

Momentum ResNet

Ким Михаил
Станкевич Матвей
Каратаева Екатерина
Чураков Игорь

Постановка проблемы

- Память необходимая для обучения (forward-backward pass):
 $O(\mathbf{k} \times d \times n_batch)$, где k – глубина сети, d – размерность признакового пространства, n_batch – размер батча.
- Часто используют глубокие сети: $\mathbf{k} > 100$

Постановка проблемы

- Память необходимая для обучения (forward-backward pass):
 $O(\mathbf{k} \times d \times n_batch)$, где k – глубина сети, d – размерность признакового пространства, n_batch – размер батча.
- Часто используют глубокие сети: $\mathbf{k} > 100$

Хотим уменьшить зависимость необходимой памяти от \mathbf{k} !

Возможные решения

- Хранить не все активации, а только некоторые, остальные – пересчитывать
- Обратимые (reversible) модели: RevNet, i-RevNet
- Neural ODE

Возможные решения: Обратимые модели

- Строим такую архитектуру модели, что предыдущую активацию можно пересчитать через последующие
- Не всегда понятно, как получить обратимую модель

Возможные решения: Обратимые модели

ResNet:

$$y = x + \mathcal{F}(x)$$

RevNet:

$$y_1 = x_1 + \mathcal{F}(x_2)$$

$$y_2 = x_2 + \mathcal{G}(y_1)$$

$$x_2 = y_2 - \mathcal{G}(y_1)$$

$$x_1 = y_1 - \mathcal{F}(x_2)$$

forward

reverse

x_1, x_2 – два слайса (по глубине) x

Возможные решения: Neural ODE

Если представить, что сеть содержит бесконечное количество слоёв, то получим динамику скрытых состояний:

$$\mathbf{h}_{t+1} = \mathbf{h}_t + f(\mathbf{h}_t, \theta_t) \quad \longrightarrow \quad \frac{d\mathbf{h}(t)}{dt} = f(\mathbf{h}(t), t, \theta)$$

Дискретный случай

Непрерывный случай

Возможные решения: Neural ODE

- Начинаем с $\mathbf{h}(\mathbf{0}) = \mathbf{x}$ (входных данных), заканчиваем $\mathbf{h}(\mathbf{T}) = \mathbf{y}$ (например, выход логиты)
- $\mathbf{h}(\mathbf{T})$ – решение дифференциального уравнения
- Считаем функцию потерь:

$$L(\mathbf{z}(t_1)) = L \left(\mathbf{z}(t_0) + \int_{t_0}^{t_1} f(\mathbf{z}(t), t, \theta) dt \right) = L(\text{ODESolve}(\mathbf{z}(t_0), f, t_0, t_1, \theta))$$

- Считаем градиенты с помощью метода сопряжённых состояний:
 1. По сути, сводится к решению другого ДУ
 2. Константное потребление памяти

Возможные решения: Проблемы

- Не всегда понятно, как существующую необратимую модель сделать обратимой
- Пересчёт предыдущей активации может быть не очень простым
- От точности численного решения ДУ зависит качество модели

Предлагаемый метод: Momentum ResNet

- Меняем forward следующим образом:

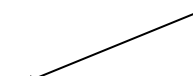
$$\begin{cases} v_{n+1} = \gamma v_n + (1 - \gamma)f(x_n, \theta_n) \\ x_{n+1} = x_n + v_{n+1}, \end{cases}$$

- Во время backward:

$$\begin{cases} x_n = x_{n+1} - v_{n+1}, \\ v_n = \frac{1}{\gamma} (v_{n+1} - (1 - \gamma)f(x_n, \theta_n)) \end{cases}$$

Предлагаемый метод: Momentum ResNet

Потеря информации

$$\begin{cases} v_{n+1} = \gamma v_n + (1 - \gamma) f(x_n, \theta_n) \\ x_{n+1} = x_n + v_{n+1}, \end{cases}$$


