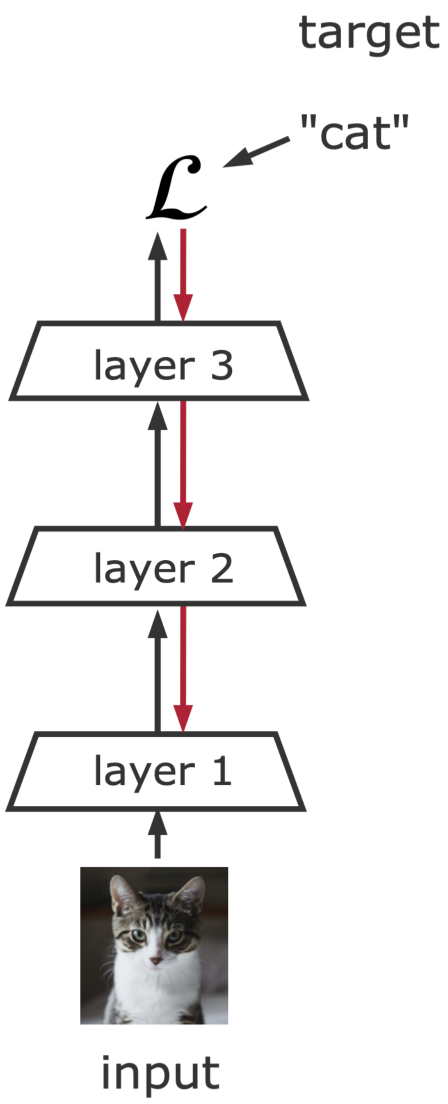
Putting An End to End-to-End: Gradient-Isolated Learning of Representations

Sindy Löwe*, Peter O'Connor, Bastiaan S. Veeling*
AMLab, University of Amsterdam

NeurIPS 2019

*equal contribution

We can train a neural network
without end-to-end backpropagation
and achieve competitive performance.



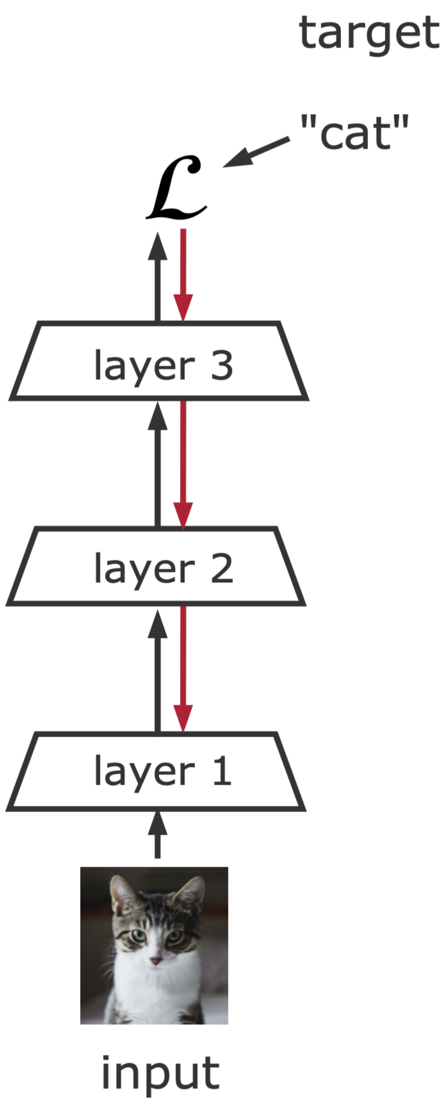
Computational Issues of End-to-End Backpropagation

