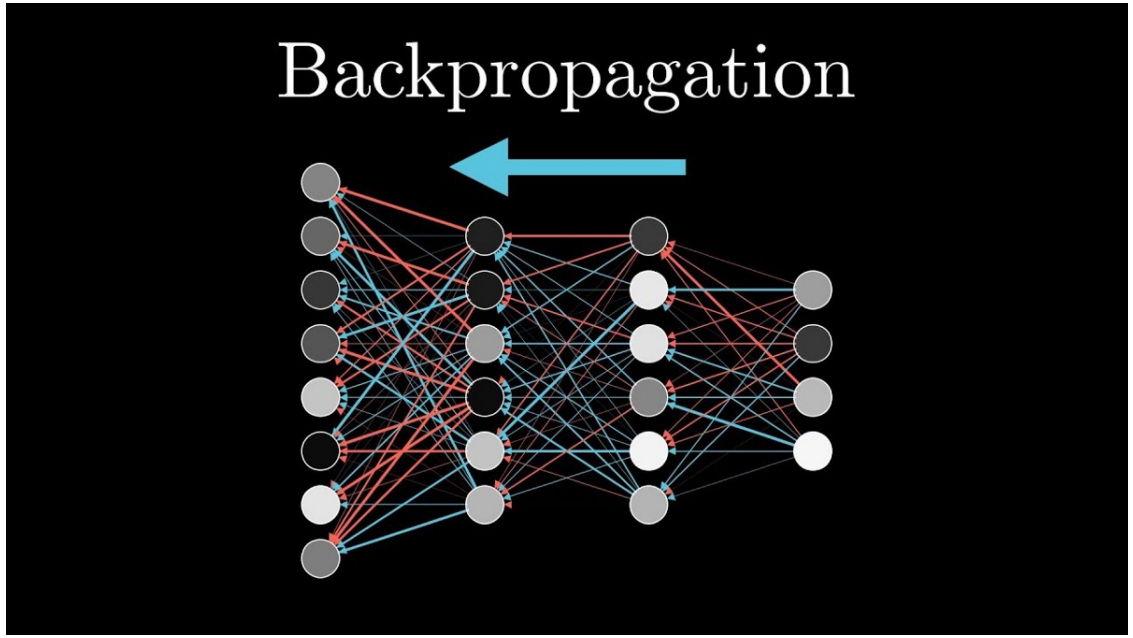
Putting An End to End-to-End: Gradient-Isolated Learning of Representations

Анищенко Илья

Основной подход к обучению: backpropagation и глобальная функция потерь



Проблемы, с которыми этот метод сталкивается:

При supervised обучении нужно много размеченных данных (размечать дорого и долго)

Все объекты оптимизации: веса, активации, градиенты, сама модель со слоями, - должны вмещаться в единую память GPU

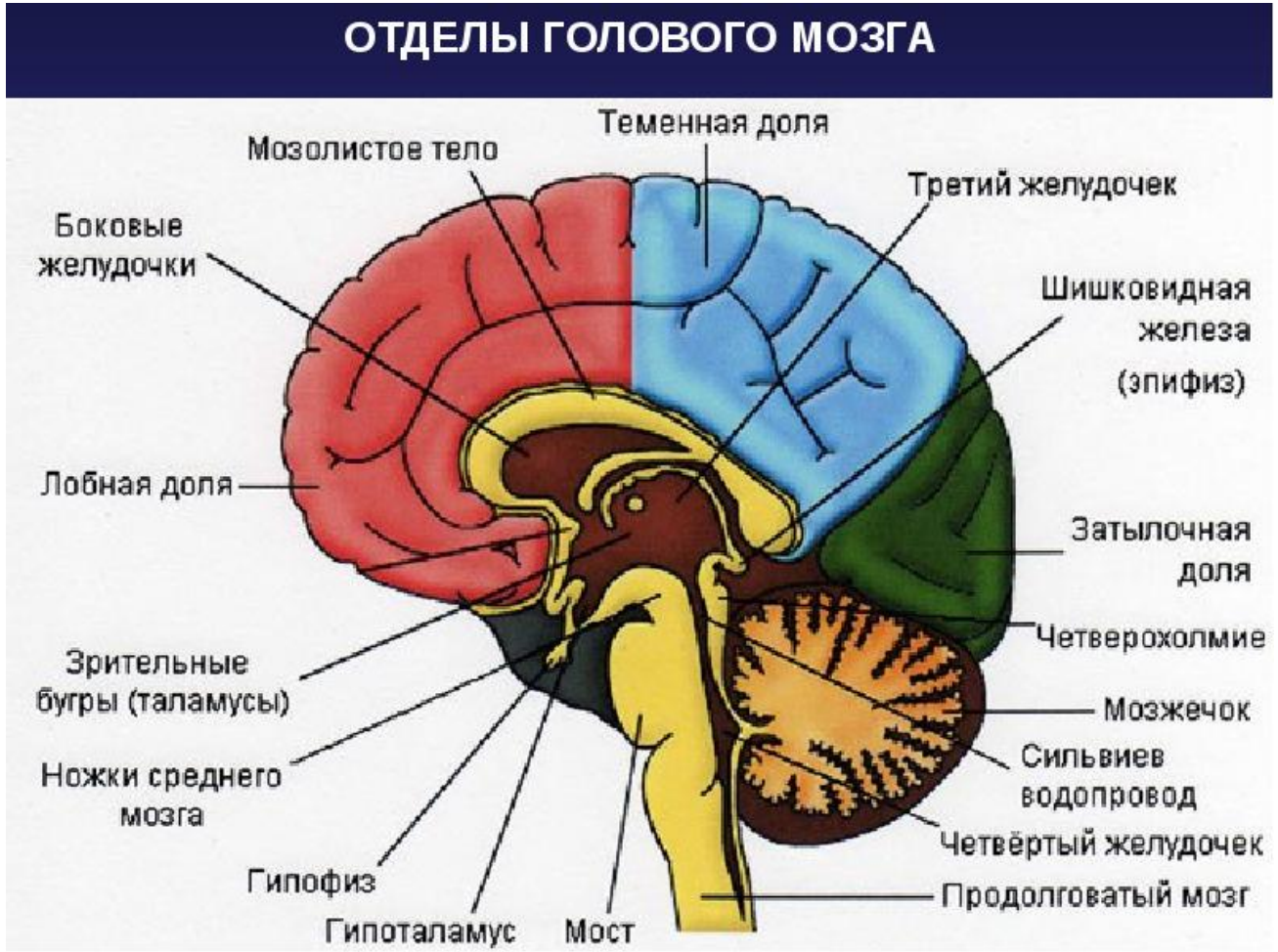
Проблема исчезающего градиента

Нет возможности асинхронного обучения слоев, т.к. либо они ждут свои входы, либо градиент

