Momentum Residual Neural Networks[1]

Докладчик - Алексей Цеховой, 181
Рецензент - Сибагатова Софья, 181
Практик-исследователь - Барановская Дарья, 181
Хакер - Андреев Никита, 181

[1] M. E. Sander, et al. - "Momentum Residual Neural Networks", 2021.

Мотивация

Цель:

- Хотим сократить память при обучении ResNet

Суть:

- Обратимая альтернатива ResNet: при обратном проходе активации вычисляются на ходу, при прямом не сохраняются

Обратимые остаточные архитектуры

Дискретные:

- Обратимая связь активаций
- Сложность либо с переносом модели, либо с обращением

Непрерывные:

- ResNet как динамическая система, решаем ОДУ
- Сложность с переносом модели
- Численная неустойчивость

Предложенный метод

ResNet:

- $x_{n+1} = x_n + f(x_n, \theta_n)$

Momentum ResNet:

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

обратный проход:

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

Обратимое умножение

Алгоритм[2]:

Input: Information buffer i, value c, ratio n/d
$i = i \times d$ ▷ make room for new digit
$i = i + (c \mod d)$ ▷ store digit lost by division
$c = c \div d$ ▷ divide by denominator
$c = c \times n$ ▷ multiply by numerator
$c = c + (i \mod n)$ ▷ add digit from buffer
$i = i \div n$ ▷ shorten information buffer
return updated buffer i, updated value c

- Битов на операцию: $\log_2(1/\gamma)$

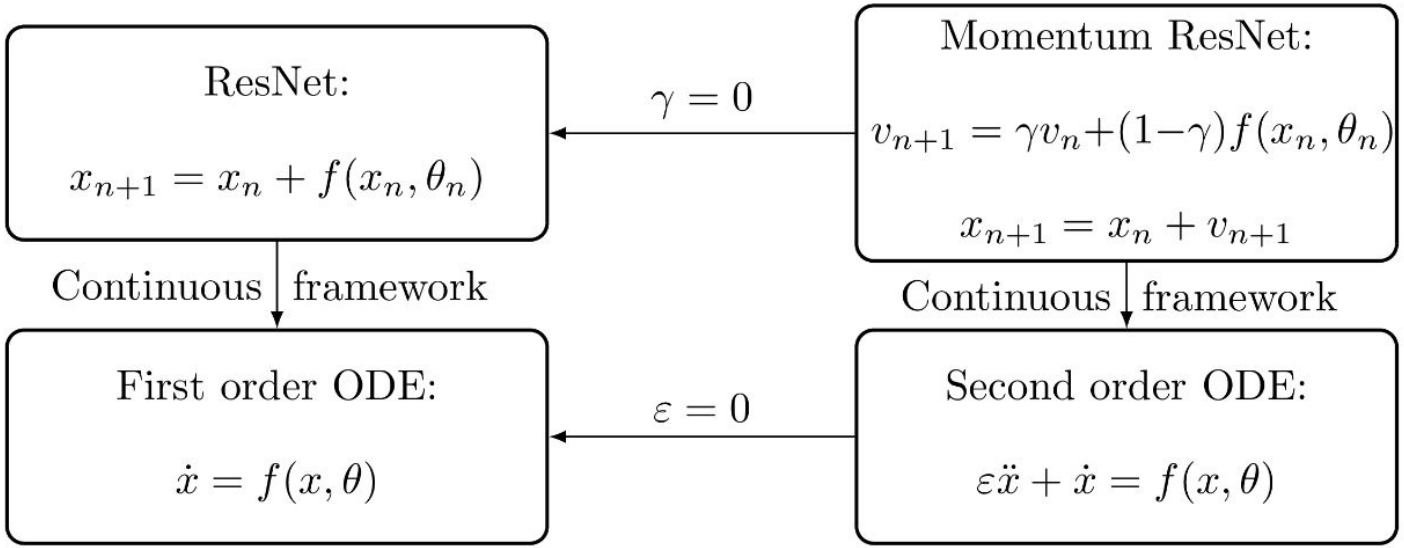
[2] D. Maclaurin, et al. - "Gradient-based hyperparameter optimization through reversible learning", 2015.