Algorithm 3 Exactly reversible multiplication by a ratio

- 1: **Input:** Information buffer i , value c , ratio n/d
 - 2: $i = i \times d$ ▷ make room for new digit
 - 3: $i = i + (c \bmod d)$ ▷ store digit lost by division
 - 4: $c = c \div d$ ▷ divide by denominator
 - 5: $c = c \times n$ ▷ multiply by numerator
 - 6: $c = c + (i \bmod n)$ ▷ add digit from buffer
 - 7: $i = i \div n$ ▷ shorten information buffer
 - 8: **return** updated buffer i , updated value c
-

$$\gamma = n/d$$

Предлагаемый метод: Momentum ResNet

- При $\gamma < 1$ и глубине k для хранения буфера для одного элемента выхода нужно $k \log(1/\gamma)$
- При γ близкой к 1: $k(1 - \gamma) / \ln(2)$

Пример:

$$f(x) = W_2 \operatorname{sigmoid}(W_1 x + b)$$

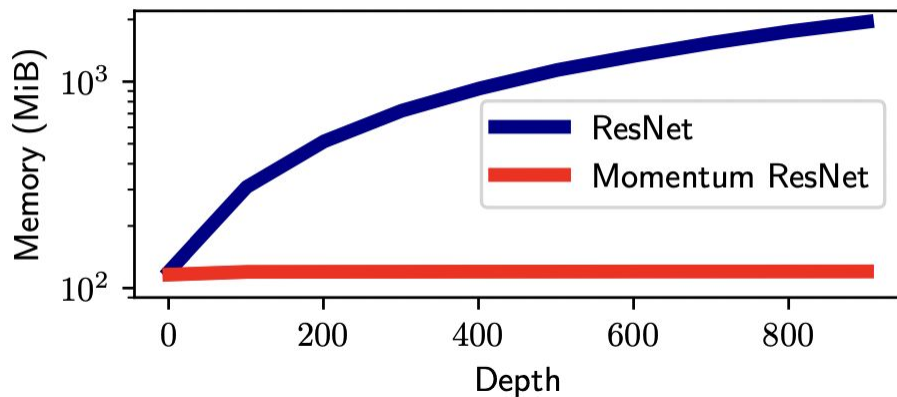
$d \times p \qquad p \times d \quad d \times 1 \quad p \times 1$

ResNet: $O(k \times d \times n_{\text{batch}})$

Momentum ResNet: $O((1 - \gamma) \times k \times d \times n_{\text{batch}})$

Предлагаемый метод: свойства

- При $\gamma < 1$ и глубине k для хранения буфера для одного элемента выхода нужно $k \log(1/\gamma)$
- При γ близкой к 1: $k(1 - \gamma) / \ln(2)$



Предлагаемый метод: свойства

$$\begin{cases} v_{n+1} = \gamma v_n + (1 - \gamma)f(x_n, \theta_n) \\ x_{n+1} = x_n + v_{n+1}, \end{cases}$$

$$\downarrow \frac{1}{1-\gamma} = \varepsilon$$

$$v_{n+1} = v_n + \frac{f(x_n, \theta_n) - v_n}{\varepsilon}, \quad x_{n+1} = x_n + v_{n+1}$$

Предлагаемый метод: свойства

$$v_{n+1} = v_n + \frac{f(x_n, \theta_n) - v_n}{\varepsilon}, \quad x_{n+1} = x_n + v_{n+1}$$



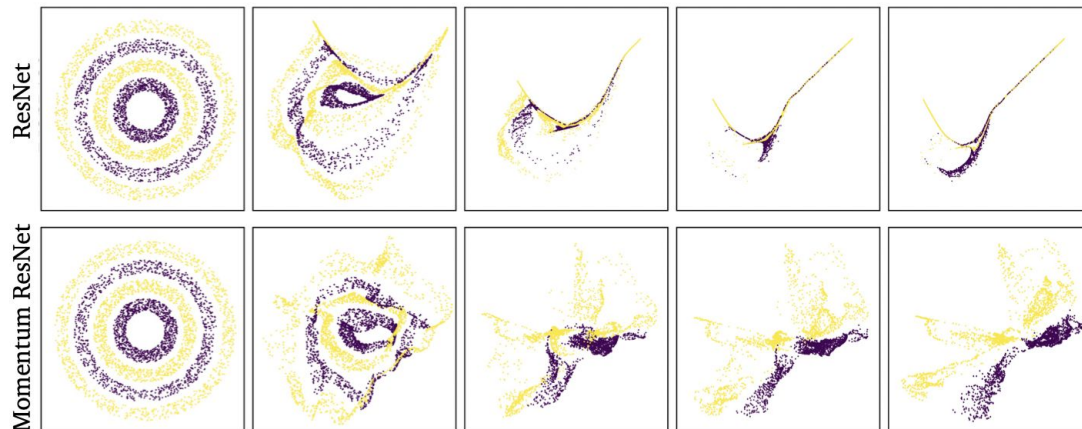
$$\varepsilon \ddot{x} + \dot{x} = f(x, \theta) \quad \text{with} \quad (x(0), \dot{x}(0)) = (x_0, v_0)$$

