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

Зубанов Виктор

Задача

Нужно извлечь признаки из данных большой размерности(изображения, аудио).

Некоторые решения:

- Предобученные веса
- Генеративные модели
- Предсказывать по контексту

Contrastive Predicting Coding

x_t - sequential data

$z_t = g_{\text{enc}}(x_t)$ - encoder

$c_t = g_{\text{ar}}(z_{\leq t})$ - autoregressive model

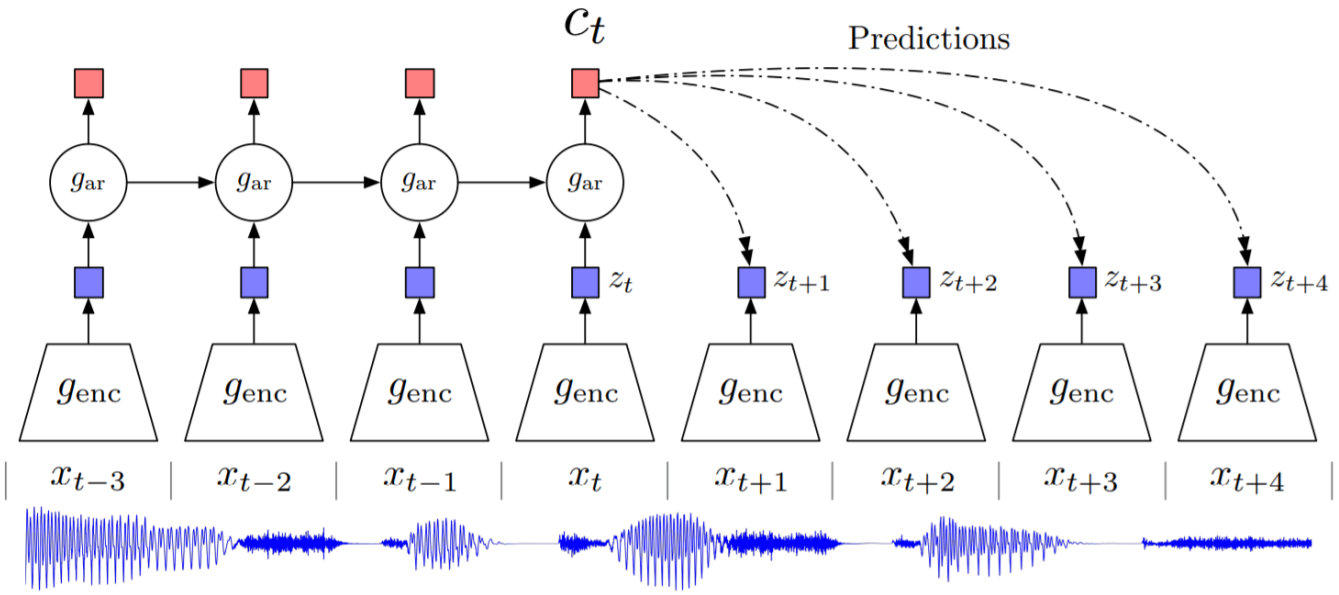


Figure 1: Overview of Contrastive Predictive Coding, the proposed representation learning approach. Although this figure shows audio as input, we use the same setup for images, text and reinforcement learning.

Contrastive Predicting Coding

x_t - sequential data

$z_t = g_{\text{enc}}(x_t)$ - encoder

$c_t = g_{\text{ar}}(z_{\leq t})$ - autoregressive model

Noise-Contrastive Estimation:

$\{z_{t+k}, z_{j_1}, z_{j_2}, \ldots z_{j_{N-1}}\}$

z_{t+k} - positive sample

$z_{j_1}, z_{j_2}, \ldots z_{j_{N-1}}$ - negative samples

Training data: pairs (z_j, c_t)

Scoring function: $f_k(z_j, c_t) = \exp\left(z_j^T W_k c_t\right)$

Contrastive Predicting Coding

x_t - sequential data

$z_t = g_{\text{enc}}(x_t)$ - encoder

$c_t = g_{\text{ar}}(z_{\leq t})$ - autoregressive model

Noise-Contrastive Estimation:

$\{z_{t+k}, z_{j_1}, z_{j_2}, \ldots z_{j_{N-1}}\}$

z_{t+k} - positive sample

$z_{j_1}, z_{j_2}, \ldots z_{j_{N-1}}$ - negative samples

Training data: pairs (z_j, c_t)