- Creates substantial memory overhead

Computational Issues of End-to-End Backpropagation

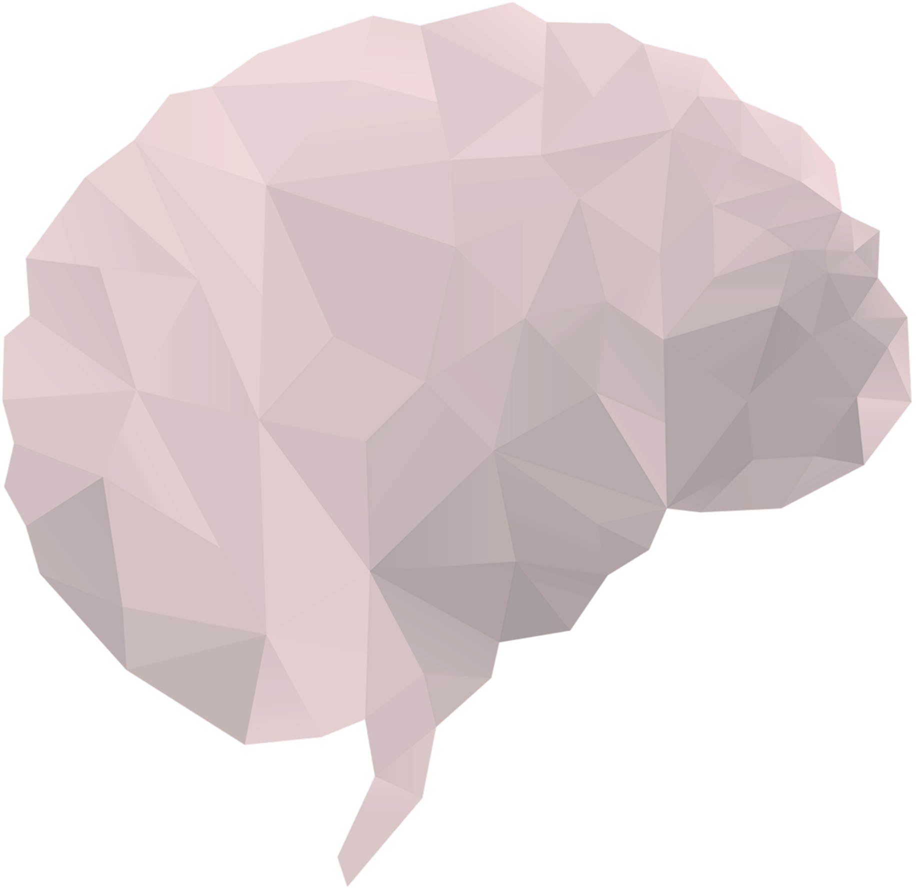
- Creates substantial memory overhead

GPU memory



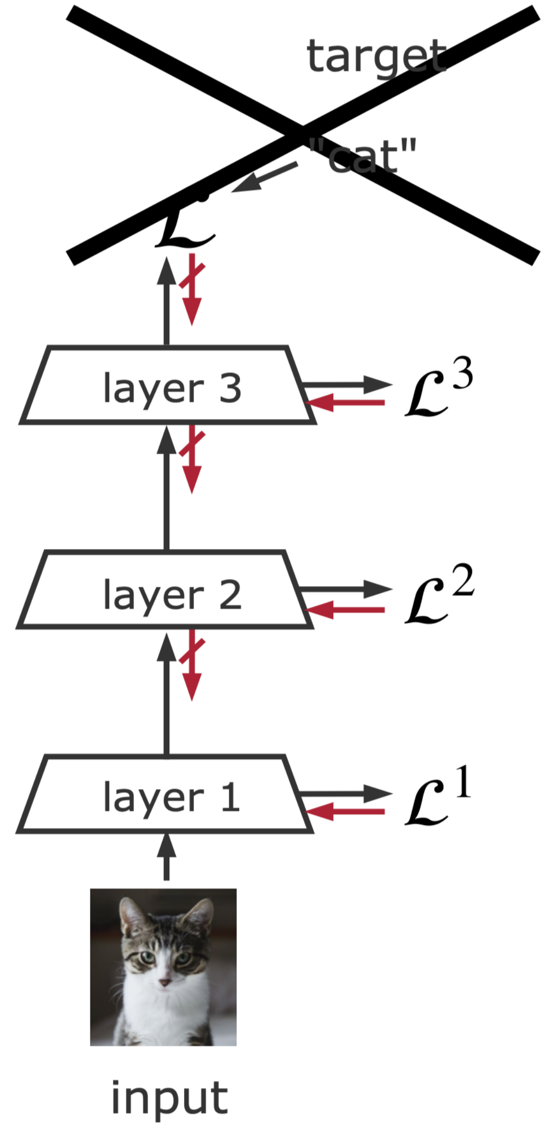
Computational Issues of End-to-End Backpropagation

- Creates substantial memory overhead
- Locking prevents massive parallelization of training



Biological Inspiration

- Brain learns predominantly based on local information



Greedy InfoMax (GIM)

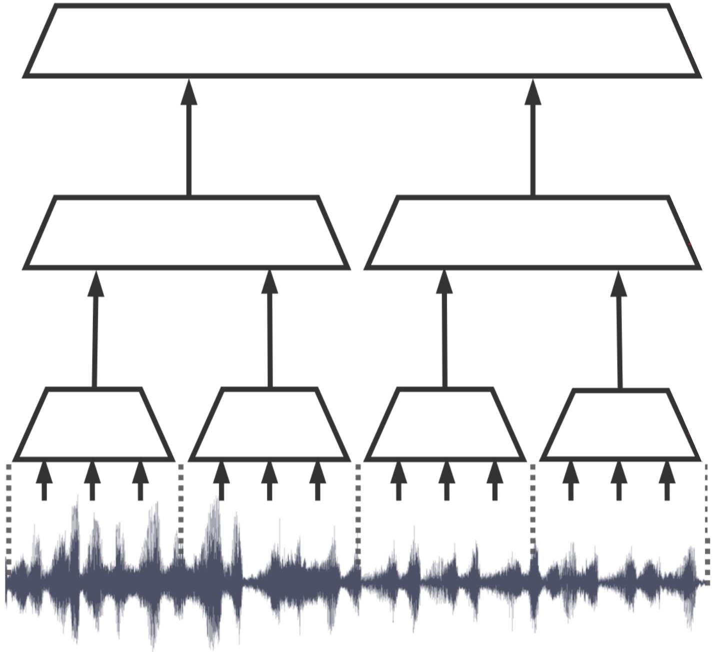
- Divide architecture into separate modules that are trained greedily with local loss per module