ОДУ второго порядка

- Можно показать, что ОДУ первого порядка не универсальные аппроксиматоры (Секция 4.1)
- Выразительность ОДУ второго порядка растёт с ростом ε (Секция 4.2 Предложение 3)

$$\frac{1}{1-\gamma} = \varepsilon.$$

Эксперименты: сегментация 2D облака

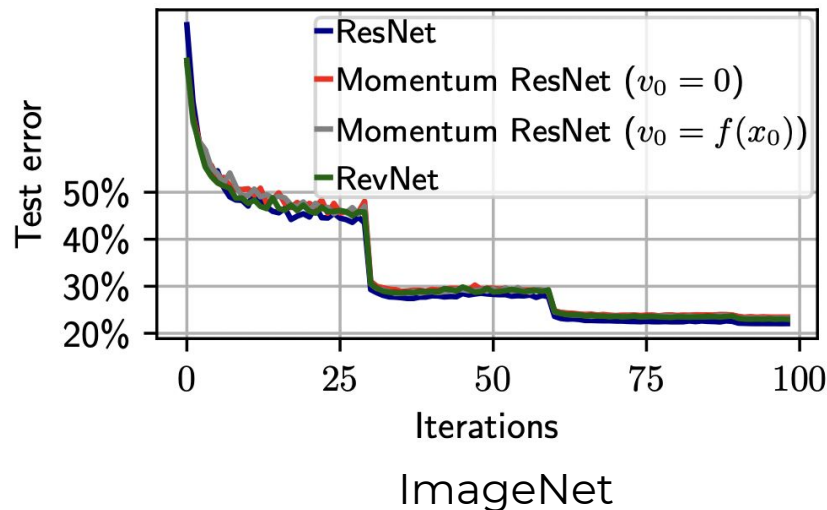


Визуализация трансформации облака точек
слоем с номером 3k.