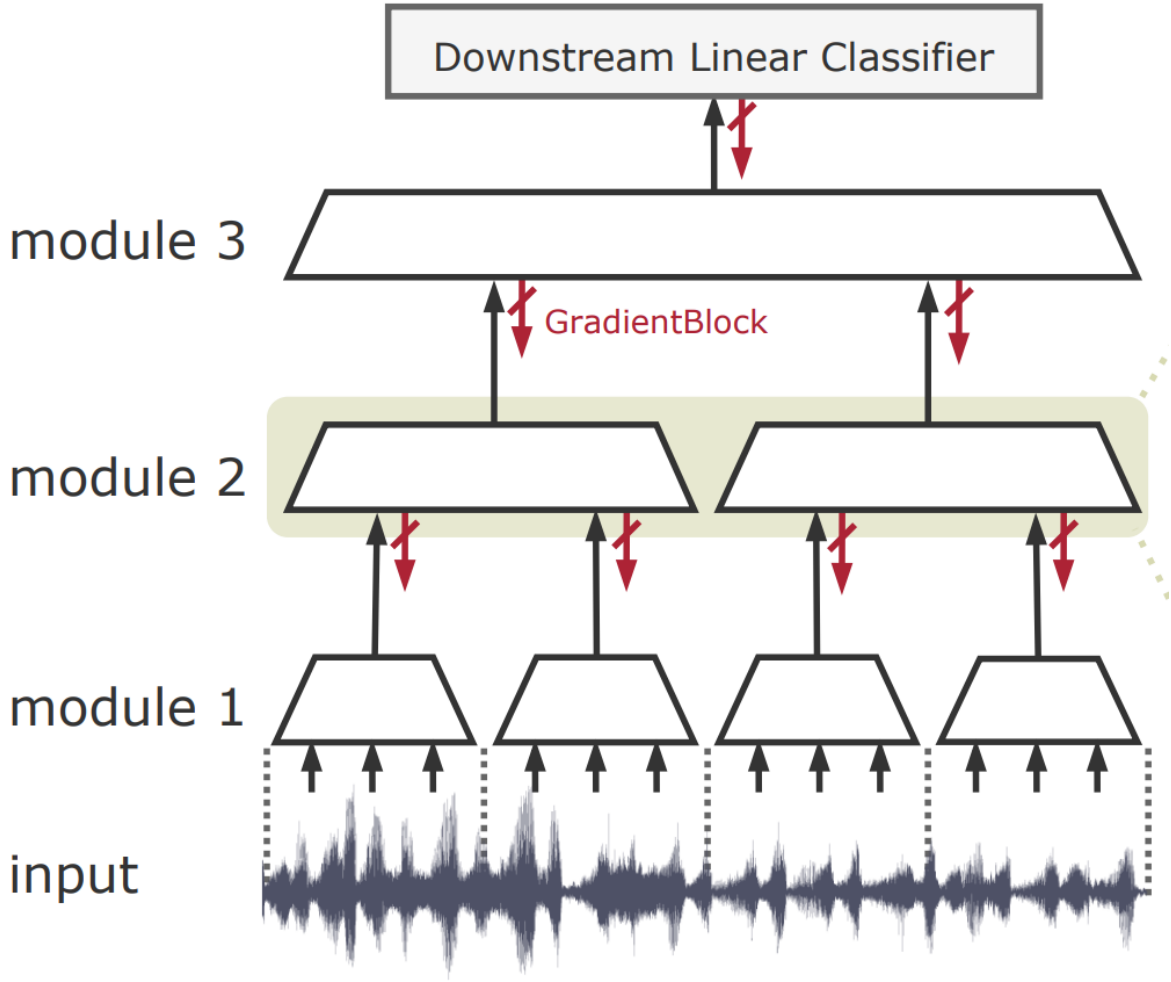
Ассоциации с работой мозга

Мозг человека обладает высокой модульностью и обучается на основе локальной информации.

Дети могут научиться распознавать новую категорию на основе нескольких образцов.