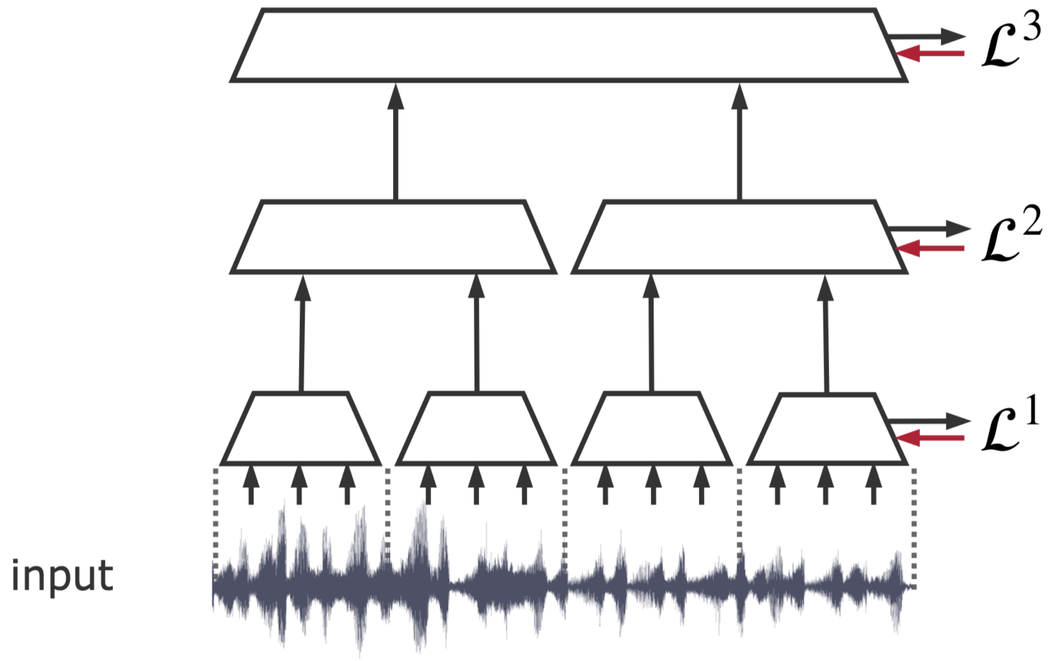
---

- Employ self-supervised loss for representation learning

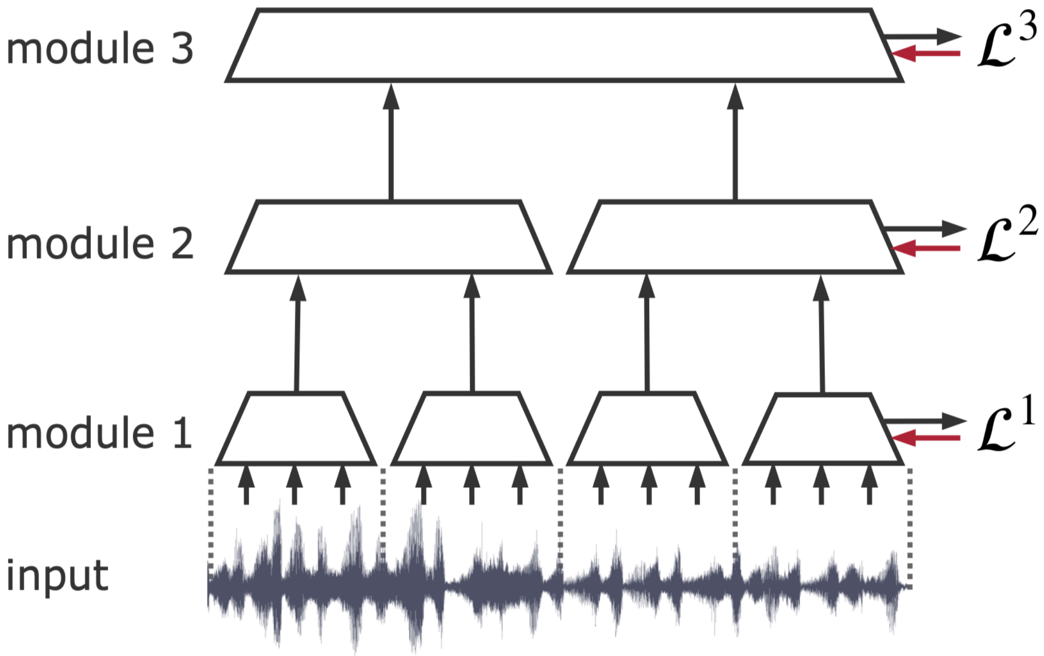
We can train a neural network **without end-to-end backpropagation** and achieve competitive performance.

input



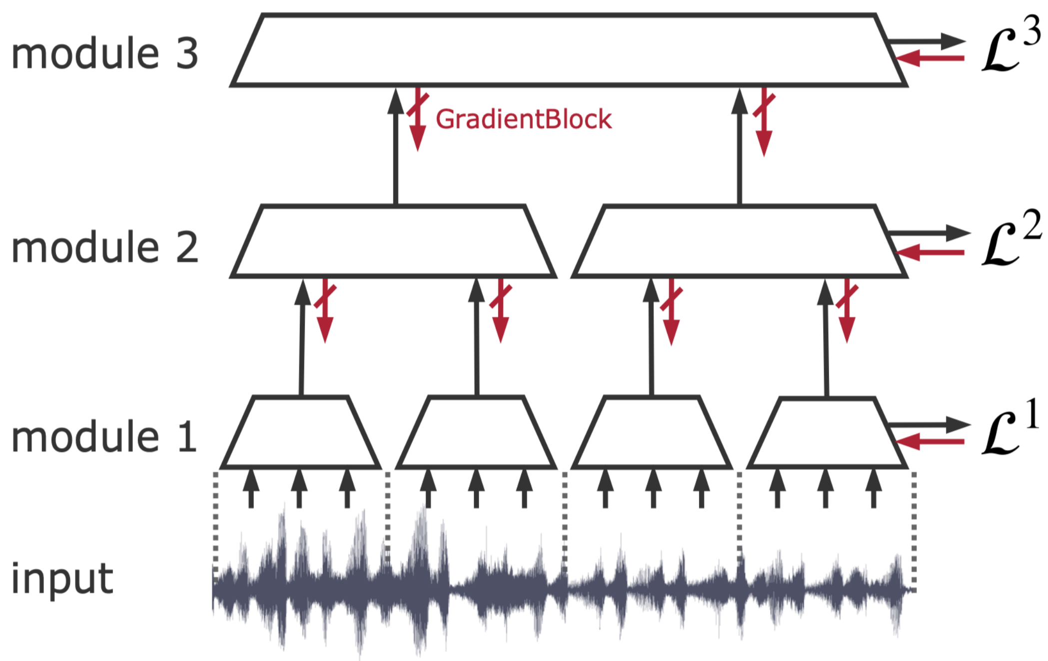


Use local losses



Use local losses

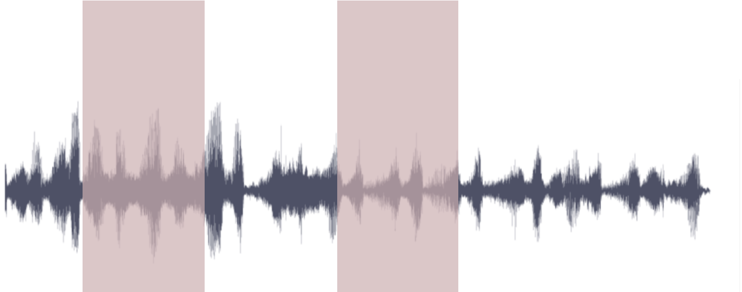
Split architecture at the “module” level