Scoring function: $f_k(z_j, c_t) = \exp\left(z_j^T W_k c_t\right)$

$$\mathcal{L}_N = -\sum_k \mathop{\mathbb{E}}_X \left[\log \frac{f_k(z_{t+k}, c_t)}{\sum_{z_j \in X} f_k(z_j, c_t)}\right]$$

Contrastive Predicting Coding

x_t - sequential data

$z_t = g_{\text{enc}}(x_t)$ - encoder

$c_t = g_{\text{ar}}(z_{\leq t})$ - autoregressive model

Можно показать, что оптимальное решение для f

$$f_k(z_{t+k}, c_t) \propto \frac{p(z_{t+k}|c_t)}{p(z_{t+k})}$$

Тогда loss является нижней оценкой mutual information:

$$I(x; c) = \sum_{x,c} p(x, c) \log \frac{p(x|c)}{p(x)}$$

Найденные признаки в теории должны быть slow features(примера(интонация голоса, объект на изображении, etc)

Contrastive Predicting Coding

После обучения функция f(..) отбрасывается.

В качестве признаков можно использовать z_t , c_t

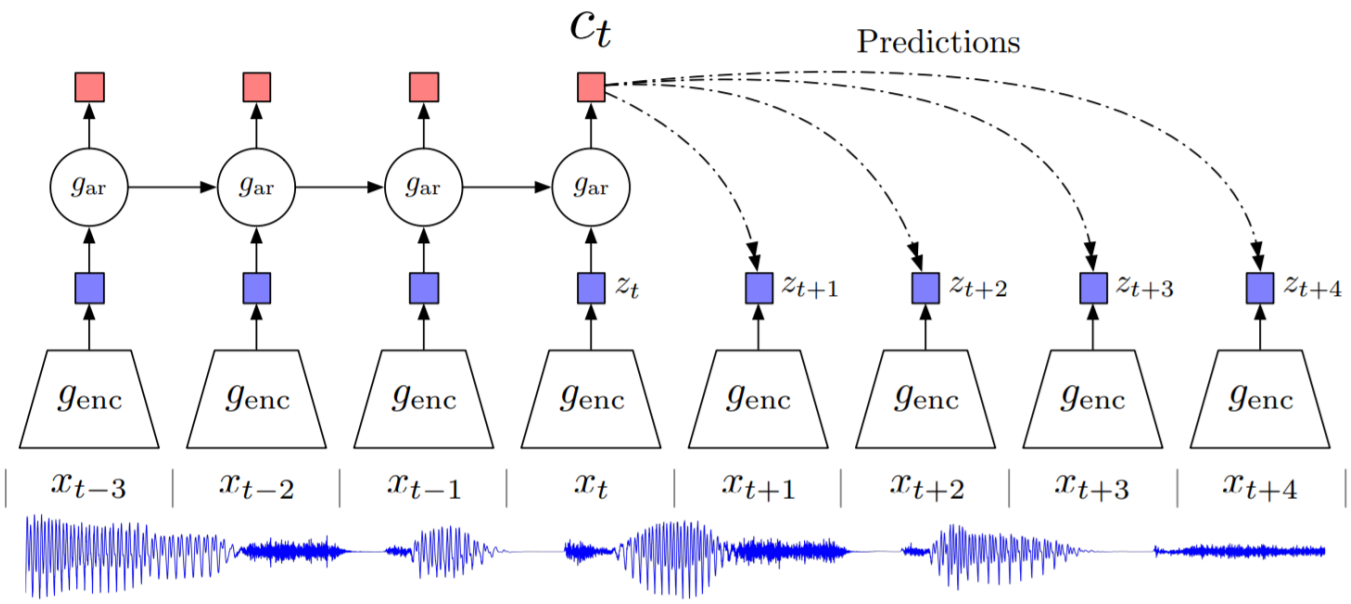
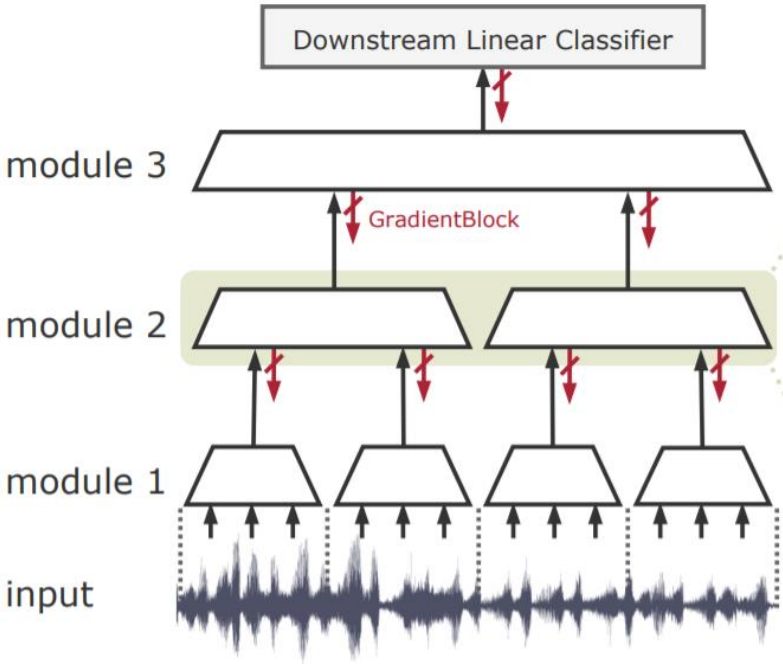
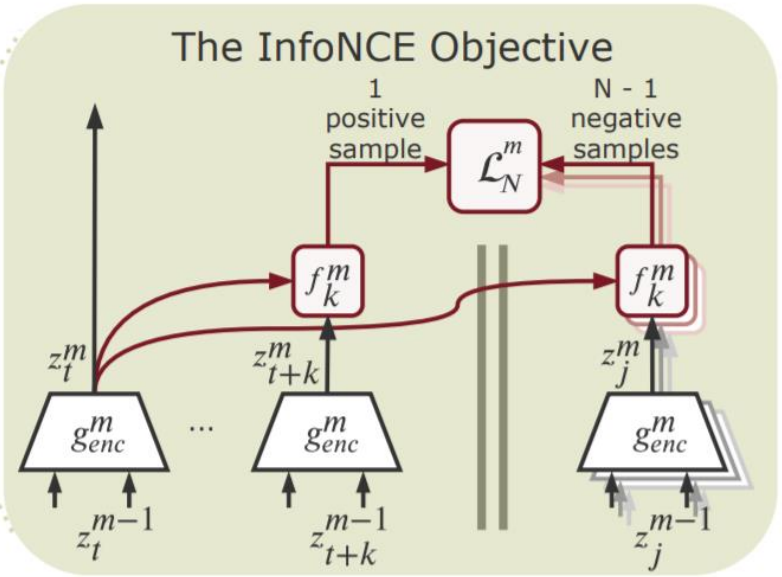


