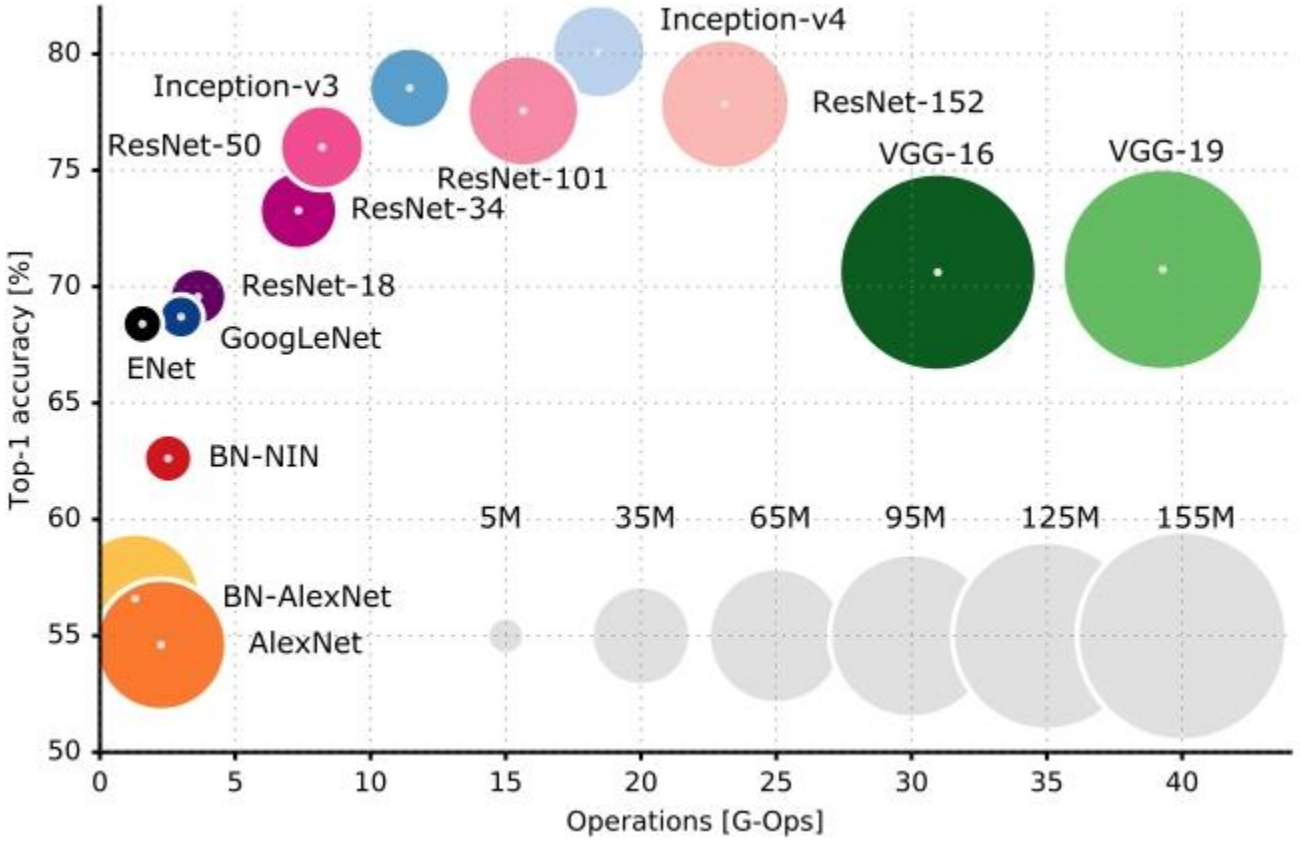
Training ImageNet in 1 Hour

Гарницкий Марк

Устали ждать?



Solution:

- We can train ResNet-50 in 1 hour using 256 GPUs
- The per-worker workload must be large
- momentum SGD