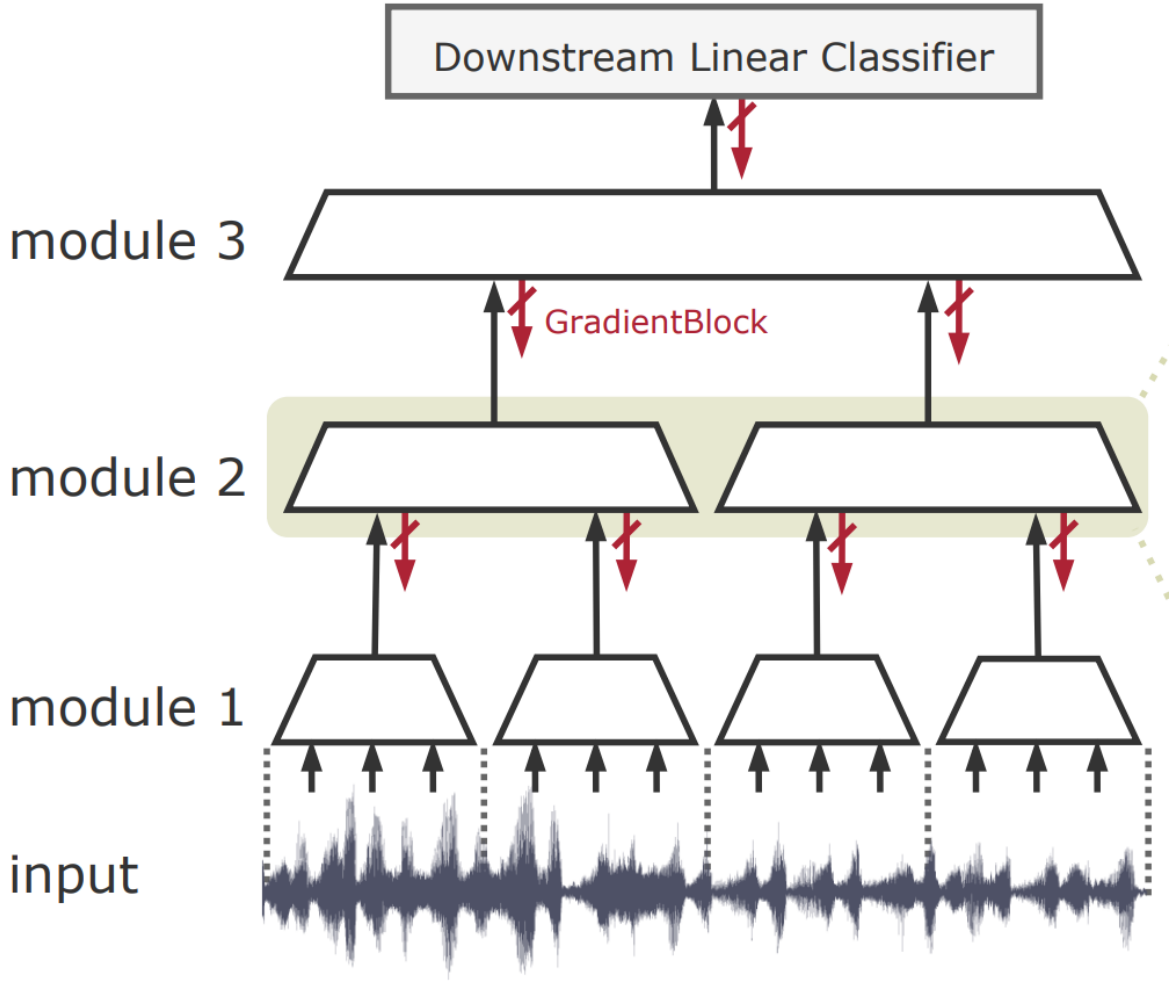


Предложенный метод:
Greedy InfoMax (GIM)



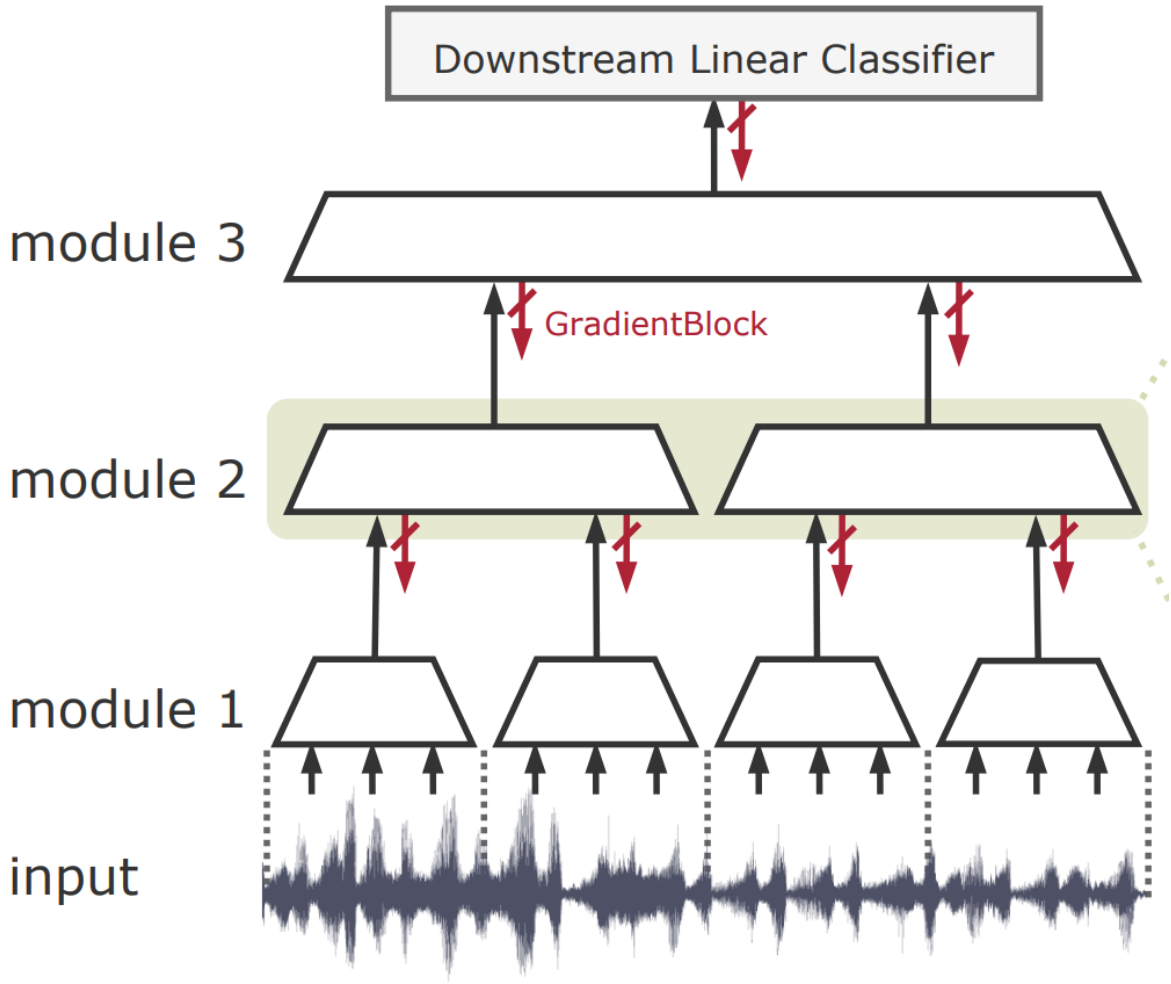
- Предложенный жадный алгоритм InfoMax обеспечивает высокую производительность при выполнении задач классификации аудио и изображений, несмотря на жадное самоконтролируемое обучение.

