Simplicial Embeddings

In Self-supervised Learning and downstream classification

Samuel Lavoie, Christos Tsirigotis, Max Schwarzer,
Kenji Kawaguchi, Ankit Vani, Aaron Courville

Mila
Université de Montréal

Self-Supervised Learning

Learn a representation of the data using a pretext task.

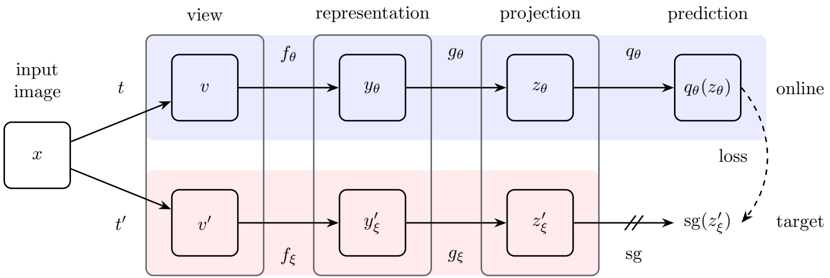
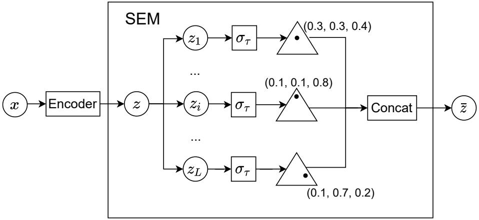


Figure 1: Taken from Grill et al. 2020.

BYOL's objective

$$\mathcal{L} := 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{||q_\theta(z_\theta)||_2 \cdot ||z'_\xi||_2}$$

Simplicial Embeddings

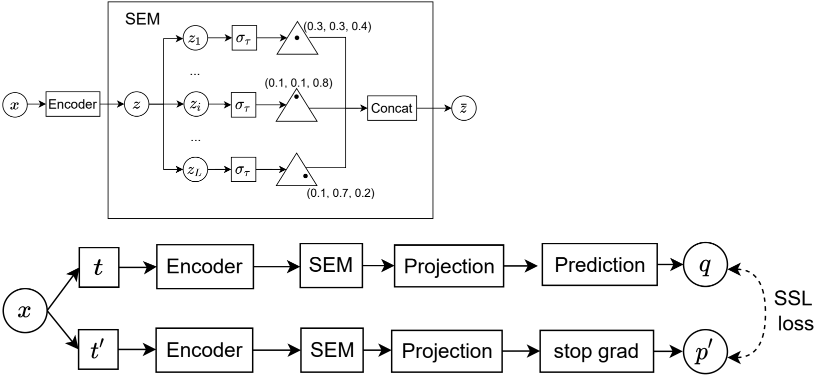


Simplicial Embeddings

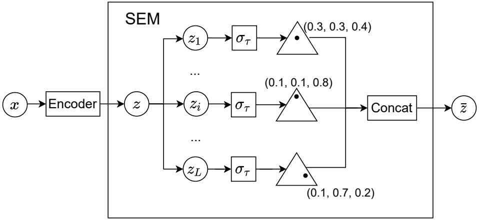
$$z \in \mathbb{R}^{LV},\ \bar{z}_i := [\sigma_\tau(z_{i1}),\ \ldots, \sigma_\tau(z_{iV})], \sigma_\tau(z_{ij}) = \frac{e^{z_{ij}/\tau}}{\sum_{k=1}^{V} e^{z_{ik}/\tau}},$$

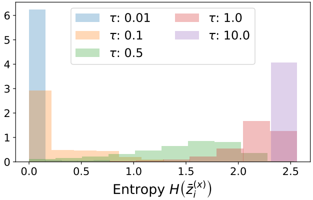
$$\bar{z} := \text{Concat}(\bar{z}_1, \ldots, \bar{z}_L)$$

Simplicial Embeddings in SSL



Simplicial Embeddings – An inductive bias





Simplicial Embeddings

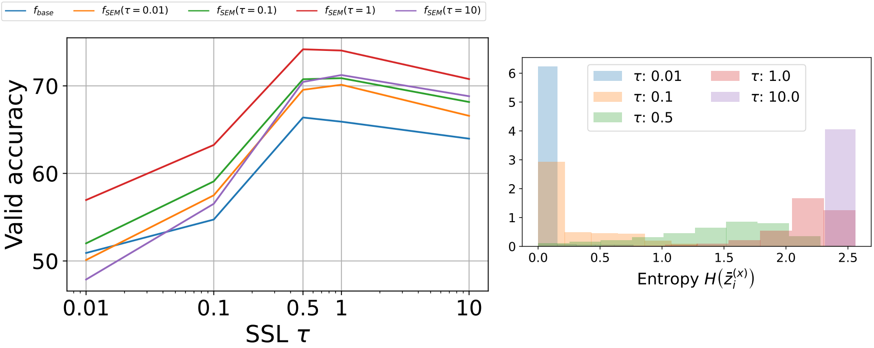
Each simplicial embedding $\bar{z}_i^{(x)}$ for a sample x represents a p.m.f. We can compute the entropy of a simplex $\bar{z}_i^{(x)}$ for a sample x:

$$H(\bar{z}_i^{(x)}) := -\sum_{j=1}^{V} p(\bar{z}_{ij}^{(x)}) \log p(\bar{z}_{ij}^{(x)}).$$