Обратимые остаточные архитектуры:

	Neur.ODE	*i-ResNet*	*i-RevNet*	*RevNet*	***Mom.Net***
Closed-form inversion	✓	✗	✓	✓	✓
Same parameters	✗	✓	✗	✗	✓
Unconstrained training	✓	✗	✓	✓	✓

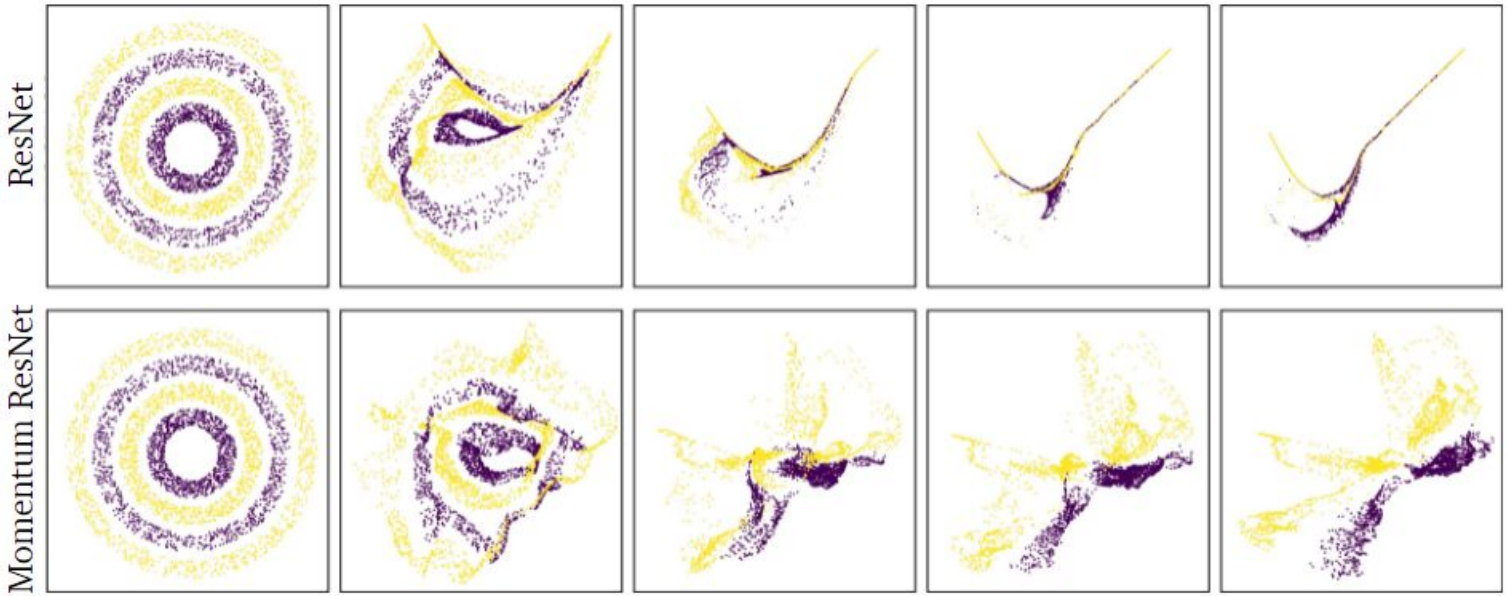
Возможность представления

- Обобщение ResNet (при моменте 0)
- Представляет ОДУ второго порядка, лучше ResNet - ОДУ первого порядка:



Опыт 1: лучшая возможность представления

- $f(x,\ \theta) = W_2^T tanh(W_1 x + b):\ W_1,\ W_2 \in \mathbb{R}^{16\times 2},\ b \in \mathbb{R}^{16}$
- 15 слоёв
- Разделение колец (каждые 3 слоя):