Предложенный метод:
Greedy InfoMax (GIM)



- Это обеспечивает асинхронное, несвязанное обучение нейронных сетей, позволяя обучать произвольно глубокие сети на входных данных, превышающих объем памяти

Предложенный метод:
Greedy InfoMax (GIM)



- Мы показываем, что взаимная максимизация информации особенно подходит для послойной жадной оптимизации, и утверждаем, что это уменьшает проблему исчезающих градиентов.

База авторского метода

Понимание объекта как конструкции из патчей:

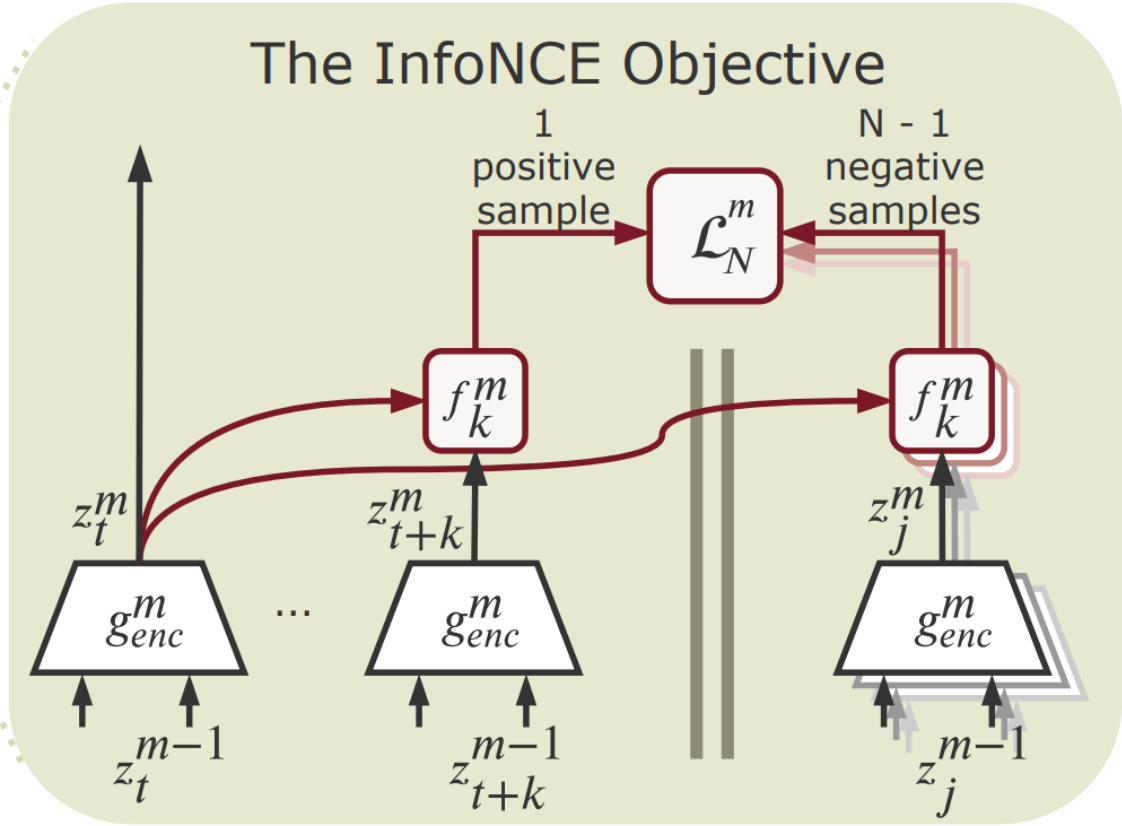
- Соседние части объекта имеют много общих аспектов, информации
- Далекие объекты необязательно при этом имеют общую информацию

База авторского метода

Contrastive Predictive Coding (CPC) – метод для изучения представлений, которые максимизируют взаимную инфу, разделяемую между соседями

$g_{enc}(x_t) = z_t$ Обработка входного сигнала

$g_{ar}(z_{0:t}) = c_t$ Агрегация информации от шага 0 до t



По полученным данным хотим отследить связь информации между z_{t+k} и c_t

База авторского метода

Используемая локальная функция потерь - Noise Contrastive Estimation (NCE):

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

Набор из одного «положительного» эл-та (кодируемый сигнал через k шагов) и N-1 «отрицательного» эл-та (берутся равномерно из всех имеющихся закодированных входных сигналов)