Figure 1: Overview of Contrastive Predictive Coding, the proposed representation learning approach. Although this figure shows audio as input, we use the same setup for images, text and reinforcement learning.

Greedy InfoMax

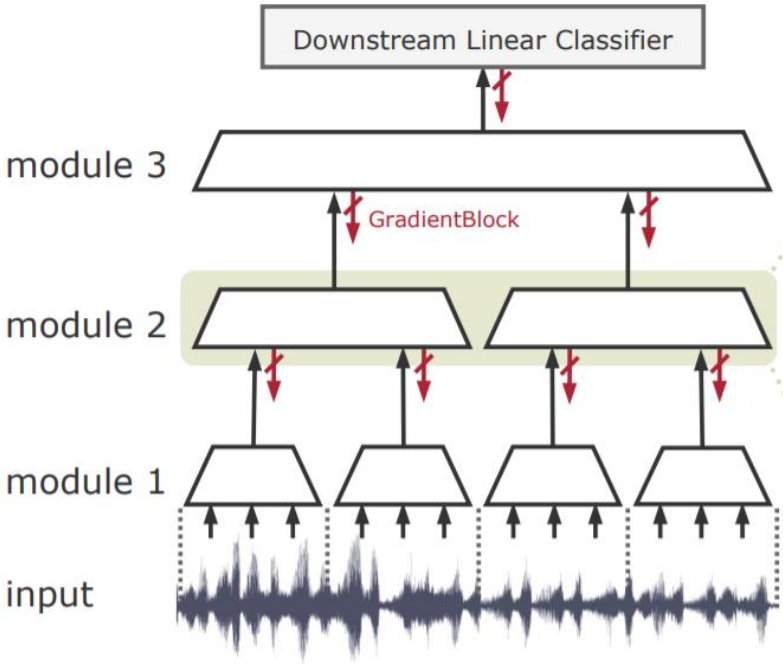


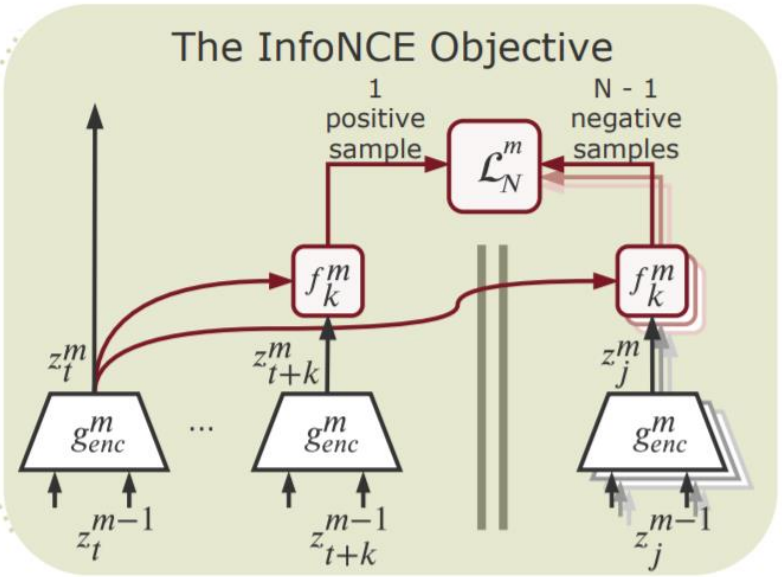


$$f_k^m(z_{t+k}^m, z_t^m) = \exp\left(z_{t+k}^{m\,T} W_k^m z_t^m\right)$$

$$\mathcal{L}_N^m = -\sum_k \mathop{\mathbb{E}}_X \left[\log \frac{f_k^m(z_{t+k}^m, z_t^m)}{\sum_{z_j^m \in X} f_k^m(z_j^m, z_t^m)}\right]$$