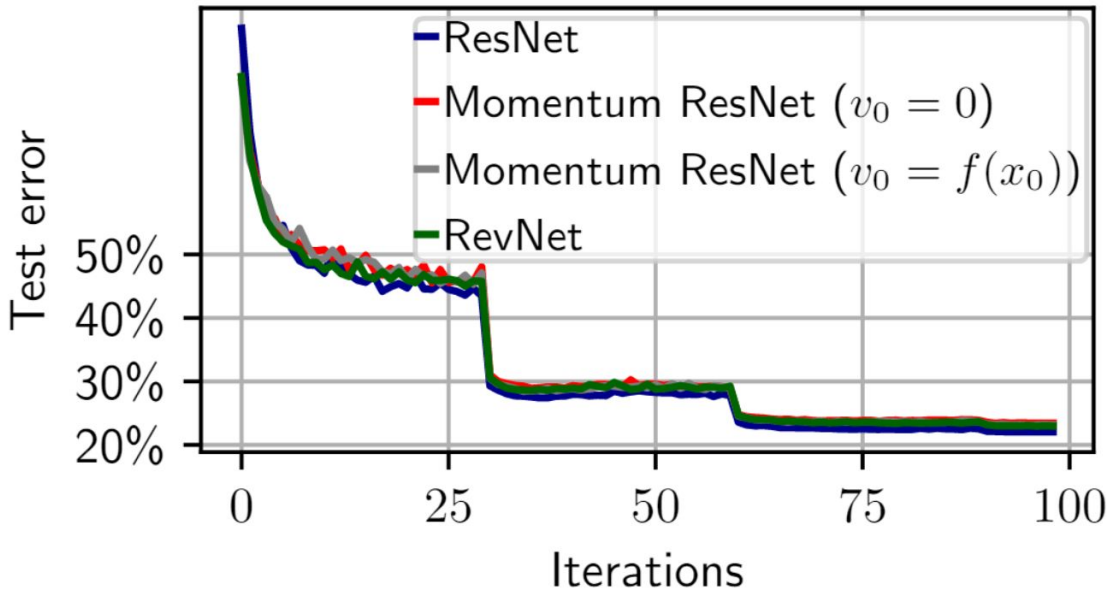
Опыт 2: точность

- ResNet-101 на CIFAR-10 и CIFAR-100
- $\gamma = 0.9$

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

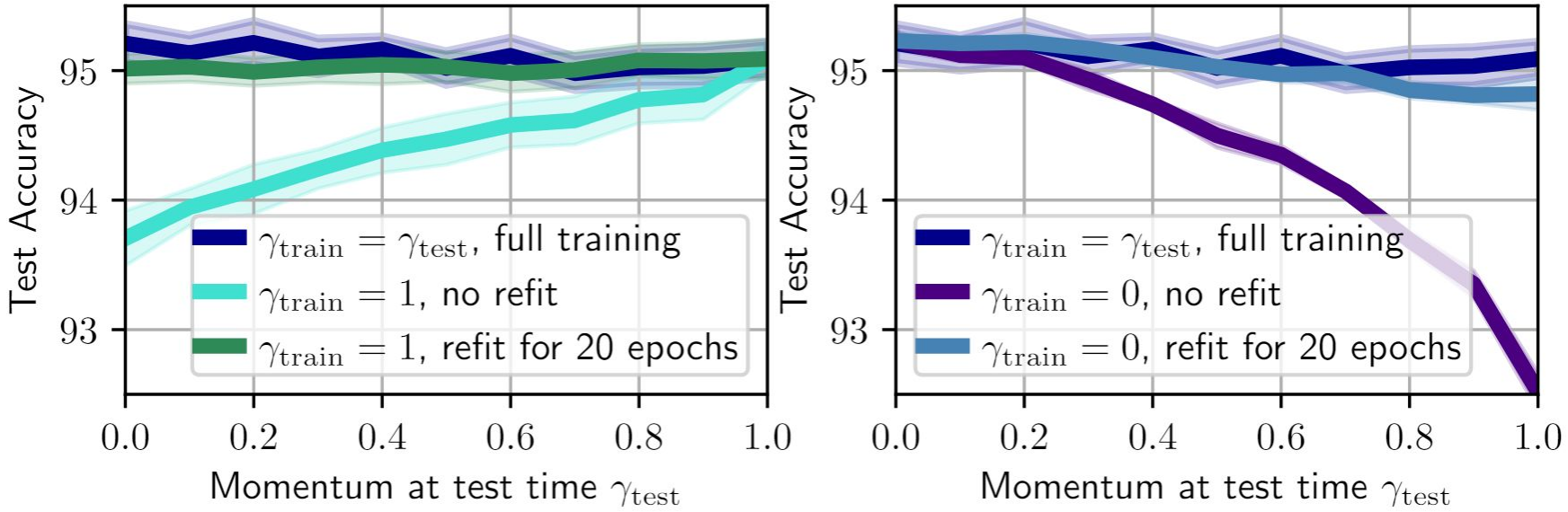
Опыт 3: точность

- ResNet-101 на ImageNet
- $\gamma = 0.9$
- ResNet: 22%, остальные: 23%



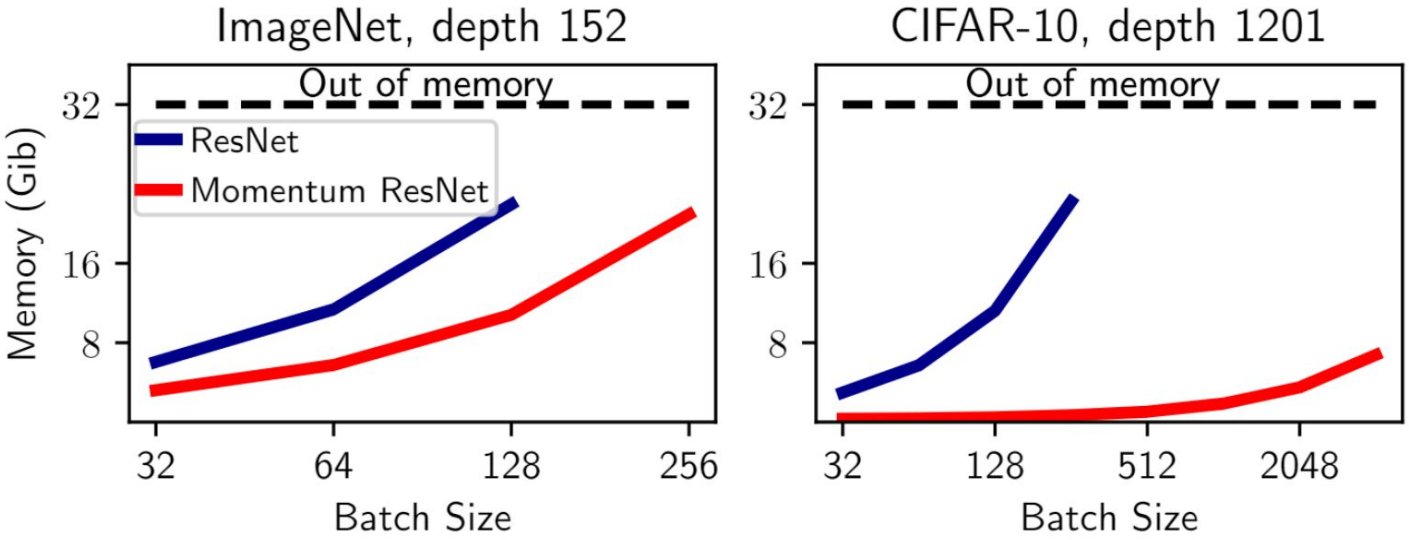
Опыт 4: роль моментов

- ResNet-101 на CIFAR-10
- $\gamma = 0.9$



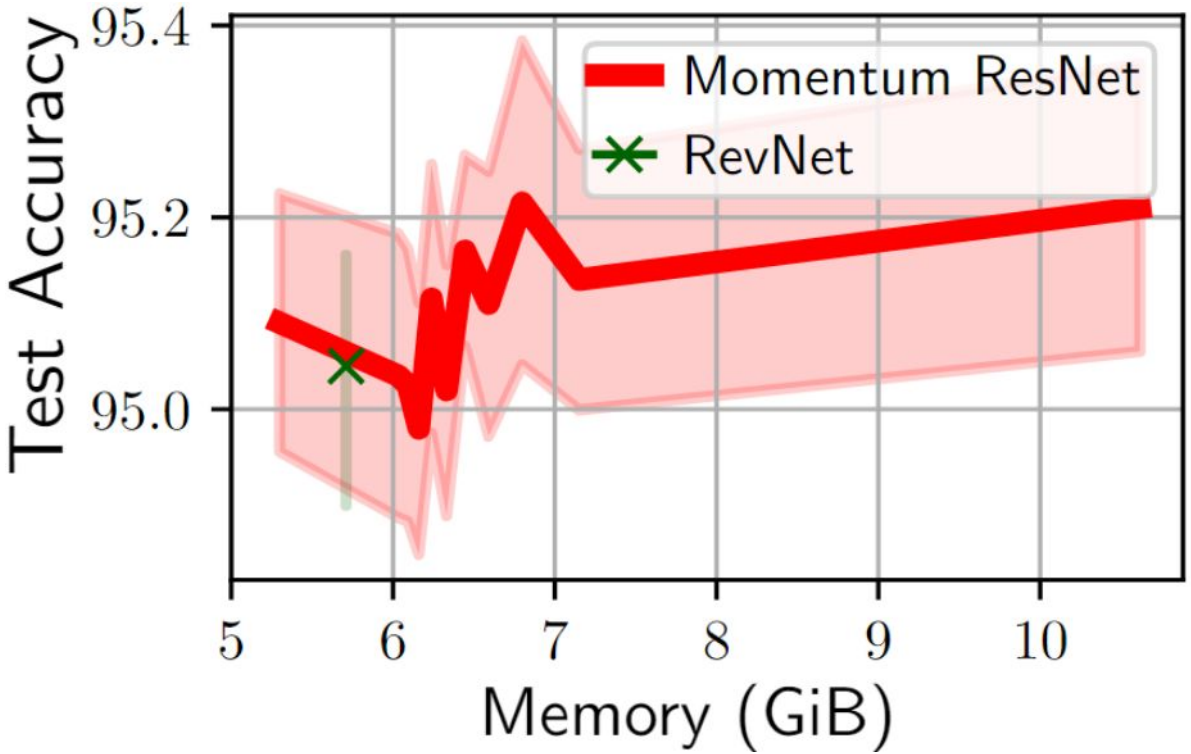
Опыт 5: память

- ImageNet, глубина 152
- CIFAR-10, глубина 1201
- $\gamma = 0.9$



Опыт 6: память

- ResNet-101 на CIFAR-10

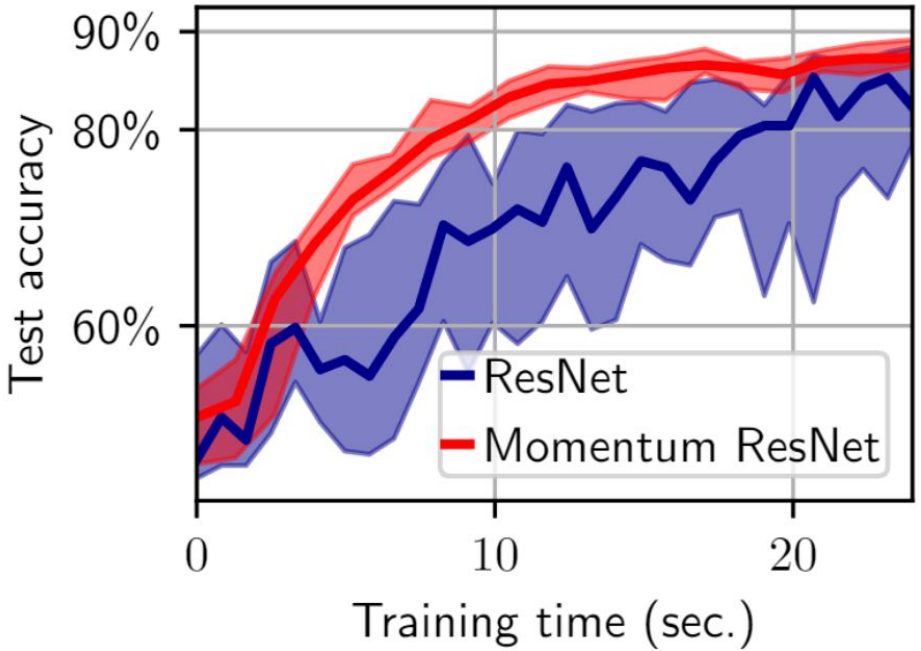


Опыт 7: перенос весов