$$f_k(z_j, c_t) = \exp\left(z_j^T W_k c_t\right)$$ Вариант используемой ф-ии f

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

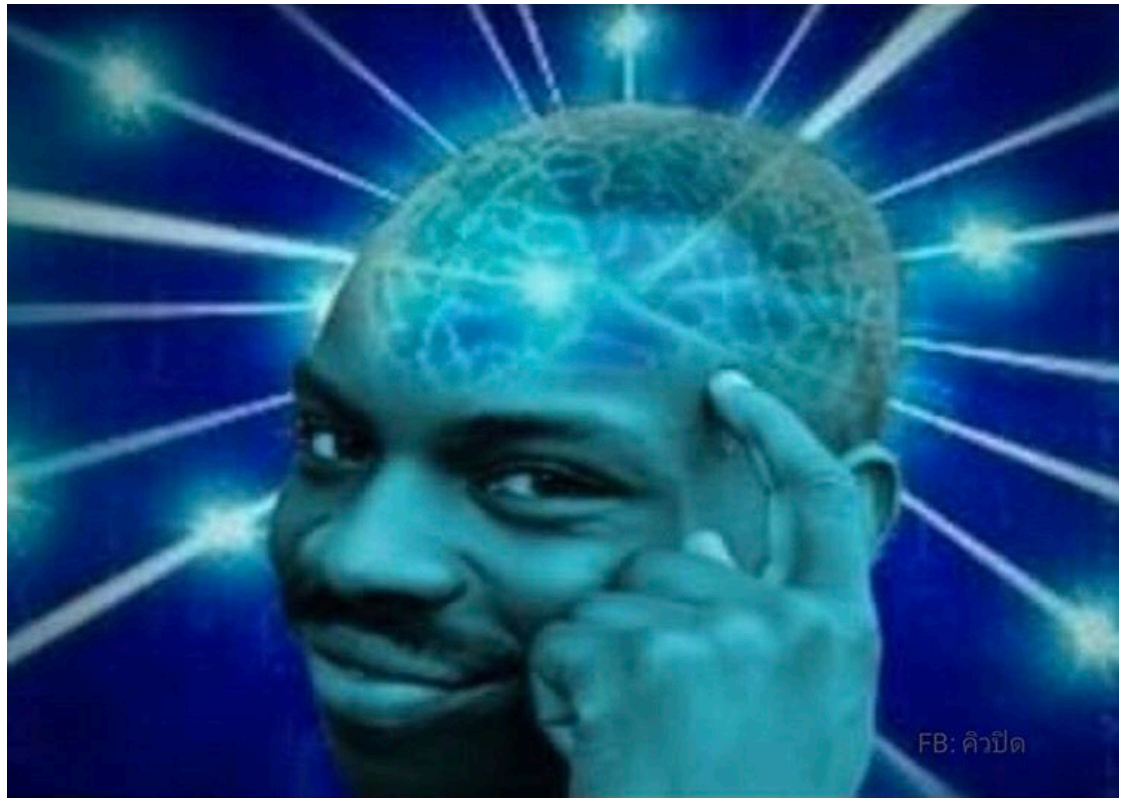
База авторского метода

Аналитическое оптимальное f: $$f_k(z_{t+k}, c_t) \propto \frac{p(z_{t+k}|c_t)}{p(z_{t+k})}.$$

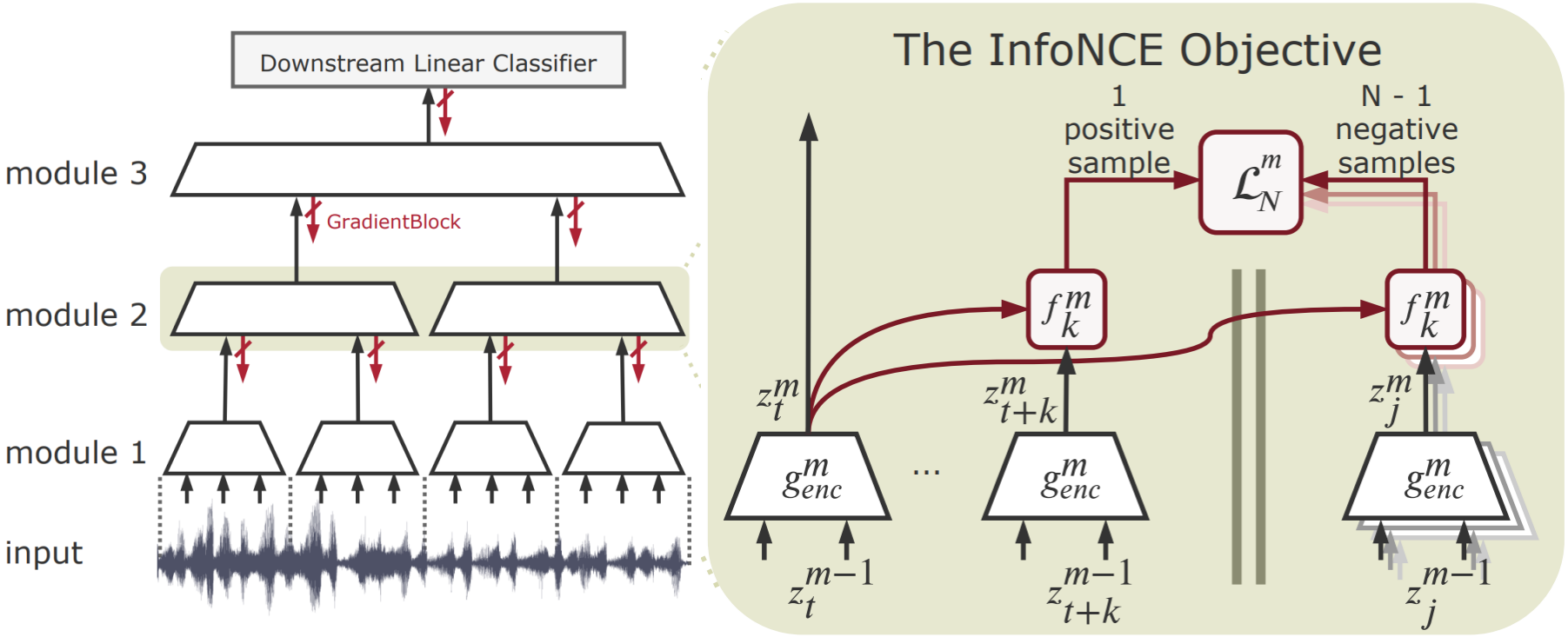
На основе этого можно сформулировать, что L_N это нижняя граница $I(z_{t+k}, c_t)$

Это дает ограничение на взаимную информацию $I(x_{t+k}, c_t)$

Концепция этих принципов взята из нейробиологии (Linsker [1988])



