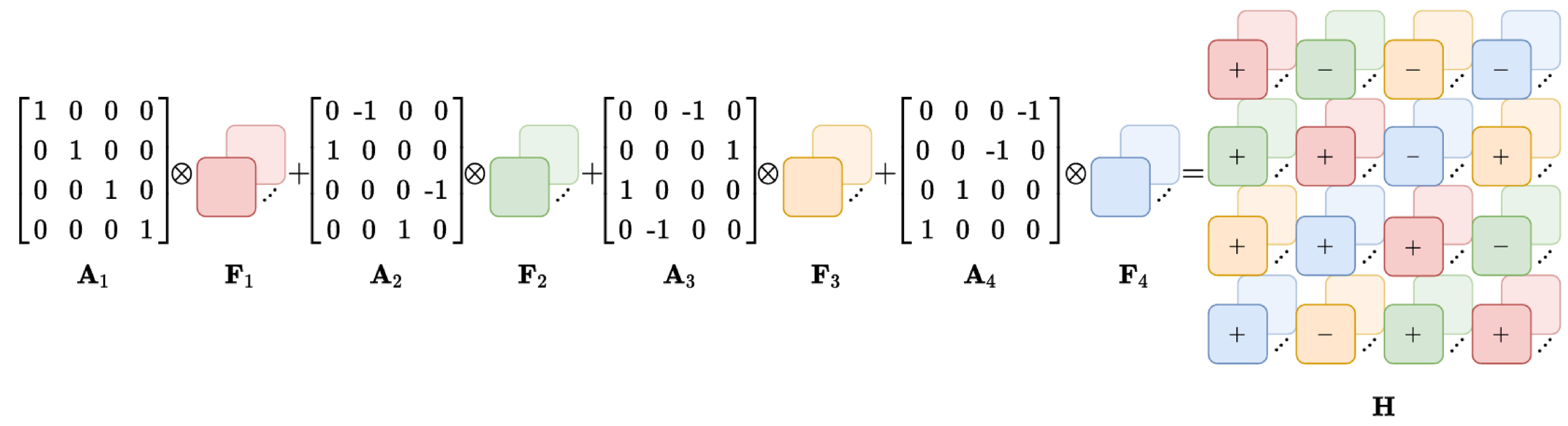
Neural models based on **hypercomplex algebra systems** are growing and prolificating for a plethora of applications, ranging from computer vision to natural language processing. Hand in hand with their adoption, **parameterized hypercomplex neural networks (PHNNs)** are growing in size and no techniques have been adopted so far to **control their convergence at a large scale**. In this paper, we study PHNNs convergence and propose parameterized hypercomplex identity initialization (PHYDI), a method to improve their convergence at different scales, leading to **more robust performance** when the number of layers scales up, while also reaching the same performance **with fewer iterations**. We show the effectiveness of this approach in different benchmarks and with common PHNNs with ResNets- and Transformer-based architecture.

PHNNs

The parameterized hypercomplex (PH) layer builds the weight matrix as a sum of Kronecker products of two sets of **learnable matrices**:

$$\mathbf{H} = \sum_{i=1}^{n} \mathbf{A}_i \otimes \mathbf{F}_i$$

PH layers can be defined with **$1/n$ parameters** with respect to ral-valued ones where n is data dimensionality. For $n = 4$ PHC is:



PHYDI: Initializing Parameterized Hypercomplex Neural Networks as Identity Functions

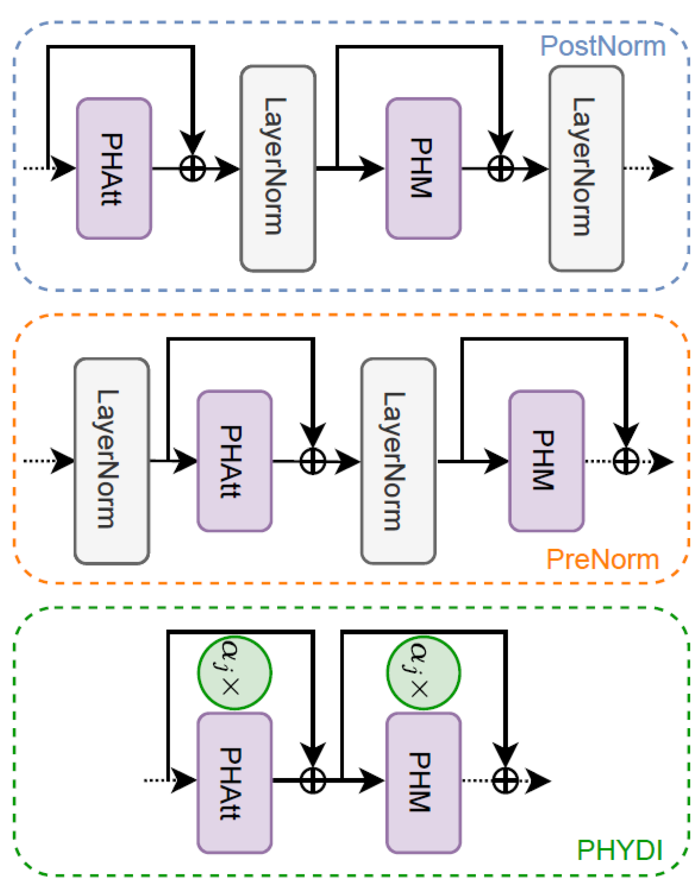
PHResNets

PHResNets can be defined through PHC layers:

$$\mathbf{x}_{j+1} = \mathbf{x} + \mathrm{PHC}(\mathrm{ReLU}(\mathrm{PHC}((\mathbf{x}))))$$

To simplify the gradient propagation at the initialization, the signal should not propagate on the PH set layer of layers, but rather on its residual connection $\mathbf{x}$. To do that, a parameter α is set to multiply the set of PH layers and initialized to 0, so that only the residual connection remains active during the first iteration:

$$\mathbf{x}_{j+1} = \mathbf{x} + \boldsymbol{\alpha}_j \mathrm{PHC}(\mathrm{ReLU}(\mathrm{PHC}((\mathbf{x}))))$$



PHTransformers with PHYDI.