- Обученный на Imagenet ResNet-152 (te=22%) перенесли в Momentum ResNet-152, $\gamma = 0.9$
- 1 эпоха дообучения: te=26.5%
- 5 эпох: te=23.5%

Опыт 8: fine-tuning

- Сети из опыта 7 на hymenoptera (500x500px)
- Батчи по 2 и 4 изображения соответственно



Заключение

- Значительно меньше памяти
- Сравнимое качество
- Простое построение из готовых ResNet
- Более широкий класс отображений

Momentum Residual Neural Networks

Michael E. Sander [1 2] **Pierre Ablin** [1 2] **Mathieu Blondel** [3] **Gabriel Peyré** [1 2]

ICML 2021 poster

Submission history

From: Michael E. Sander [view email]
[v1] Mon, 15 Feb 2021 22:24:52 UTC (2,224 KB)
[v2] Thu, 13 May 2021 12:29:54 UTC (4,555 KB)
[v3] Thu, 22 Jul 2021 08:18:05 UTC (4,554 KB)

Авторы

Michael E. Sander

Ecole Normale Supérieure

Machine Learning Applied Mathematics Optimal transport Differential equations

Pierre Ablin

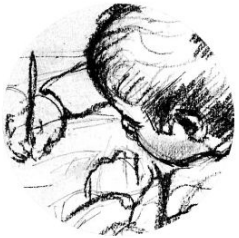
CNRS

Optimization Machine Learning Statistics

Mathieu Blondel

Google

Machine Learning



Gabriel Peyré

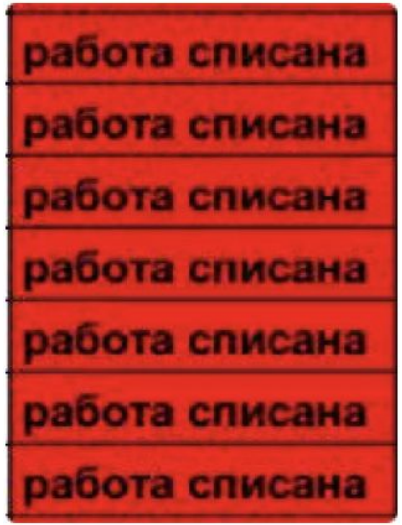
CNRS, DMA, Ecole Normale Supérieure

Applied Mathematics Imaging sciences Machine learning

[Submitted on 10 Dec 2015]

Deep Residual Learning for Image Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun



Плагиат

Alexander Chernov

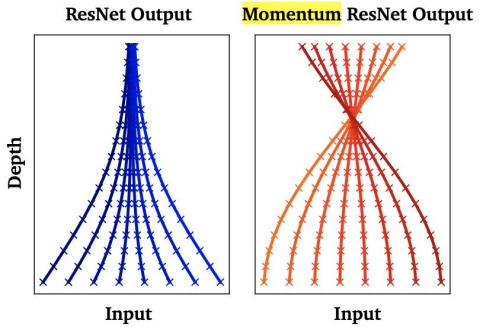
вас предупредили

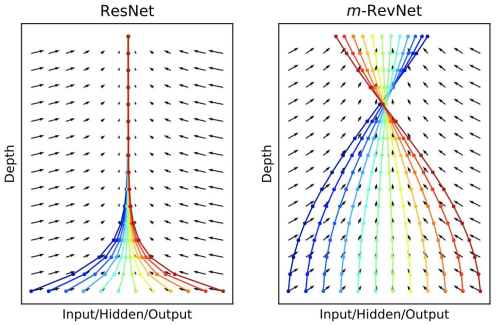
m-RevNet: Deep Reversible Neural Networks with Momentum

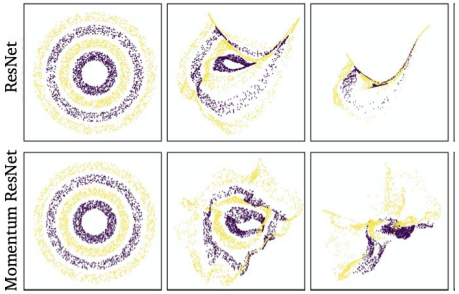
Duo Li[†] Shang-Hua Gao[‡]
[†]The Hong Kong University of Science and Technology [‡]Nankai University
duo.li@connect.ust.hk

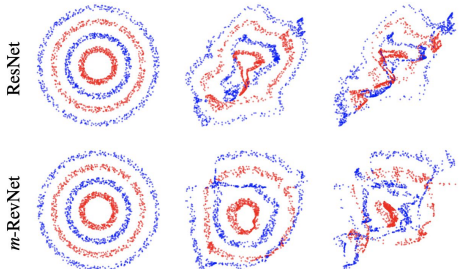
Submission history
From: Duo Li [view email]
[v1] Thu, 12 Aug 2021 17:14:32 UTC (4,231 KB)
[v2] Mon, 16 Aug 2021 13:04:04 UTC (0 KB)

- Только не скатывай точь в точь
- Хорошо









Рецензия. Сильные/слабые стороны

+ Актуальность решаемой проблемы
+ Применимые на практике результаты
+ Большое количество обширных теоретических и экспериментальных исследований

− Недостаточно подробное описание некоторых экспериментов
− В ключевых экспериментах о сравнении качества и потребляемой памяти Momentum ResNet и ResNet представлены результаты только для ResNet101 и его Momentum-аналога

Рецензия. Качество текста

\+ Текст написан последовательно, подробно, доходчиво

\− Большое количество орфографических и лингвистических ошибок, затрудняющих восприятие

3.2. Memory cost

Instead of storing the full data at each layer, we only need to store the bits lost at each multiplication by γ (cf. "intert-ibility"). For an architecture of depth k, this corresponds

⇨ invertibility

Top-1 classification error on ImageNet (single crop) for 4 different residual architectures of ımeters. Final test accuracy is 22% for the ResNet-101 and 23% for the 3 other invertible models. ıe performance as a RevNet with the same number of parameters.

⇨ error

Рецензия. Воспроизводимость

+ Предоставлен пакет для PyTorch, позволяющий получить Momentum-аналог любой модели ResNet (на самом деле также подходит для трансформеров)
+ Полное описание проводимых экспериментов

– Недостаточно подробное для воспроизведения описание эксперимента с облаками точек, но предоставлен код эксперимента, в котором есть все нужные гиперпараметры

Рецензия. Дополнительные комментарии

- Предоставить результаты для ResNet различной глубины
- Провести сравнительные эксперименты времени обучения ResNet и Momentum ResNet
- Провести аналогичные для ResNet эксперименты по количеству памяти, качеству, возможности для fine-tuning моделей, содержащих residual блоки (например, трансформеры)

Оценка статьи по критериям НИПСа:

- Оценка: 9 из 10
- Уверенность: 5 из 5

Хакер. Применение

Можно обращать любые архитектуры с residual блоками. Для этого есть понятные функции и документация[3].

```
import torch
from momentumnet import transform_to_momentumnet
from torchvision.models import resnet101
resnet = resnet101(pretrained=True)
mresnet101 = transform_to_momentumnet(resnet, gamma=0.9, use_backprop=False)
```