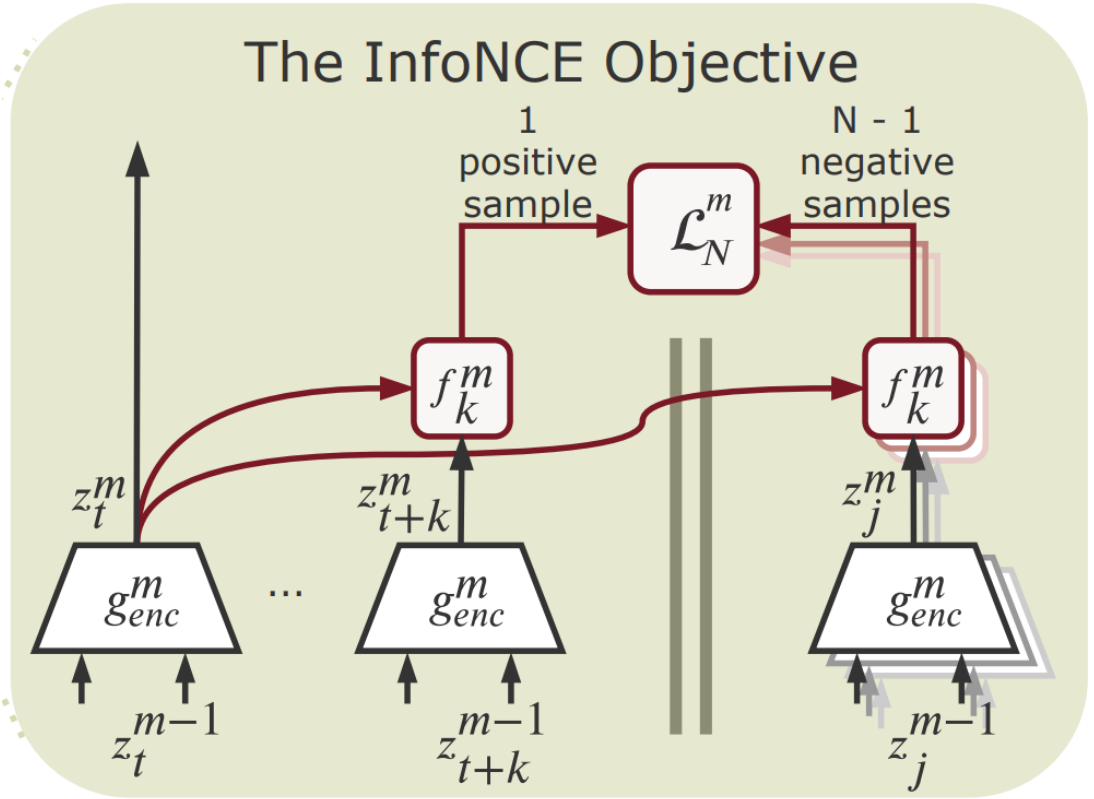
Greedy InfoMax



Деление глубинной сетки на модули:

- На уровне каждого отдельного слоя
- На уровне блоков (несколько сверток, к примеру)
- Выходной слой для задачи классификации

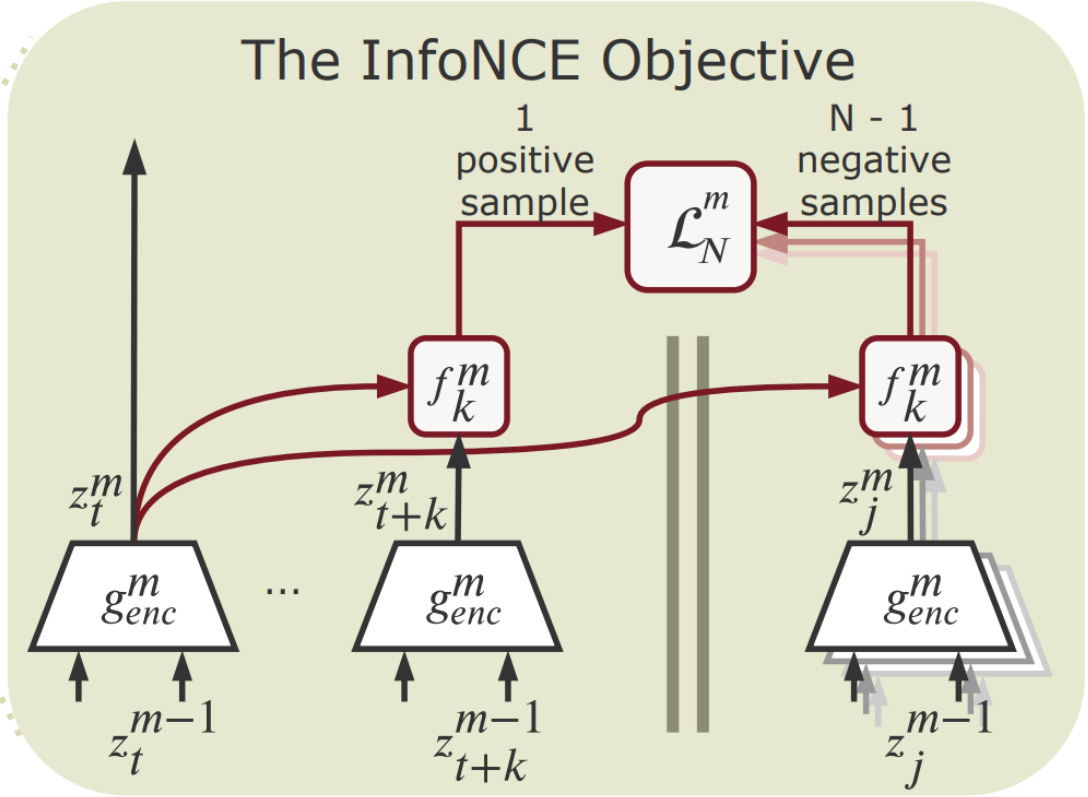
Greedy InfoMax



$$z_t^m = g_{enc}^m(\text{GradientBlock}(z_t^{m-1}))$$ Кодировка данных с предыдущего модуля текущим слоем

Связь между z_{t+k} и c_t В концепции GIM не смотрится, так как авторы в экспериментах не получили с таким отслеживанием информации хорошие результаты.