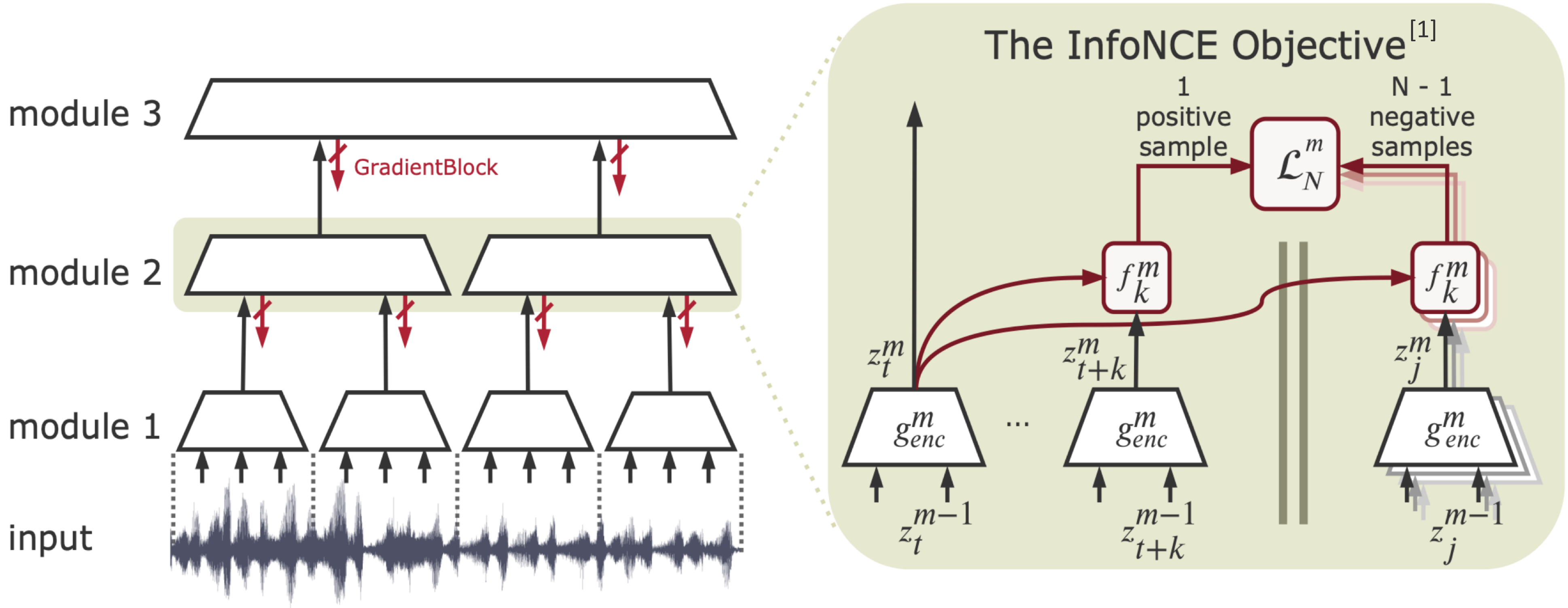
Use local losses

Split architecture at the “module” level

Block gradient flow

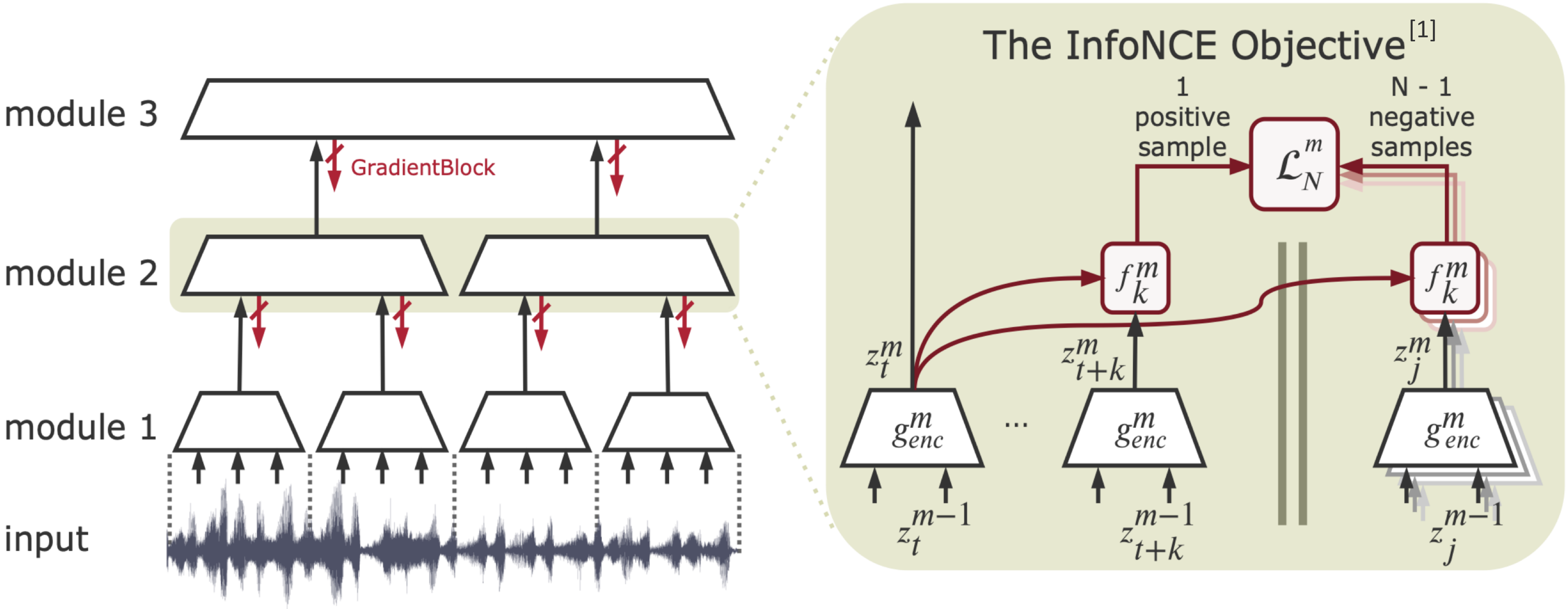


”Cats are awesome.”



[1] Oord et.al.: Representation learning with contrastive predictive coding. arXiv, 2018