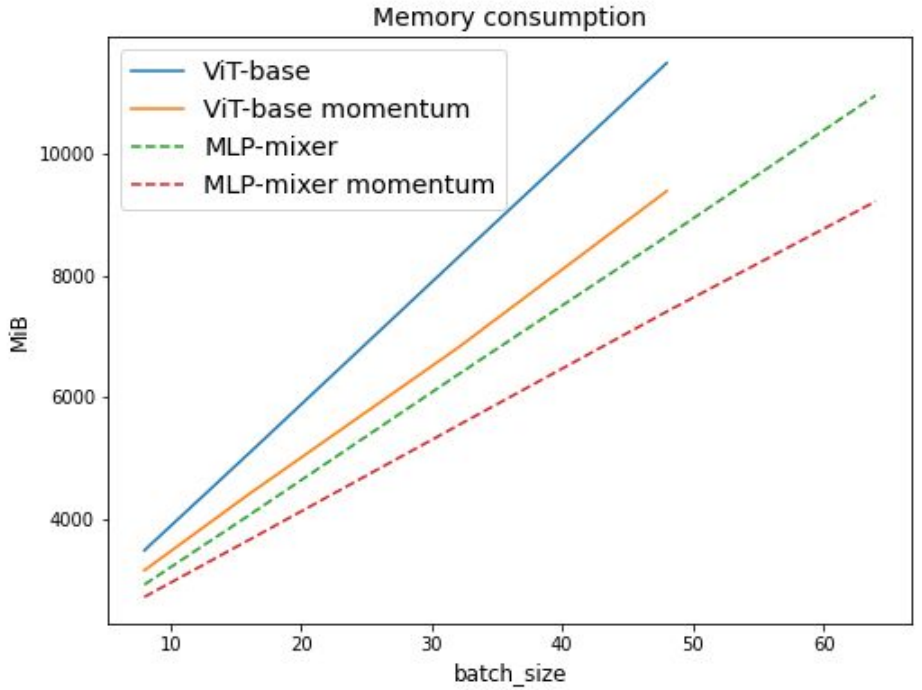
```
import torch
from momentumnet import transform_to_momentumnet
transformer = torch.nn.Transformer(num_encoder_layers=6, num_decoder_layers=6)
mtransformer = transform_to_momentumnet(transformer, sub_layers=["encoder.layers", "decoder.layers"],
                                        gamma=0.9, use_backprop=False,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

Таким образом можно экспериментировать с различными архитектурами и их момемнтум версиями.

[3] https://github.com/michaelsdr/momentumnet

Хакер. Эксперименты с памятью

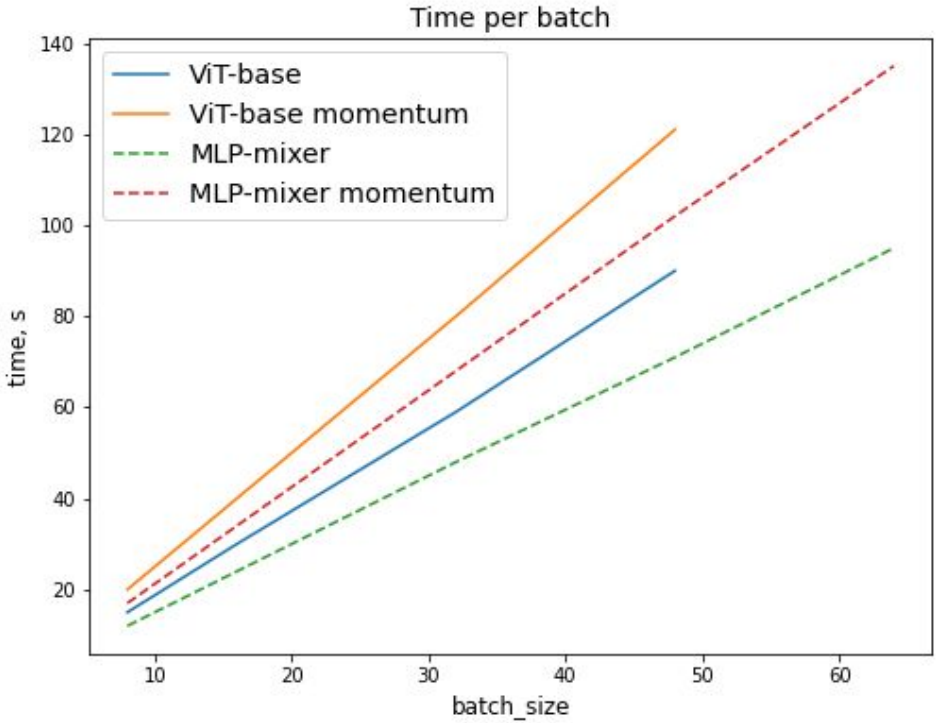
- Прогоняем модель на одном батче размера bs x 3 x 224 x 224
- Эксперименты проводим с архитектурами Vision Transformer[4] и MLP mixer[5]
- Batch_size 64 не влез для Vision Transformer
- Видим, что память растет линейно, однако некоторая экономия все же наблюдается