Greedy InfoMax



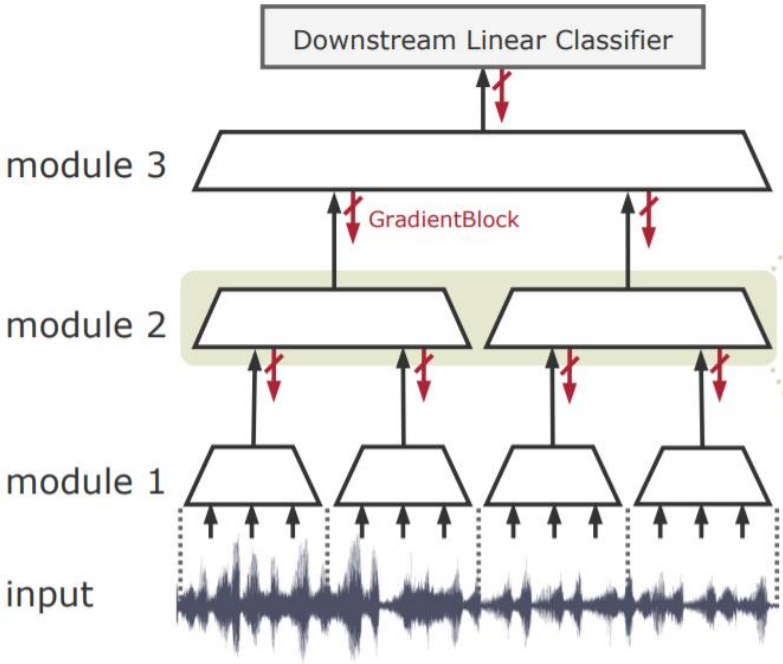


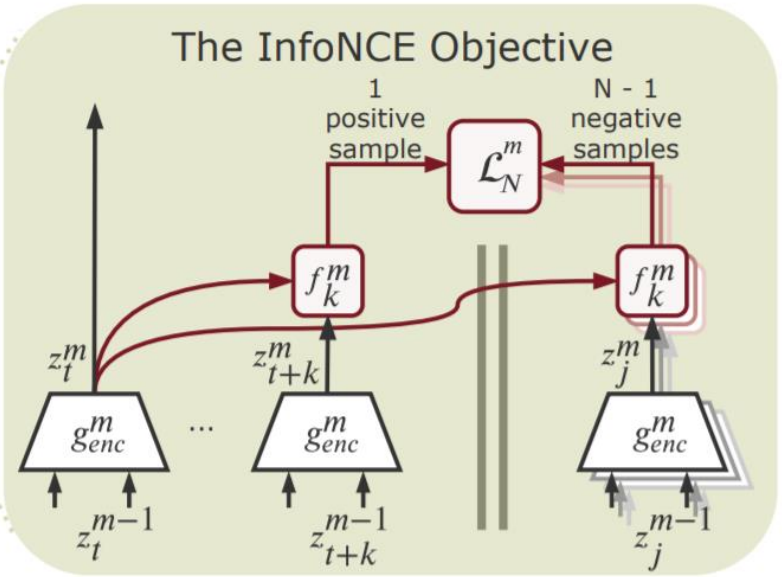
$$f_k^m(z_{t+k}^m, z_t^m) = \exp\left(z_{t+k}^{m\,T} W_k^m z_t^m\right)$$

$$\mathcal{L}_N^m = -\sum_k \mathop{\mathbb{E}}_X \left[\log \frac{f_k^m(z_{t+k}^m, z_t^m)}{\sum_{z_j^m \in X} f_k^m(z_j^m, z_t^m)}\right]$$

$$z_t^M = g_{enc}^M\left(g_{enc}^{M-1}\left(\cdots g_{enc}^1(x_t)\right)\right)$$

Greedy InfoMax





- Уменьшена проблема vanishing gradients
- Асинхронное обучение модулей

Table 2: GPU memory consumption during training. All models consist of the ResNet-50 architecture and only differ in their training approach. GIM allows efficient greedy training.

Method	GPU memory (GB)
Supervised	6.3
CPC	7.7
GIM - all modules	7.0
GIM - 1st module	**2.5**

Эксперименты: изображения

STL-10 dataset:

Изображения 256x256 нарезаются на гриды.

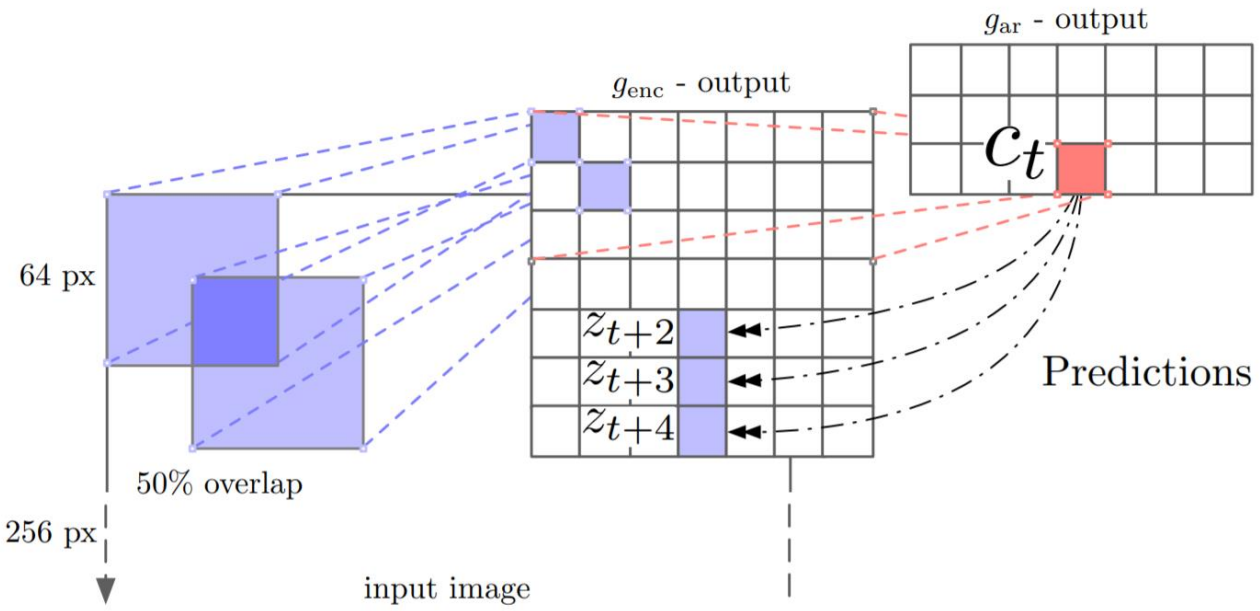


Figure 4: Visualization of Contrastive Predictive Coding for images (2D adaptation of Figure 1).