InfoNCE Objective preserves information between temporally nearby patches



[1] Oord et.al.: Representation learning with contrastive predictive coding. arXiv, 2018

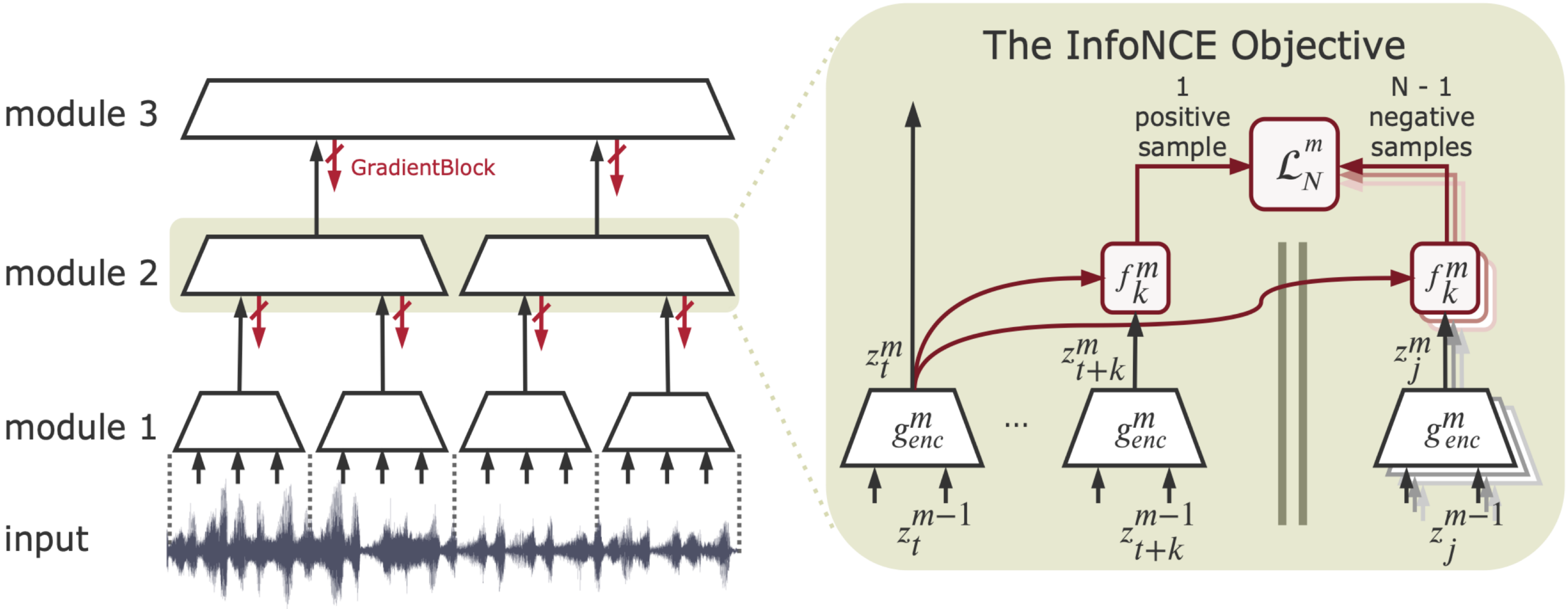
InfoNCE Objective maximizes Mutual Information between temporally nearby representations:

$$\max I(z_t^m, z_{t+k}^m) \overset{[2]}{\leq} \max I(z_t^m, z_{t+k}^{m-1})$$

Output at *t*

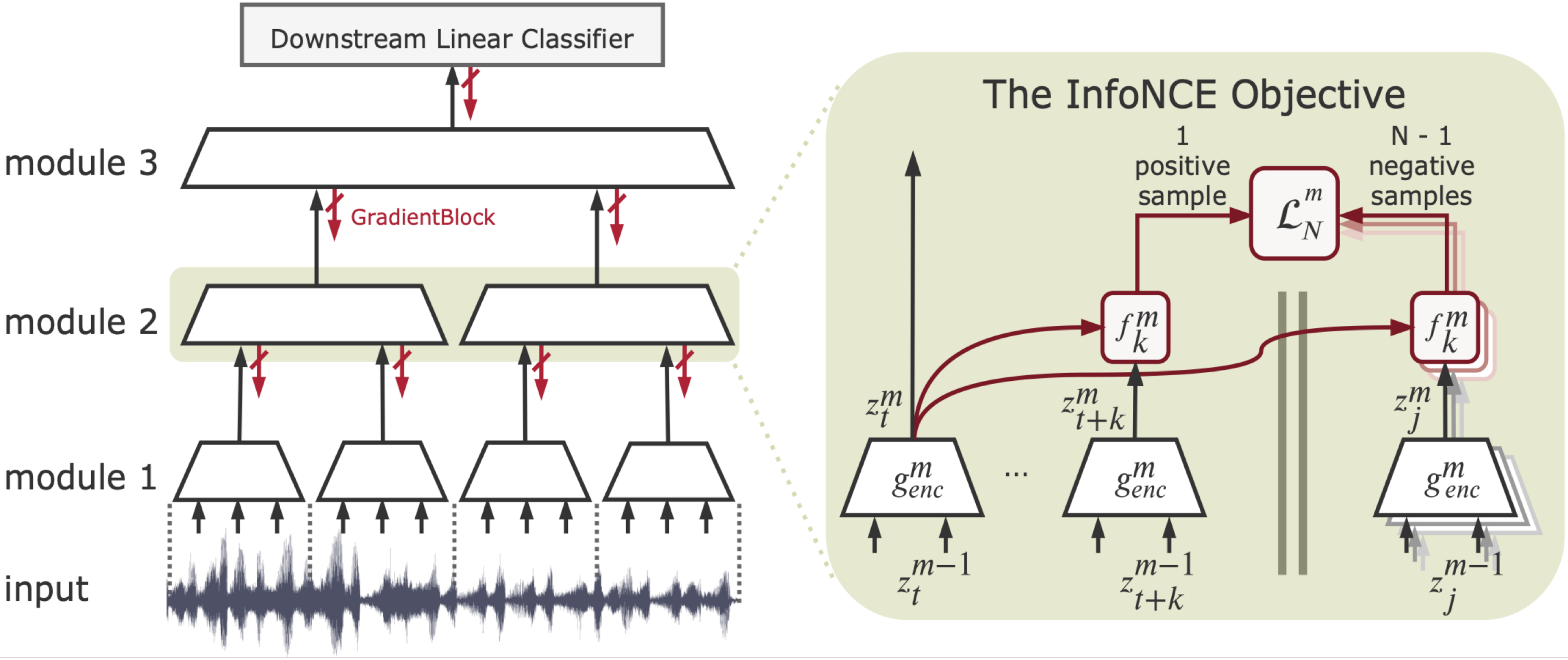
Input at *t+k*

[2] Tschannen et.al.: On mutual information maximization for representation learning. arXiv, 2019



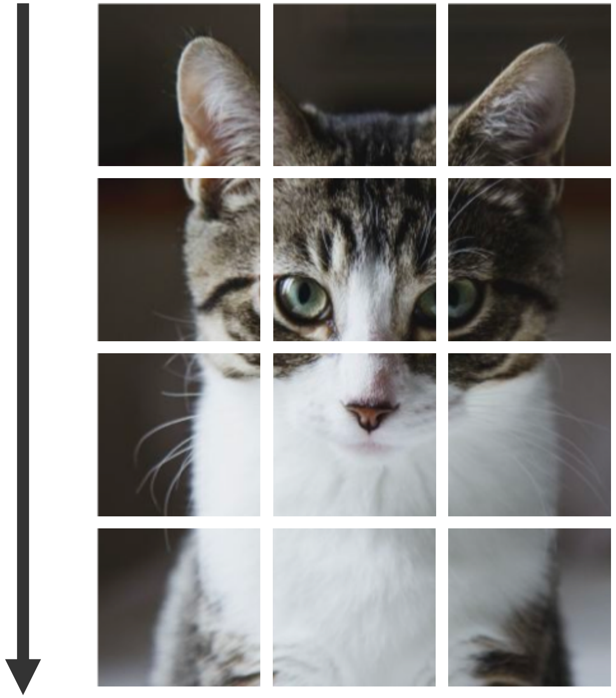
module 3
GradientBlock
module 2
module 1
input
The InfoNCE Objective
1
positive
sample
N - 1
negative
samples
$\mathcal{L}_N^m$
f_k^m
f_k^m
z_t^m
z_{t+k}^m
z_j^m
g_{enc}^m
...
g_{enc}^m
g_{enc}^m
z_t^{m-1}
z_{t+k}^{m-1}
z_j^{m-1}

Measure quality of representations using linear classifier



We can train a neural network
without end-to-end backpropagation
and achieve competitive performance.

Top-down
ordering



Performance on STL-10 Images

GIM outperforms CPC

Method	Accuracy (%)
Randomly initialized	27.0
Supervised	71.4
Greedy Supervised	65.2
CPC	$80.5^{*} \pm 3.1$
Greedy InfoMax (GIM)	$\mathbf{81.9}^{*} \pm 0.3$

*leveraging unlabeled part of STL-10 dataset

Performance on STL-10 Images

GIM outperforms comparable SOTA models

Method	Accuracy (%)
Randomly initialized	27.0
Supervised	71.4
Greedy Supervised	65.2
CPC	80.5* ± 3.1
Greedy InfoMax (GIM)	**81.9*** ± 0.3
Deep InfoMax (Hjelm et al., 2019)	78.2*
Predsim (Nøkland and Eidnes, 2019)	80.8