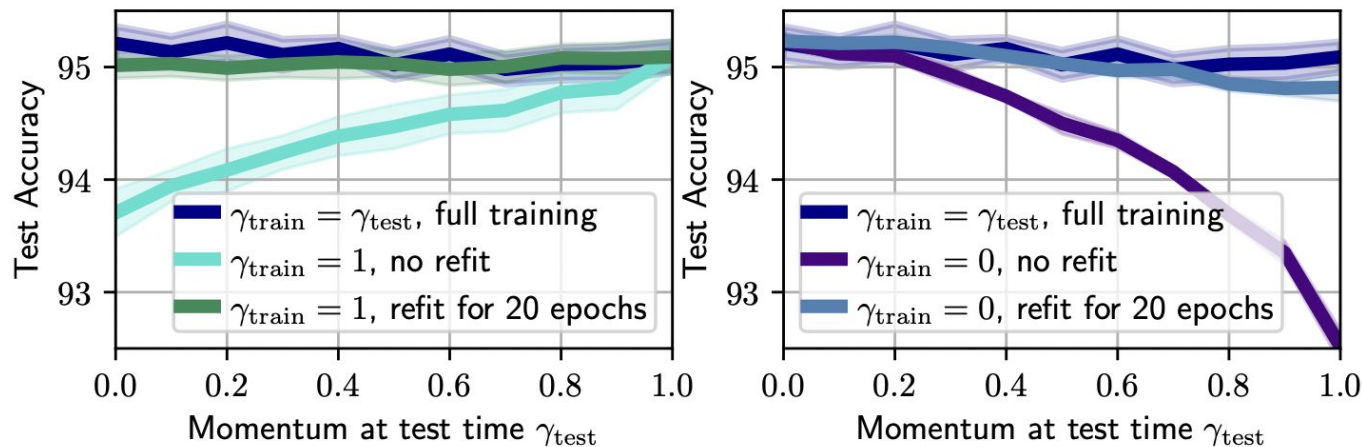
Эксперименты: классификация

| Model | CIFAR-10 | CIFAR-100 |
|---------------------------------|------------------|------------------|
| Momentum ResNet, $v_0 = 0$ | 95.1 ± 0.13 | 76.39 ± 0.18 |
| Momentum ResNet, $v_0 = f(x_0)$ | 95.18 ± 0.06 | 76.38 ± 0.42 |
| ResNet | 95.15 ± 0.12 | 76.86 ± 0.25 |

Классификация изображений, $\gamma = 0.9$ +
разная инициализация v_0 .

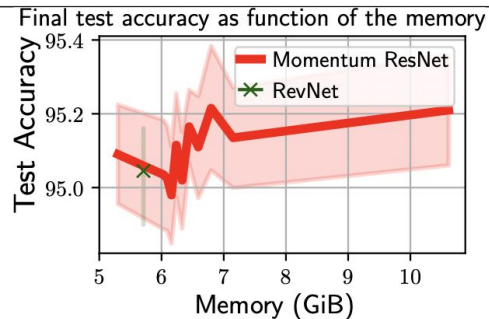
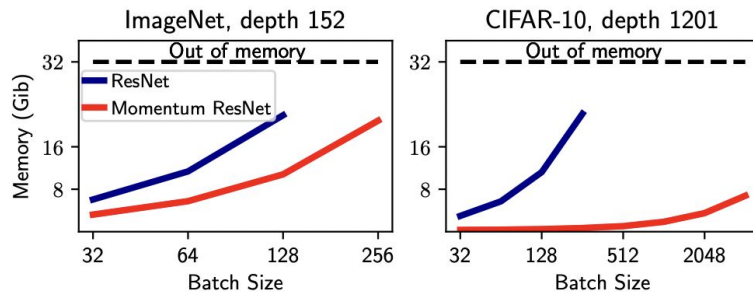


Эксперименты: влияние гаммы



full training – одна γ для обучения и теста, no refit – разные γ во время обучения и теста, refit – дообучение с новым γ .

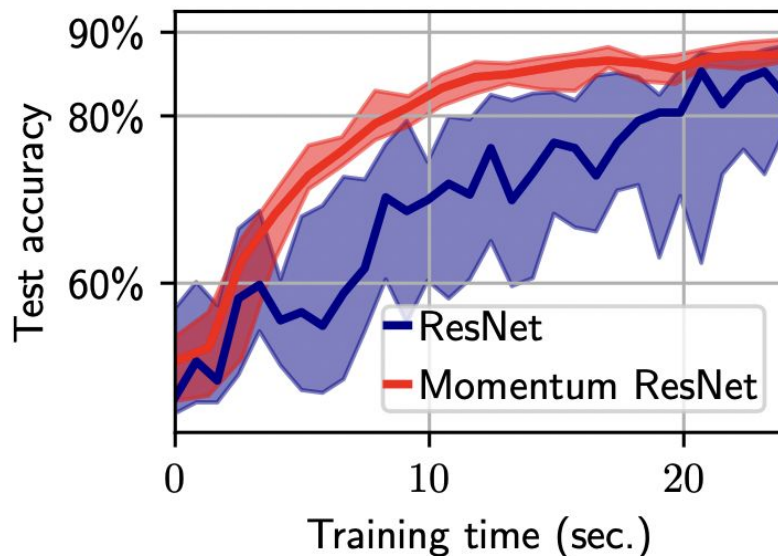
Эксперименты: потребление памяти



Верхний ряд – потребление памяти от размера батча.

Нижний ряд – точность от количества памяти.