[4] https://arxiv.org/pdf/2010.11929.pdf
[5] https://arxiv.org/pdf/2105.01601.pdf

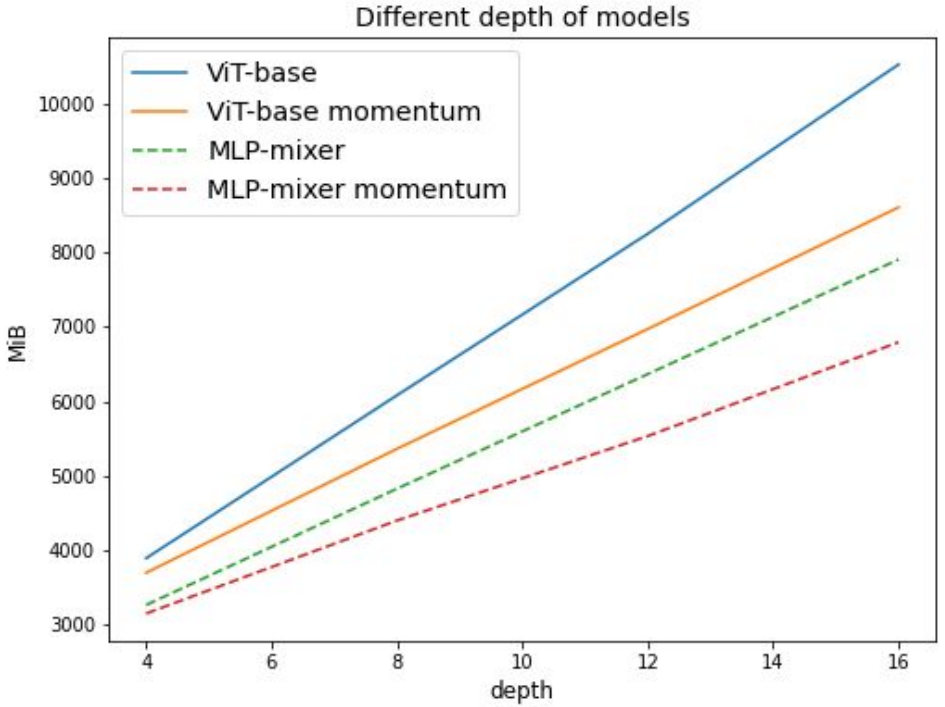
Хакер. Эксперименты со временем

- Прогоняем модель на одном батче размера bs x 3 x 224 x 224
- Эксперименты проводим с архитектурами Vision Transformer и MLP mixer
- Batch_size 64 не влез для Vision Transformer
- Видим, что время растет линейно, однако моментум версии проигрывают обычным моделям



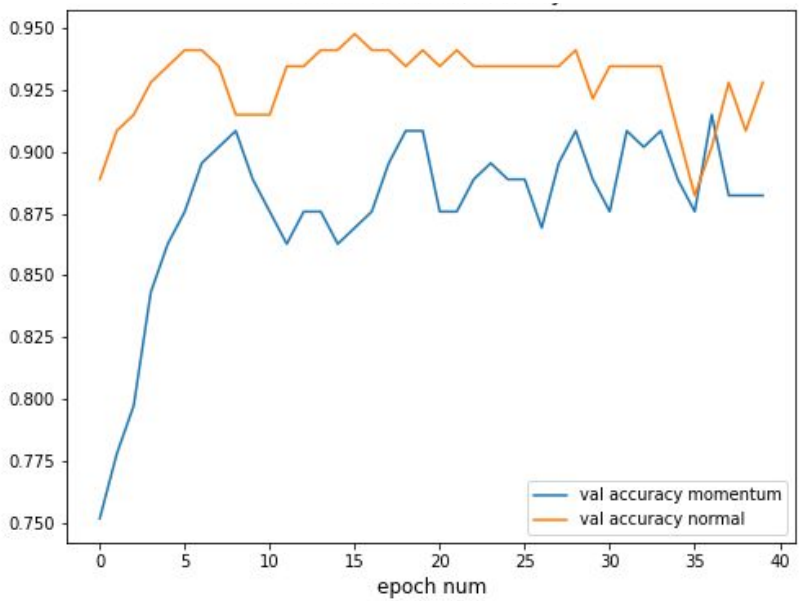
Хакер. Эксперименты с глубиной модели

- Прогоняем модель на одном батче размера 32 x 3 x 224 x 224
- Эксперименты проводим с архитектурами Vision Transformer и MLP mixer
- Видим, что при росте глубины модели (параметр depth) также есть эффект от применения моментум модификации



Хакер. Fine-tuning

ResNet 18. Validation accuracy



ResNet 18. Validation loss