*leveraging unlabeled part of STL-10 dataset

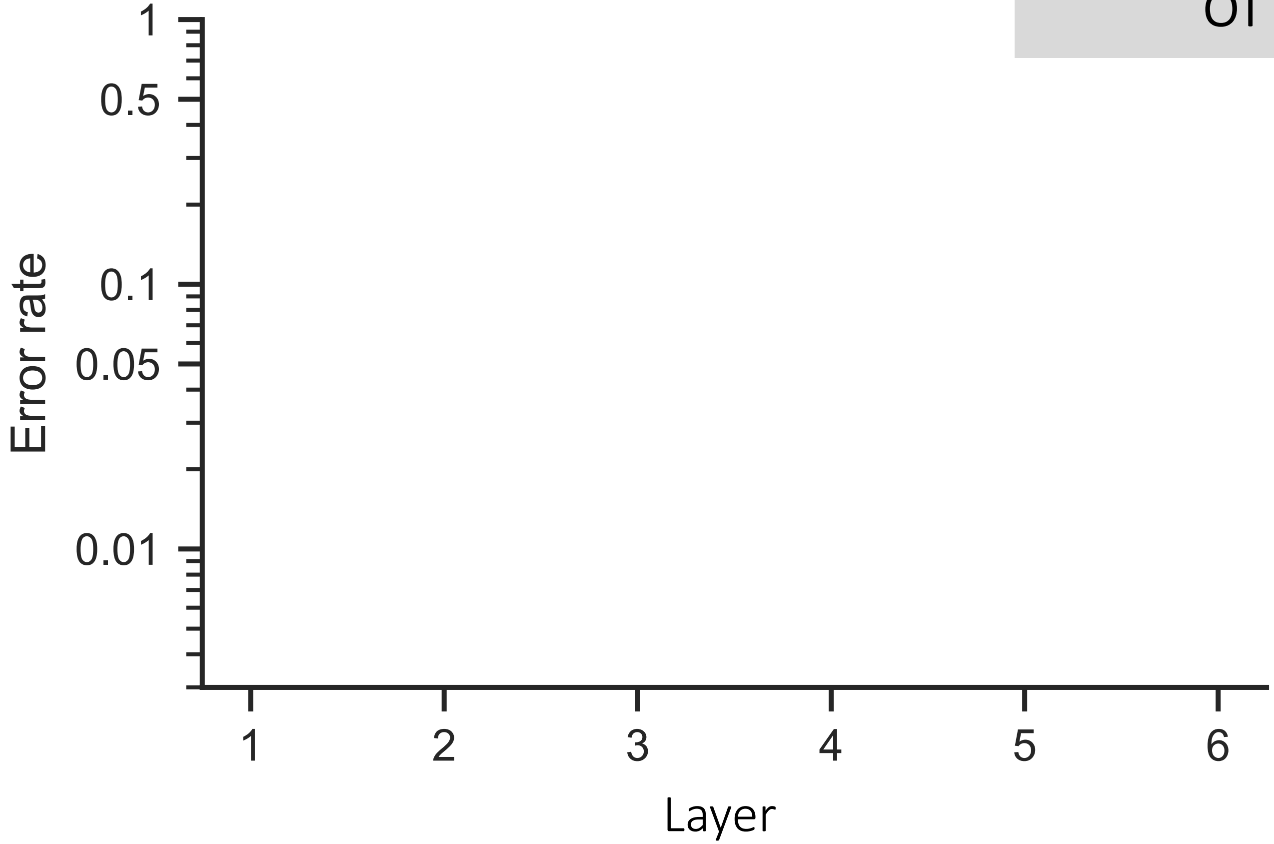
Performance on LibriSpeech

GIM slightly outperformed by CPC on phone classification

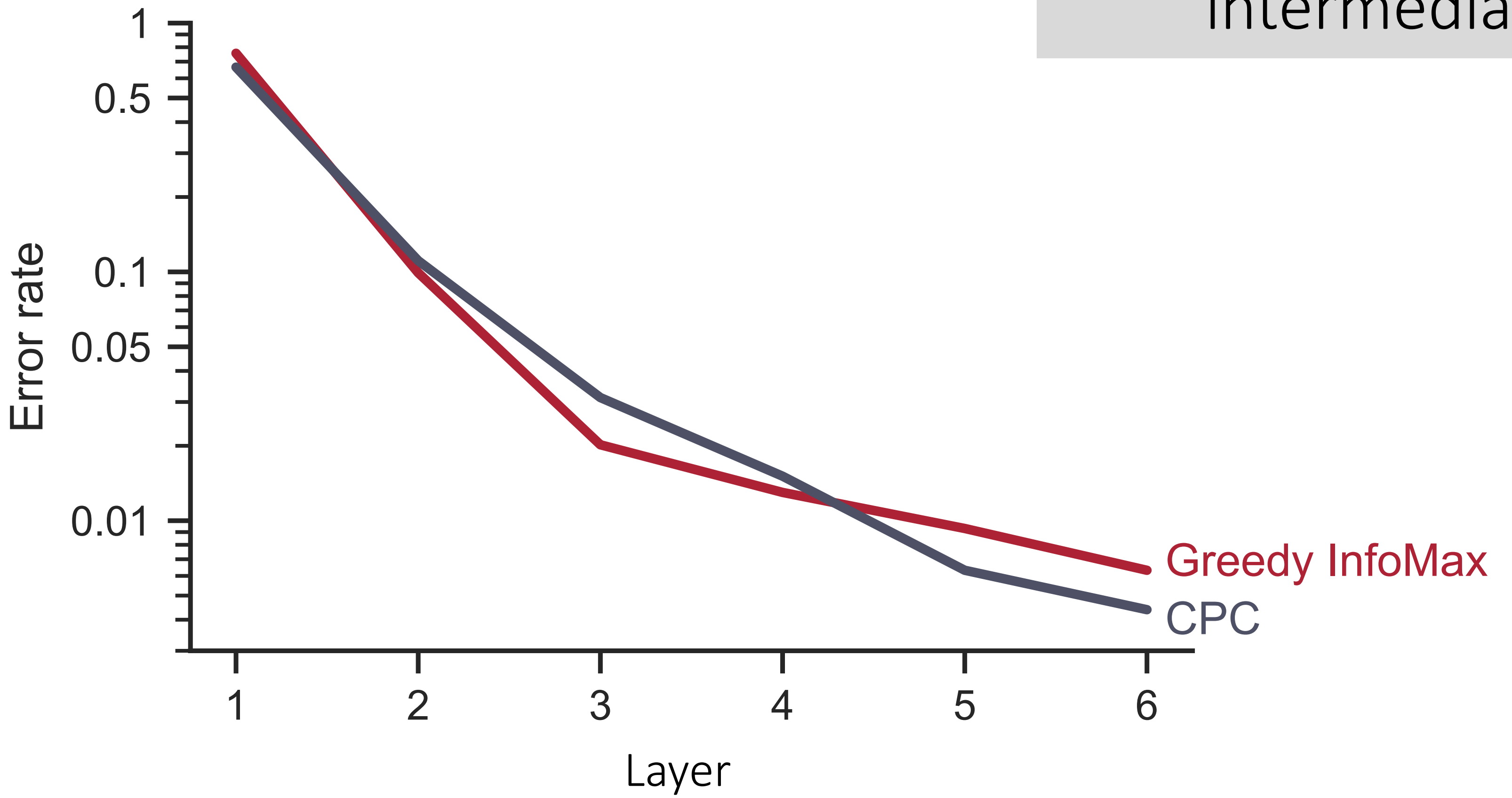
Method	Speaker Classification Accuracy (%)	Phone Classification Accuracy (%)
Randomly initialized	1.9	27.6
MFCC features	17.6	39.7
Supervised	98.9	77.7
Greedy Supervised	98.7	73.4
CPC (Oord et al., 2018)	99.6	64.9
Greedy InfoMax (GIM)	99.4	62.5

