Supplementary Material for Deconstructing the Ladder Network Architecture

Mohammad Pezeshki[*] MOHAMMAD.PEZESHKI@UMONTREAL.CA
Linxi Fan[⋆] LINXI.FAN@COLUMBIA.EDU
Philémon Brakel[*] PBPOP3@GMAIL.COM
Aaron Courville[*†] AARON.COURVILE@UMONTREAL.CA
Yoshua Bengio[*†] YOSHUA.BENGIO@UMONTREAL.CA
[*]Université de Montréal, [⋆]Columbia University, [†]CIFAR

1. Hyperparameters for different variants

Here we provide the best hyperparameter combinations we have found for different variants in different settings. We consider the standard deviation of additive Gaussian noise and the reconstruction penalty weights in the decoder as the hyperparameters. For each variant, we fix the best hyperparameters tuned on the validation set and run the variant 10 times with 10 different but fixed data seeds (used to choose 100 or 1000 labeled examples).

Depending on each variant and its hyperparameter space, we used either random search or grid search. Table 1 specifies the search space for hyperparameters and tables 2, 3, and 4 collect the best hyperparameter combinations for each experiment setting. In the case of MLP and AMLP combinator functions, standard deviation of the Gaussian initialization η is chosen from a grid of $(0.0001, 0.006, 0.0125, 0.025, 0.05)$. The best η values are listed in Table 5.

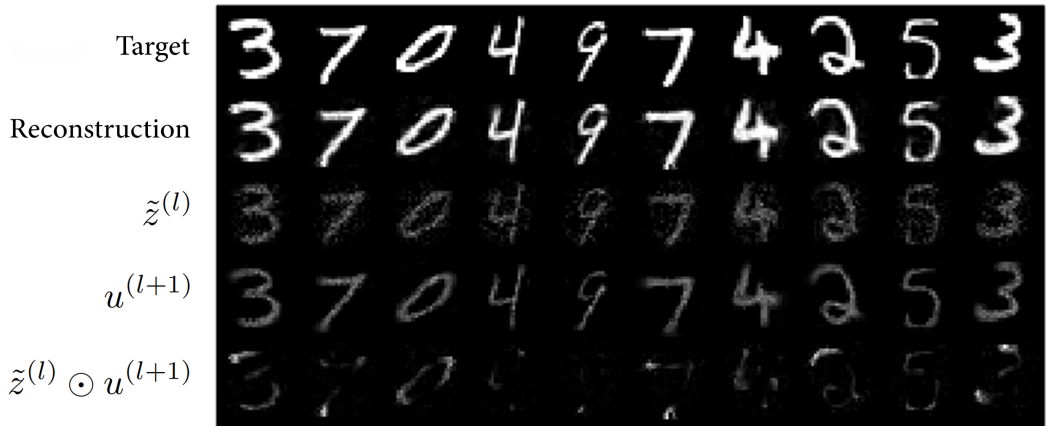

Figure 1. Visualization of three different inputs of a learned combinator function on the input layer.

Table 1. Two different hyperparameter search methods. For random search, we run 20 random hyperparameter combinations for each variant and in each task.

Search method	Noise stddev search space ($\times 10^{-1}$)	Reconstruction weights search space
Random search	(0, 0, 0, 0, 0, 0, 0) (1, 1, 1, 1, 1, 1, 1) (2, 2, 2, 2, 2, 2, 2) (3, 3, 3, 3, 3, 3, 3) (4, 4, 4, 4, 4, 4, 4) (5, 5, 5, 5, 5, 5, 5) (6, 6, 6, 6, 6, 6, 6) (7, 7, 7, 7, 7, 7, 7) (8, 8, 8, 8, 8, 8, 8)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (10.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (50.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (100.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (500.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (800.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (1000.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (2000.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (4000.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (6000.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (500, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1) (1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1) (2000, 20.0, 0.2, 0.2, 0.2, 0.2, 0.2)
Grid 100 & 1000	(2, 2, 2, 2, 2, 2, 2) (3, 3, 3, 3, 3, 3, 3) (4, 4, 4, 4, 4, 4, 4)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1) (2000, 20.0, 0.2, 0.2, 0.2, 0.2, 0.2) (5000, 50.0, 0.5, 0.5, 0.5, 0.5, 0.5) (10000, 100.0, 1.0, 1.0, 1.0, 1.0, 1.0)
Grid 60000	(2, 2, 2, 2, 2, 2, 2) (3, 3, 3, 3, 3, 3, 3) (4, 4, 4, 4, 4, 4, 4)	(500, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (1000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (2500, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) (5000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)