Greedy InfoMax

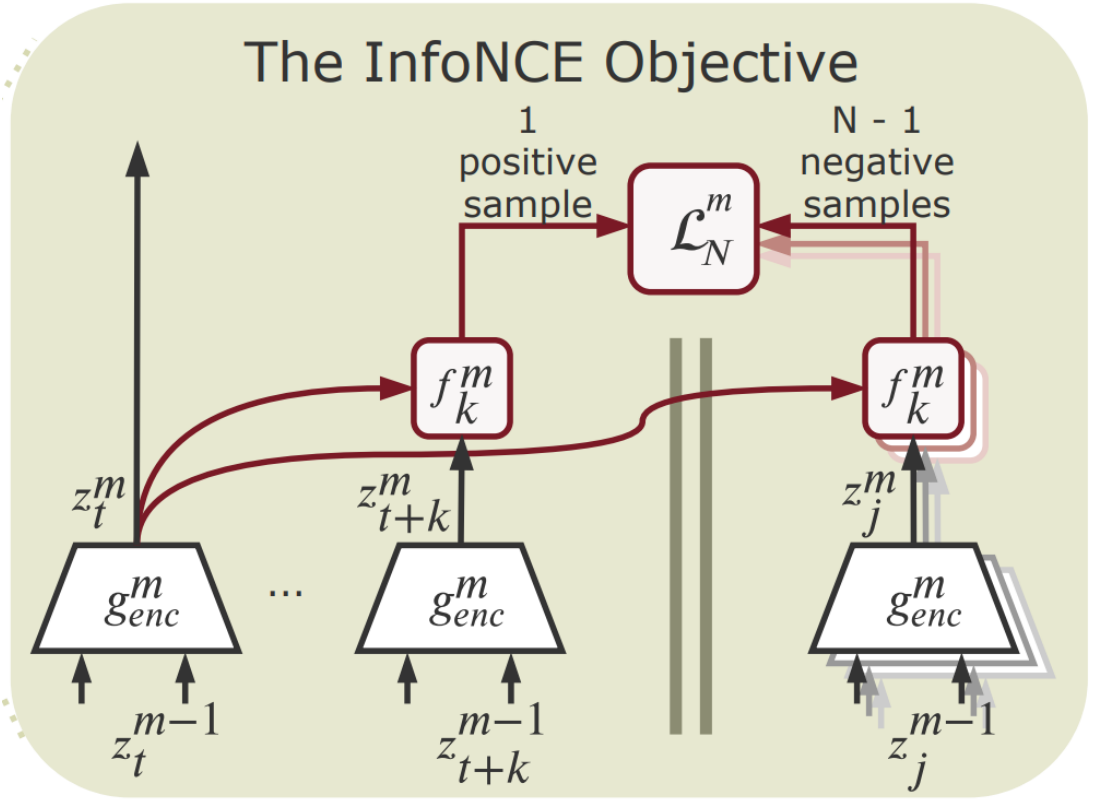


Поэтому функция обучения и локальной лосс в методе авторов претерпевают след. изменения:

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

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

Greedy InfoMax



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

Фактический результат работы моделей после сходимости всех модулей

$$c_t^M = g_{ar}^M\left(\text{GradientBlock}\left(z_{0:t}^{M-1}\right)\right)$$

Вводим авторег. Блок для формирования контекста под наши задачи

$$f_k^M(z_{t+k}^{M-1}, c_t^M) = \exp\left(\text{GradientBlock}\left(z_{t+k}^{M-1}\right)^T W_k^M c_t^M\right)$$

Функция обучения авторег. блока

Greedy InfoMax

Получившаяся локальная функция потерь – тоже нижняя граница $I(z_{t+k}^m, z_t^m)$

Также важно, что InfoNCE loss максимизирует нижнюю границу $I(z_{t+k}^{m-1}, z_t^m)$

Практические выводы:

- Последовательная оптимизация модулей (не держим в памяти всю сетку)
- Обучать модель на очень больших данных (на начальных слоях данные будут «сжиматься»)
- Обучать модули сетки с разной частотой, обучать их асинхронно
- Решение проблемы с исчезанием градиента

Эксперименты

Классификация изображений

Сеть обучается на STL-10 (неразмеченные данные), после получения каких-то паттернов на конечном слое, сеть замораживается и отдельно дообучается линейный классификатор

Работа с картинками:
Изображения 96 x 96 обрезаются до 64 x 64 с разными аугментациями, и представляют себя набор перекрещивающихся патчей размера 7x7. Каждый патч размера 16 x 16

За основу берется ResNet-50 v2 без батчнорма и делится на 3 отдельных модуля.

Эксперименты

Результаты:

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.)