GIM and CPC achieve equivalent performance on speaker classification

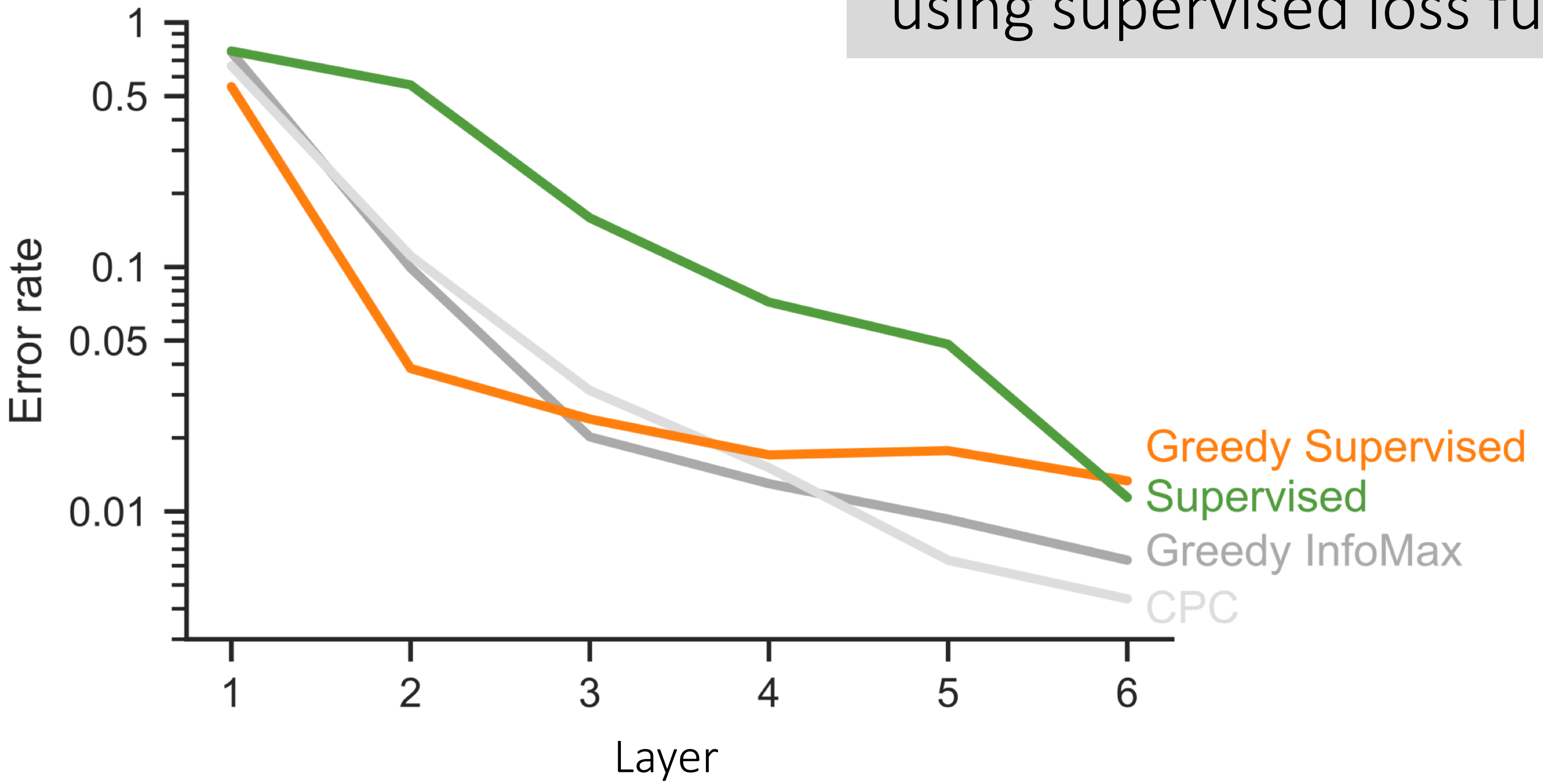
Measure speaker classification performance of intermediate layers



GIM and CPC achieve similar performance in intermediate layers



Performance gap for intermediate layers when using supervised loss function



We can train a neural network
without end-to-end backpropagation
and achieve competitive performance.

Thanks!

Sindy Löwe

sindy_loewe

Peter O’Connor

Bastiaan Veeling

BasVeeling