Table 2. Best hyperparameters for the semi-supervised task with 100 labeled examples.

Variant	Search method	Best noise stddev ($\times 10^{-1}$)	Best reconstruction weights
Baseline+noise	Random	(3, 3, 3, 3, 3, 3, 3)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
Vanilla	Grid	(3, 3, 3, 3, 3, 3, 3)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
FirstNoise	Random	(6, 0, 0, 0, 0, 0, 0)	(1000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
FirstRecons	Random	(3, 3, 3, 3, 3, 3, 3)	(1000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
FirstN&R	Random	(6, 0, 0, 0, 0, 0, 0)	(1000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
NoLateral	Random	(7, 0, 0, 0, 0, 0, 0)	(100.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
RandInit	Grid	(3, 3, 3, 3, 3, 3, 3)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
RevInit	Grid	(3, 3, 3, 3, 3, 3, 3)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
NoSig	Grid	(3, 3, 3, 3, 3, 3, 3)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
NoMult	Grid	(3, 3, 3, 3, 3, 3, 3)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
Linear	Grid	(3, 3, 3, 3, 3, 3, 3)	(2000, 20.0, 0.2, 0.2, 0.2, 0.2, 0.2)
Gaussian	Grid	(3, 3, 3, 3, 3, 3, 3)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
GatedGauss	Grid	(2, 2, 2, 2, 2, 2, 2)	(2000, 20.0, 0.2, 0.2, 0.2, 0.2, 0.2)
MLP[4]	Grid	(2, 2, 2, 2, 2, 2, 2)	(5000, 50.0, 0.5, 0.5, 0.5, 0.5, 0.5)
MLP[2,2]	Grid	(2, 2, 2, 2, 2, 2, 2)	(2000, 20.0, 0.2, 0.2, 0.2, 0.2, 0.2)
MLP[2,2,2]	Grid	(2, 2, 2, 2, 2, 2, 2)	(10000, 100.0, 1.0, 1.0, 1.0, 1.0, 1.0)
AMLP[4]	Grid	(3, 3, 3, 3, 3, 3, 3)	(2000, 20.0, 0.2, 0.2, 0.2, 0.2, 0.2)
AMLP[2,2]	Grid	(3, 3, 3, 3, 3, 3, 3)	(2000, 20.0, 0.2, 0.2, 0.2, 0.2, 0.2)
AMLP[2,2,2]	Grid	(3, 3, 3, 3, 3, 3, 3)	(1000, 10.0, 0.2, 0.2, 0.2, 0.2, 0.2)

Table 3. Best hyperparameters for the semi-supervised task with 1000 labeled examples.