PHTransformers

PHTransformers can be defined through PHM and PHAtt layers as:

$$\mathbf{x}_{j+1} = \mathrm{LayerNorm}\{\mathbf{x}_j + \mathrm{PHM}(\mathrm{LayerNorm}(\mathbf{x}_j + \mathrm{PHAtt}\,(\mathbf{x}_j))\}$$

To initialize the layer as the identity function, we can remove the layer normalization and insert the PHYDI parameters as multipliers for the sub-layers:

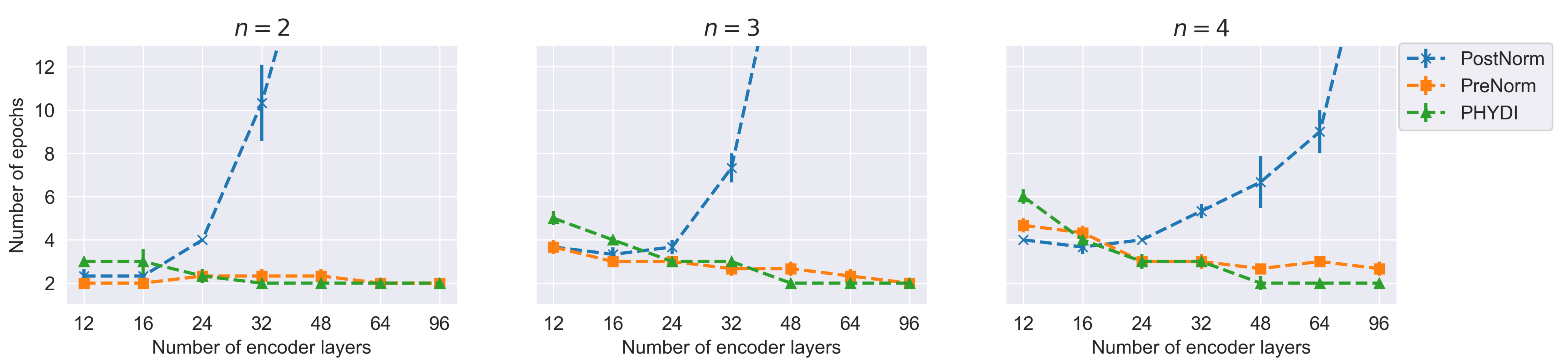
$$\mathbf{x}_{j+1} = \mathbf{x}_j + \boldsymbol{\alpha}_j\,\mathrm{PHM}(\mathbf{x}_j + \boldsymbol{\alpha}_j\,\mathrm{PHAtt}\,(\mathbf{x}_j))$$

Results

PHResNets with standard, WKP, and PHYIDI initialization for different values of the hyperparameter n in the CIFAR10 dataset. Metrics **M1**: Epochs to reach 80% of Accuracy, **M2**: # Epochs to beat one w/ PHYDI. The uncertainties correspond to standard error.

Model $n=2$	M1↓	M2	Model $n=3$	M1↓	M2	Model $n=4$	M1↓	M2
PHResNet18	6.00 ± 0.58	2	PHResNet18	6.33 ± 0.33	2	PHResNet18	$\mathbf{3.75} \pm 0.48$	1
+WKP	$\mathbf{5.75} \pm 0.25$	2	+WKP	6.00 ± 0.00	1	+WKP	5.67 ± 0.67	2
+PHYDI	6.00 ± 0.00	-	+PHYDI	$\mathbf{5.00} \pm 0.58$	-	+PHYDI	4.50 ± 0.50	-
PHResNet50	10.67 ± 1.20	3	PHResNet50	8.67 ± 1.20	2	PHResNet50	8.33 ± 0.88	2
+WKP	10.67 ± 0.67	2	+WKP	9.00 ± 0.58	2	+WKP	9.33 ± 0.88	2
+PHYDI	$\mathbf{7.00} \pm 0.58$	-	+PHYDI	$\mathbf{6.33} \pm 0.67$	-	+PHYDI	$\mathbf{7.00} \pm 1.15$	-
PHResNet152	32.67 ± 2.03	4	PHResNet152	26.67 ± 1.76	4	PHResNet152	22.67 ± 3.71	3
+WKP	29.80 ± 3.68	4	+WKP	20.00 ± 2.12	4	+WKP	20.60 ± 1.60	3
+PHYDI	$\mathbf{6.33} \pm 1.33$	-	+PHYDI	$\mathbf{4.67} \pm 0.33$	-	+PHYDI	$\mathbf{5.33} \pm 0.33$	-

The proposed PHYDI initialization speeds up the convergence of parameterized hypercomplex neural networks.



Number of epochs to reach a perplexity value ≥ 200 in the WikiText2 dataset for PH Transformers with increasing depth of the encoder model from 12 to 96. The three plots refer to different values of the hyperparameter $n = 2, 3, 4$.