Эксперименты: дообучение



Дообучение на humenoptera. Каждая картинка имеет размер 500x500, максимальный батч для ResNet – 2, для Momentum ResNet – 4.

Итог

- Универсальный способ инвертировать resblock в любой архитектуре
- Сокращение потребления памяти – практически не зависит от глубины
- В непрерывном случае получается Neural ODE второго порядка
- Производительность сопоставима с обычным ResNet

Рецензия

Плюсы:

- Статья предлагает эффективный метод уменьшения потребления памяти без заметной потери качества
- Все преимущества новой архитектуры теоретически обоснованы
- Возможность использования в других задачах, таких как оптимизация
- Большое количество разнообразных экспериментов
- Воспроизводимость: все заявления авторов статьи подтверждаются на практике

Рецензия

Минусы:

- Отсутствуют эксперименты с сетями небольшой глубины
- Статья довольно сильно перегружена математикой, из-за чего в некоторых моментах её тяжело воспринимать

Рецензия

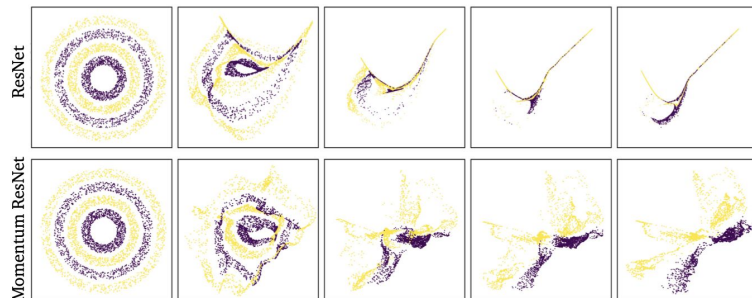
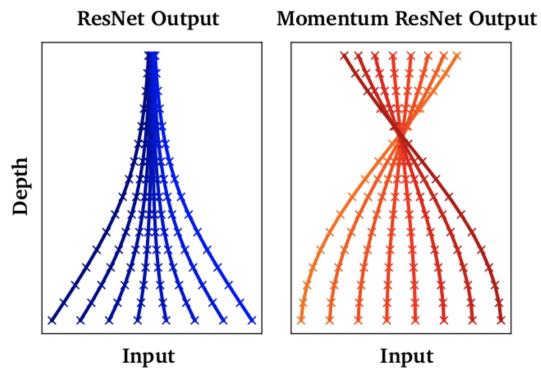
Оценка: 8

Уверенность: 4

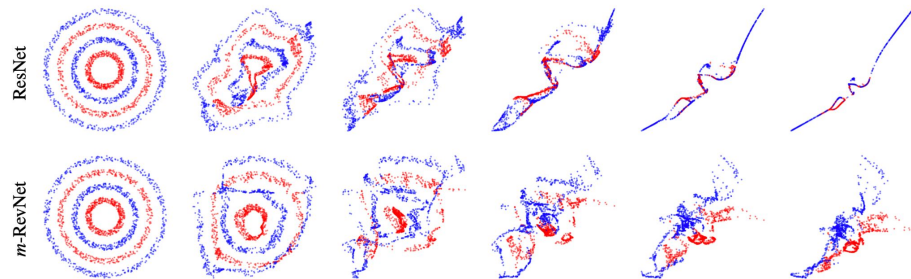
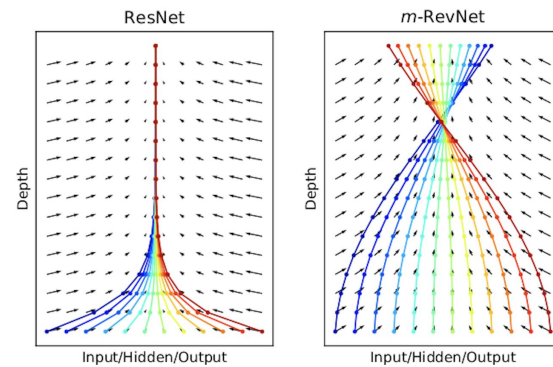
Исследование контекста работы