Variant	Search method	Best noise stddev ($\times 10^{-1}$)	Best reconstruction weights
Baseline+noise	Random	(2, 2, 2, 2, 2, 2, 2)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
Vanilla	Grid	(2, 2, 2, 2, 2, 2, 2)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
FirstNoise	Random	(6, 0, 0, 0, 0, 0, 0)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
FirstRecons	Random	(3, 3, 3, 3, 3, 3, 3)	(4000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
FirstN&R	Random	(6, 0, 0, 0, 0, 0, 0)	(1000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
NoLateral	Random	(6, 0, 0, 0, 0, 0, 0)	(100.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
RandInit	Grid	(2, 2, 2, 2, 2, 2, 2)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
RevInit	Grid	(2, 2, 2, 2, 2, 2, 2)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
NoSig	Grid	(2, 2, 2, 2, 2, 2, 2)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
NoMult	Grid	(3, 3, 3, 3, 3, 3, 3)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
Linear	Grid	(3, 3, 3, 3, 3, 3, 3)	(2000, 20.0, 0.2, 0.2, 0.2, 0.2, 0.2)
Gaussian	Grid	(3, 3, 3, 3, 3, 3, 3)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
GatedGauss	Grid	(2, 2, 2, 2, 2, 2, 2)	(1000, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
MLP[4]	Grid	(3, 3, 3, 3, 3, 3, 3)	(10000, 100.0, 1.0, 1.0, 1.0, 1.0, 1.0)
MLP[2,2]	Grid	(3, 3, 3, 3, 3, 3, 3)	(5000, 50.0, 0.5, 0.5, 0.5, 0.5, 0.5)
MLP[2,2,2]	Grid	(3, 3, 3, 3, 3, 3, 3)	(2000, 20.0, 0.2, 0.2, 0.2, 0.2, 0.2)
AMLP[4]	Grid	(3, 3, 3, 3, 3, 3, 3)	(5000, 50.0, 0.5, 0.5, 0.5, 0.5, 0.5)
AMLP[2,2]	Grid	(3, 3, 3, 3, 3, 3, 3)	(2000, 20.0, 0.2, 0.2, 0.2, 0.2, 0.2)
AMLP[2,2,2]	Grid	(3, 3, 3, 3, 3, 3, 3)	(1000, 10.0, 0.2, 0.2, 0.2, 0.2, 0.2)

Table 4. Best found hyperparameters for the task of semi-supervised with 60000 labeled examples.

Variant	Search method	Best noise stddev ($\times 10^{-1}$)	Best reconstruction weights
Baseline+noise	Random	(3, 3, 3, 3, 3, 3, 3)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
Vanilla	Grid	(3, 3, 3, 3, 3, 3, 3)	(500, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
FirstNoise	Random	(6, 0, 0, 0, 0, 0, 0)	(500, 10.0, 0.1, 0.1, 0.1, 0.1, 0.1)
FirstRecons	Random	(3, 3, 3, 3, 3, 3, 3)	(500, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
FirstN&R	Random	(6, 0, 0, 0, 0, 0, 0)	(500, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
NoLateral	Random	(6, 0, 0, 0, 0, 0, 0)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
RandInit	Grid	(3, 3, 3, 3, 3, 3, 3)	(500, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
RevInit	Grid	(3, 3, 3, 3, 3, 3, 3)	(500, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
NoSig	Grid	(3, 3, 3, 3, 3, 3, 3)	(500, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
NoMult	Grid	(3, 3, 3, 3, 3, 3, 3)	(500, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
Linear	Grid	(3, 3, 3, 3, 3, 3, 3)	(500, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
Gaussian	Grid	(3, 3, 3, 3, 3, 3, 3)	(1000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
GatedGauss	Grid	(3, 3, 3, 3, 3, 3, 3)	(2000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
MLP[4]	Grid	(3, 3, 3, 3, 3, 3, 3)	(2000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
MLP[2,2]	Grid	(3, 3, 3, 3, 3, 3, 3)	(2000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
MLP[2,2,2]	Grid	(3, 3, 3, 3, 3, 3, 3)	(1000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
AMLP[4]	Grid	(3, 3, 3, 3, 3, 3, 3)	(2000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
AMLP[2,2]	Grid	(3, 3, 3, 3, 3, 3, 3)	(2000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
AMLP[2,2,2]	Grid	(3, 3, 3, 3, 3, 3, 3)	(2000, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)

Table 5. The best MLP initialization η for all settings.

MLP variant	100 labels	1000 labels	fully-labeled
MLP[4]	0.006	0.006	0.0125
MLP[2,2]	0.05	0.0125	0.05
MLP[2,2,2]	0.025	0.025	0.05
AMLP[4]	0.006	0.025	0.0125
AMLP[2,2]	0.0125	0.0125	0.025
AMLP[2,2,2]	0.006	0.006	0.006