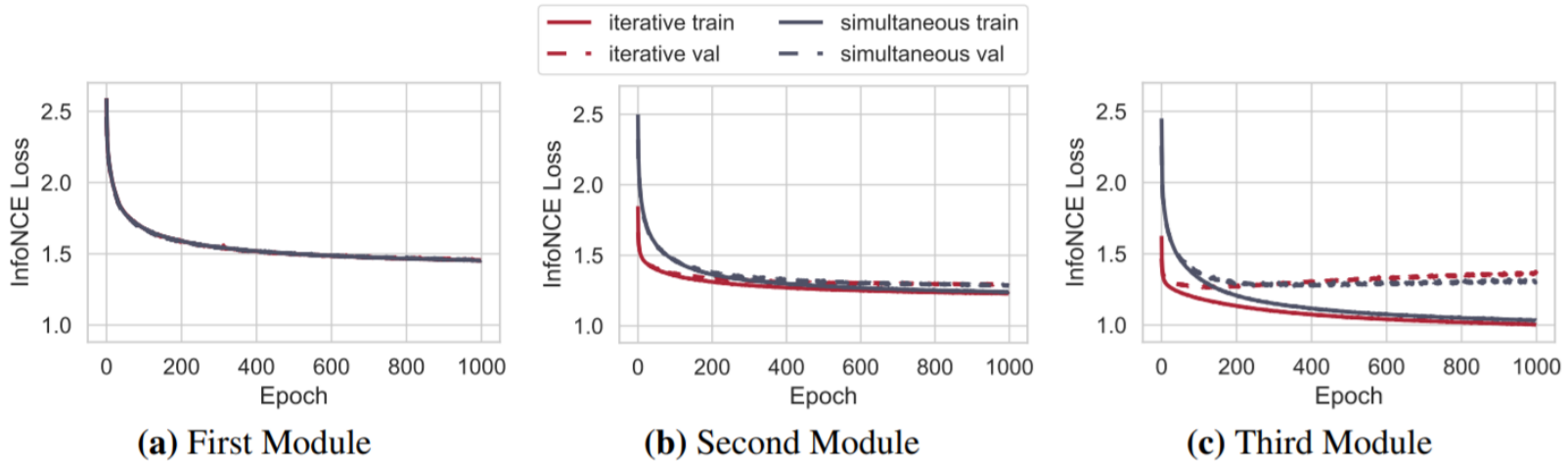
Эксперименты: изображения

Table 1: STL-10 classification results on the test set. The GIM model outperforms the CPC model, despite a lack of end-to-end backpropagation and without the use of a global objective. ($\pm$ standard deviation over 4 training runs.)

Method	Accuracy (%)
Deep InfoMax [Hjelm et al., 2019]	78.2
Predsim [Nøkland and Eidnes, 2019]	80.8
Randomly initialized	27.0
Supervised	71.4
Greedy Supervised	65.2
CPC	80.5 ± 3.1
Greedy InfoMax (GIM)	$\mathbf{81.9} \pm 0.3$

Эксперименты: асинхронное обучение

При последовательном обучении модулей качество классификации изображений снизилось с 81.9% до 79.8%.



(a) First Module **(b)** Second Module **(c)** Third Module

Figure 3: Training curves for optimizing all modules *simultaneously* (blue) or *iteratively*, one at a time (red). While there is no difference in the training methods for the first module (**a**), later modules (**b, c**) start out with a lower loss and tend to overfit more when trained iteratively on top of already converged modules.

Эксперименты: аудио

Table 3: Results for classifying speaker identity and phone labels in the LibriSpeech dataset. All models use the same audio input sizes and the same architecture. Greedy InfoMax creates representations that are useful for audio classification tasks despite its greedy training and lack of a global objective.

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

[a] In the original implementation, Oord et al. [2018] achieved 64.6% for the phone and 97.4% for the speaker classification task. [b] Baseline results from Oord et al. [2018].

Phone classification: 41 class
Speaker classification: 251 class
Dataset size: 100 hours

Вопросы

1) Приведите любой метод получения признаков из неразмеченных данных, и объясните его недостаток.
2) Как происходит обучение в модели Contrastive Predictive Coding?
3) Как возможно обучение модели Greedy Info Max без глобальной функции потерь?

References

https://arxiv.org/pdf/1905.11786.pdf

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

https://arxiv.org/pdf/1807.03748.pdf

Representation Learning with Contrastive Predictive Coding