- Статья впервые была загружена на arXiv 15 февраля 2021 года. Финальная версия была представлена на 38-ой конференции ICML
- Статья представлена на конференции в виде poster и spotlight
- Плагиат статьи Duo Li, Shang-Hua Gao “m-RevNet: Deep Reversible Neural Networks with Momentum” приняли на ICCV

| | <i>Neur.ODE</i> | <i>i-ResNet</i> | <i>i-RevNet</i> | <i>RevNet</i> | <i>Mom.Net</i> |
|------------------------|-----------------|-----------------|-----------------|---------------|----------------|
| Closed-form inversion | ✓ | ✗ | ✓ | ✓ | ✓ |
| Same parameters | ✗ | ✓ | ✗ | ✗ | ✓ |
| Unconstrained training | ✓ | ✗ | ✓ | ✓ | ✓ |



| Method | ResNet | RevNet/ <i>i-RevNet</i> | <i>i-ResNet</i> | NODE (and variants) | <i>m-RevNet (ours)</i> |
|----------------------------|--------|-------------------------|-----------------|---------------------|------------------------|
| Analytical Reversal | N/A | ✓ | ✗ | ✗ | ✓ |
| Architectural Preservation | N/A | ✗ | ✓ | ✗ | ✓ |
| End-to-End Optimization | ✓ | ✓ | ✗ | ✓ | ✓ |



Imitation is the sincerest form of flattery (Pierre Ablin)

Авторы: Michael E. Sander, Pierre Ablin, Mathieu Blondel, Gabriel Peyre

- Michael E. Sander - Ph. d. студент в ENS (Ecole Normale Supérieure); researcher в CNRS - Национальный центр научных исследований во Франции. Это первая его публикация
- Pierre Ablin - postdoc в ENS; researcher в CNRS
- Mathieu Blondel - senior research scientist at Google Research, Brain team во Франции
- Gabriel Peyré - senior researcher в CNRS и работает в ENS

Статьи

На кого ссылаются:

- The Reversible Residual Network: Backpropagation Without Storing Activations (RevNet) Gomez et al. 2017
- Neural Ordinary Differential Equations (Neural ODE) Chen et al. 2018

Кто ссылается:

- HeunNet: Extending ResNet using Heun's Methods; Maleki et al. 2021
- Neural ODE control for classification, approximation and transport, Ruiz-Balet et al., 2021