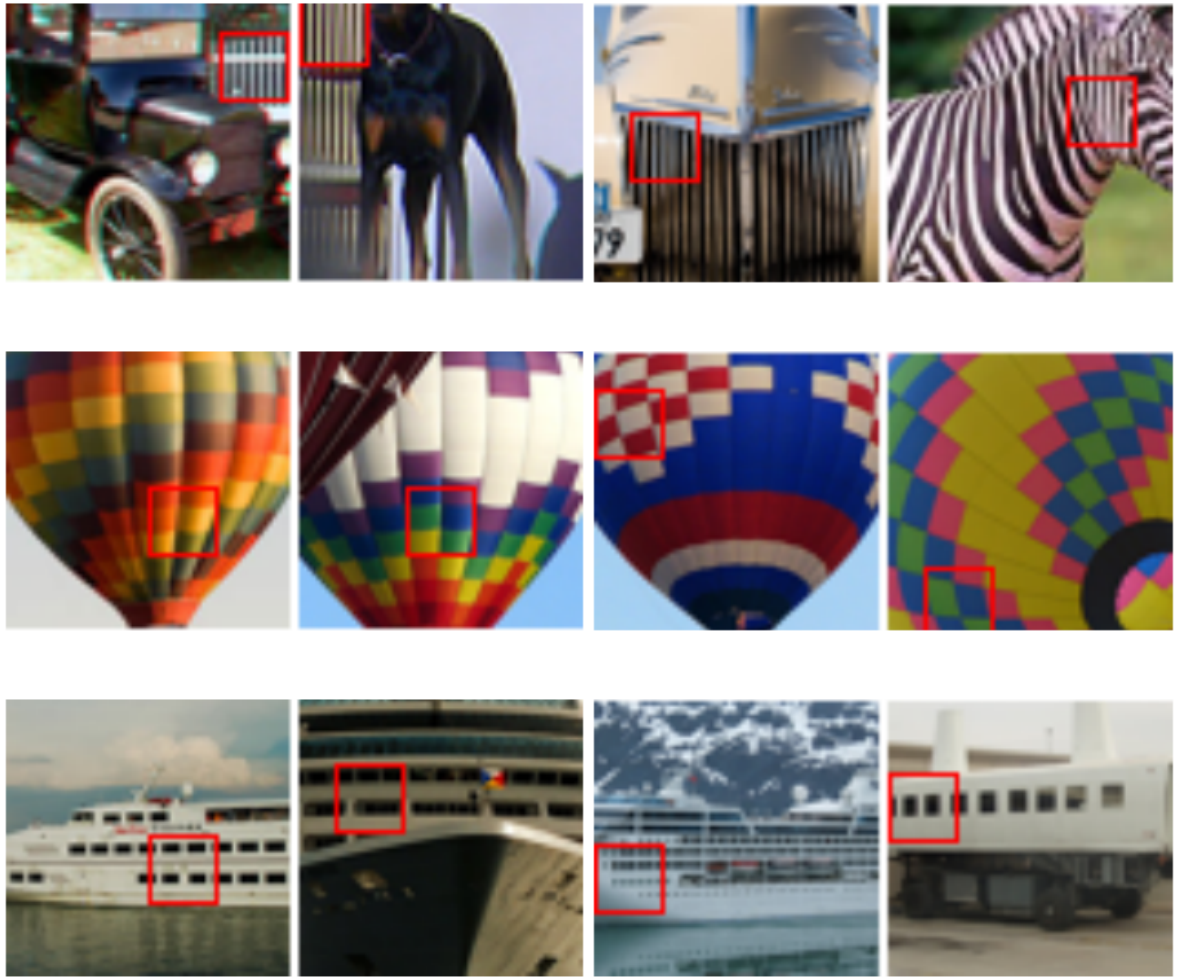
Method	Accuracy (%)
Deep InfoMax [Hjelm et al., 2019]	78.2
Predsim [Nøkland and Eidnes, 2019]	80.8
Randomly initialized	27.0
Supervised	71.4
Greedy Supervised	65.2
CPC	80.5 ± 3.1
Greedy InfoMax (GIM)	$\mathbf{81.9} \pm 0.3$

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**

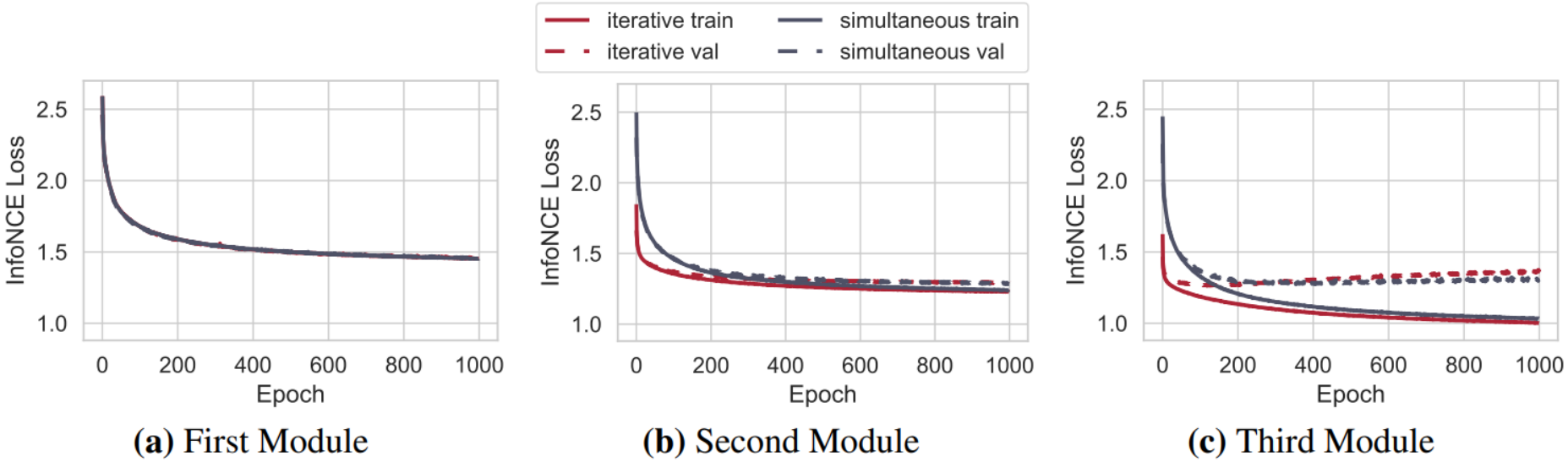
Эксперименты

Интерпретация:



Группы из 4 патчей изображения, которые возбуждают определенный нейрон, на 3 уровнях модели.
Несмотря на неконтролируемое жадное обучение, нейроны, по-видимому, извлекают все больше семантических признаков.

Эксперименты



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

Сравнение результаты работы 3 модулей сети при последовательном и асинхронном обучении

Эксперименты

Аудио: 100 часовой датасет LibriSpeech

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].

Выводы:

- Авторами был представлен новый метод обучения сетей, в перспективе имеющий очень полезные практические значения.
- Сравнительная производительность с другими методами показывает, что глубокие сетки не обязательно требуют backprop метода.
- Их метод позволят проводить обучение модели с жадной оптимизацией на неразмеченных данных, что помогает в борьбе с переобучением и исчезающим градиентом
- Дает возможность проводить асинхронное обучение отдельных частей модели.
- Такой подход к обучению имеет больше общего с работой человеческого мозга

Вопросы:

С какими проблемами можно столкнуться при обучении нейронных моделей с backprop? Какую альтернативу этому методу предлагают авторы статьи?

Как авторы создают компактное представление данных?

Опишите все практические преимущества обучения с модели с жадным InfoMax, которые мы можем получить по мнению авторов.