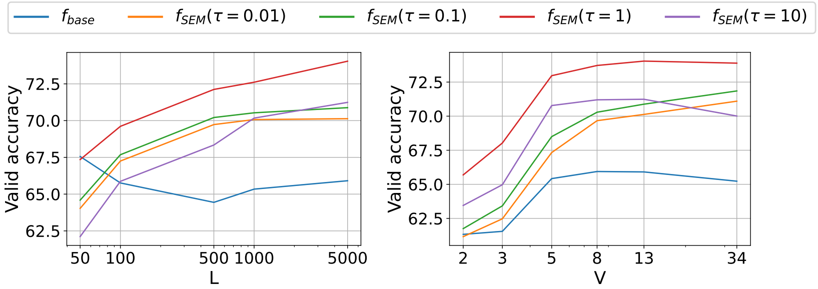
Simplicial Embeddings – For downstream classification

Two temperatures

1. SEMs to embed the data before classification: Controls the expressivity of the representation.
2. SEMs during the pre-training.



Simplicial Embeddings – For downstream classification



Comparison

Method	Accuracy
SimCLR†	65.78
MOCO†	69.89
SwAV†	64.88
DINO†	66.76
BYOL	70.46
BYOL + SEM	**74.36**

(a) CIFAR100 on ResNet18

Method	Accuracy
SimCLR‡	68.73
SimCLR + IP‡	73.99
BYOL	74.28
BYOL*	73.33
BYOL + SEM	**77.05**

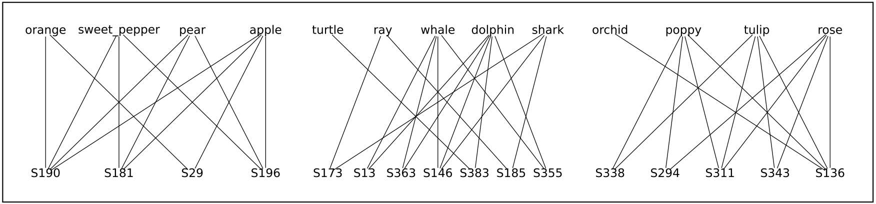
(b) CIFAR100 on ResNet50

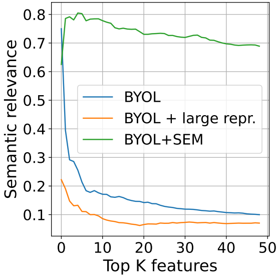
Method	Accuracy
SimCLR‡	63.1
MoCo‡	67.3
MoCo + IP‡	67.6
MoCo + SEM	69.0
SimSiam‡	70.0
SimSiam + IP‡	69.1
BYOL	70.6
BYOL + SEM	**72.8**

(c) ImageNet on ResNet50

Table 1

Emergence of semantical relevance in the features





Semantic relevance

$\mathcal{W}_K$ is a bipartite graph Features - Categories. $\mathcal{N}(c_i)$ returns the 2-neighbours of c_i in $\mathcal{W}_K$. is_super returns 1 if c_i and c_j are elements of the same super class.

$$\text{Relevance}(\mathcal{W}_K) := \sum_{i=1}^{C} \frac{\sum_{(c_i, c_j) \in \mathcal{N}(c_i)} \text{is\_super}(c_i, c_j)}{|\mathcal{N}(c_i)|}, \tag{1}$$

Transfer learning and robustness

	IN	IN-V2	IN-R	IN-A	IN-C
BYOL	68.3	55.3	16.5	0.68	35.4
BYOL + SEM	**70.6**	**57.9**	**18.1**	**0.77**	**38.9**
MoCo	66.7	53.4	14.0	**0.69**	31.1
MoCo + SEM	**68.0**	**55.0**	**15.21**	0.61	**33.8**

Table 2: Test accuracies of a linear probe trained with 100% and 1% of the ImageNet samples on a pre-trained representation trained for 100 epochs. Boldface indicates the maximal value for each evaluation set and each base model type (BYOL or MoCo).

Transfer learning and robustness

	Food	Cifar10	Cifar100	SUN	DTD	Pets	Flowers	Caltech	Cars
BYOL	71.3	89.5	71.4	57.6	71.5	85.4	84.6	77.8	45.7
BYOL + SEM	**74.1**	**92.0**	**76.3**	**60.5**	**72.5**	**87.1**	**88.6**	**82.4**	**57.3**
MoCo	70.6	88.6	69.5	57.6	**70.9**	82.3	81.5	74.3	39.8
MoCo + SEM	**71.0**	**89.6**	**72.8**	**58.6**	**70.9**	**83.8**	**84.5**	**77.5**	**45.2**

Table 3: Transfer learning accuracy by training a linear probe on a pre-trained representation with ImageNet for 100 epochss. Boldface indicates the maximal value for each transfer dataset and each base model type (BYOL or MoCo).

Conclusion

ArXiv available

https://arxiv.org/abs/2204.00616