Хакер. Интерфейсы

Библиотека: <https://github.com/michaelsdr/momentumnet>

Документация: <https://michaelsdr.github.io/momentumnet>

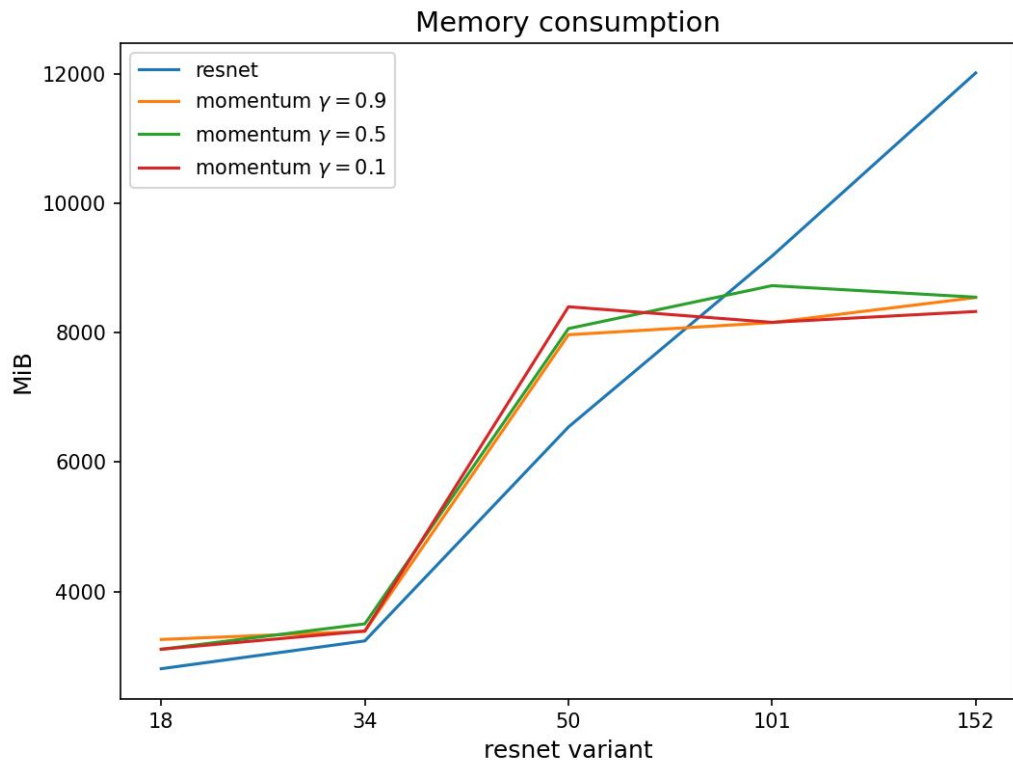
Можно обращаться любые residual блоки. Momentum сетью можно сделать в том числе torch.nn.Transformer

```
from momentumnet import transform_to_momentumnet

resnet = resnet18(pretrained=True)
mresnet = transform_to_momentumnet(resnet, gamma=0.9, use_backprop)

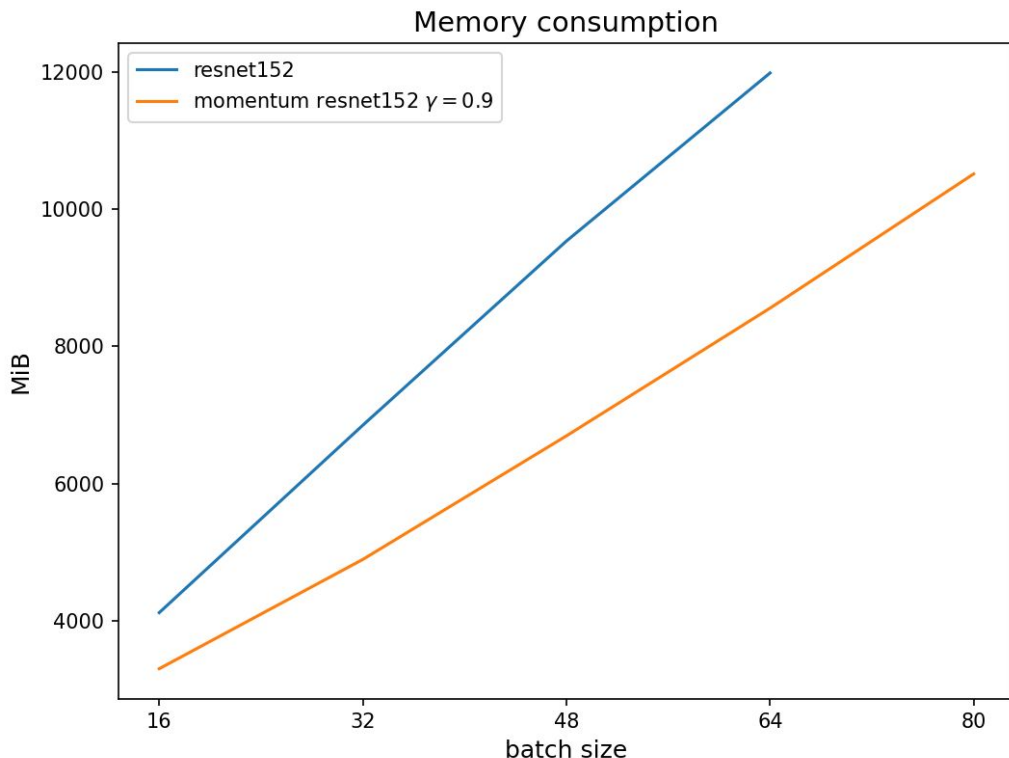
transformer = torch.nn.Transformer(num_encoder_layers=6, num_decoder_layers=6)
layers = ["encoder.layers", "decoder.layers"]
mtransformer = transform_to_momentumnet(transformer, sub_layers=layers, gamma=0.9, use_backprop=False,
keep_first_layer=False)
```

Хакер. Потребление памяти



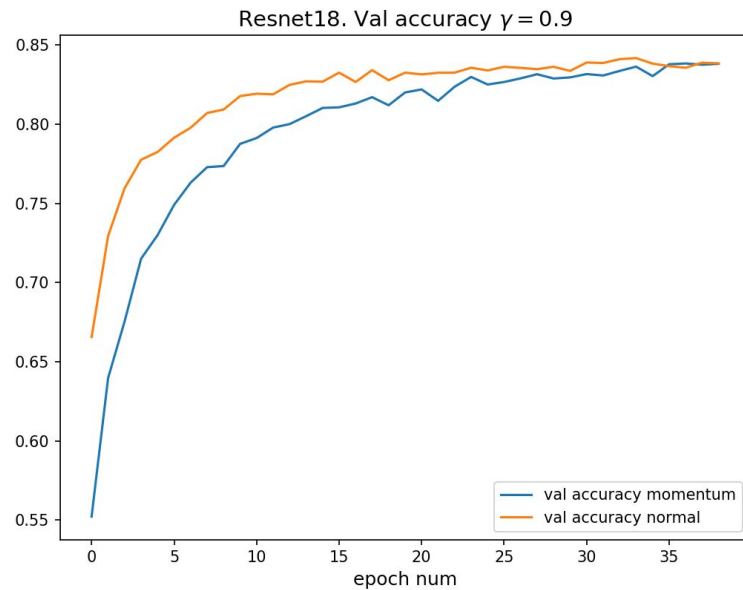
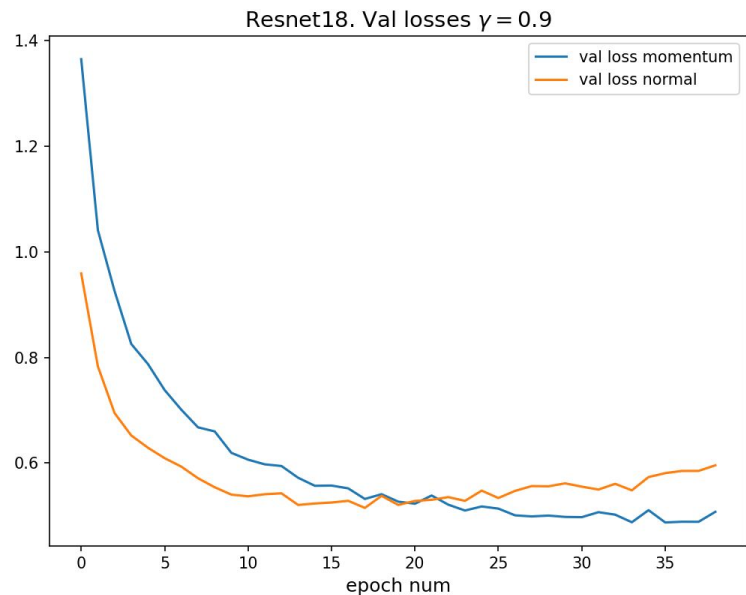
- backward() на синтетическом батче (64 x 3 x 224 x 224)
- При малой глубине momentum версия немного хуже обычной
- Начиная с resnet50 память momentum версии почти не растет
- Память обычного resnet'a растет линейно
- При глубоких моделях (101, 152 и более) можно сэкономить много памяти

Хакер. Большая модель с разными размерами батча



- backward() на синтетическом батче ($\text{batch_size} \times 3 \times 224 \times 224$)
- resnet152 и momentum resnet152
- batch_size 80 не влез в 16гб у обычного resnet
- Оценки авторов на потребление памяти:
ResNet: $O(k \times d \times n_{\text{batch}})$
Momentum ResNet: $O((1 - \gamma) \times k \times d \times n_{\text{batch}})$

Хакеп. Finetuning на CIFAR10. Resnet101



Вопросы

1. Сформулируйте формулу для прямого прохода momentum модели. Как ее обратить?
2. Приведите и поясните оценки потребляемой памяти для обычного резнета и momentum резнета.
3. Перечислите основные свойства моментум резнета.