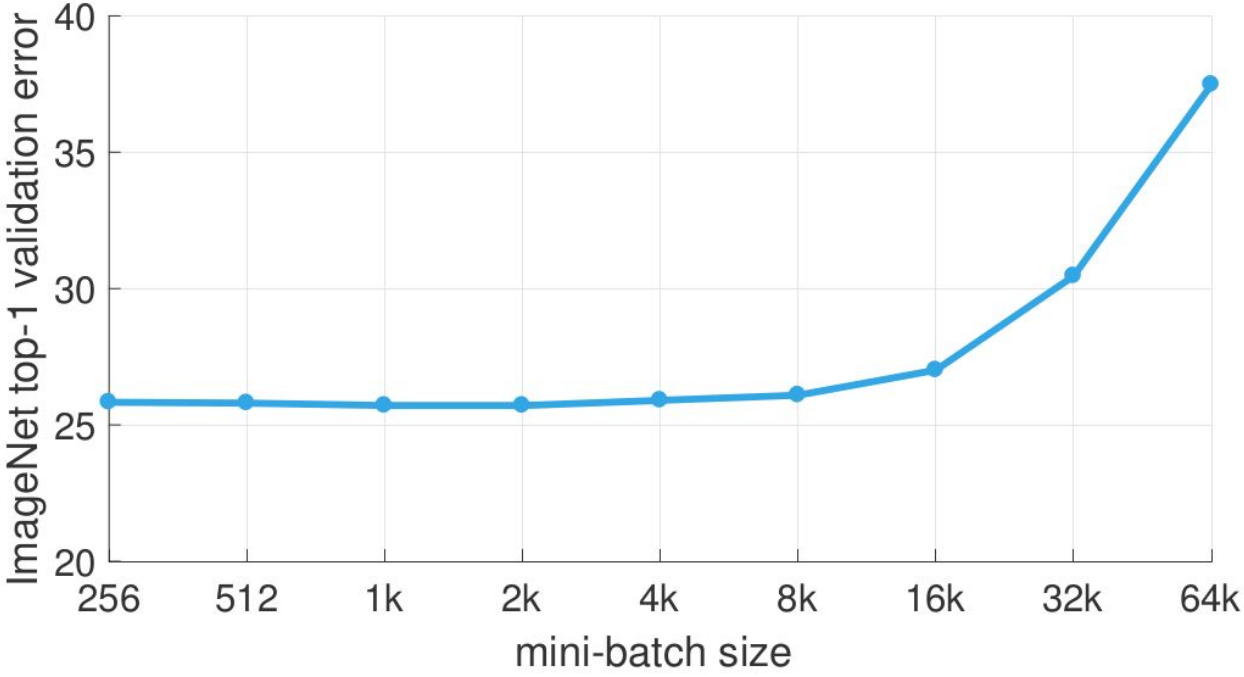


Figure 6. **ImageNet-5k** top-1 validation error *vs*. minibatch size

Learning rates for Large Minibatches

Linear Scaling Rule:
When the minibatch size is multiplied by k, multiply the learning rate by k.
All other hyper-parameters (weight decay, etc.) are kept unchanged.

Interpretation:

After k steps:

$$w_{t+k} = w_t - \eta \frac{1}{n} \sum_{j<k} \sum_{x \in \mathcal{B}_j} \nabla l(x, w_{t+j}).$$

Single step:

$$\hat{w}_{t+1} = w_t - \hat{\eta} \frac{1}{kn} \sum_{j<k} \sum_{x \in \mathcal{B}_j} \nabla l(x, w_t).$$

If $\nabla l(x, w_{t+j}) \approx \nabla l(x, w_t)$

then $\hat{\eta} = kn \rightarrow \hat{w}_{t+1} \approx w_{t+k}$

Problems:

- In initial training when the network is changing rapidly, it does not hold
- Minibatch size cannot be scaled indefinitely

Solving:

Warmup strategy

Constant warmup

Для первых m эпох мы тренируем сеть с уменьшенным learning rate=η, а затем возвращаемся к learning rate=kη.

Gradual warmup

Мы стартуем с learning rate=η, и постепенно домножаем его на константу, так чтобы после m эпох мы получили learning rate=kη

Batch Normalization with Large Minibatches

$$L(B, w) = \frac{1}{n} \sum_{x \in B} l_B(x, w) \rightarrow L(w) = \frac{1}{\|X^n\|} \sum_{B \in X^n} L(B, w)$$

$$L(B,w) = \frac{1}{n}\sum_{x\in B} l_B(x,w) \to L(w) = \frac{1}{\|X^n\|}\sum_{B\in X^n} L(B,w)$$

$$w_{t+k} = w_t - \eta \sum_{j<k} \nabla L(B_j, w_{t+j})$$

$$\hat{w}_{t+1} = w_t - \hat{\eta}\frac{1}{k}\sum_{j<k} \nabla L(B_j, w_t)$$

$$L(B, w) = \frac{1}{n} \sum_{x \in B} l_B(x, w) \to L(w) = \frac{1}{\|X^n\|} \sum_{B \in X^n} L(B, w)$$

$$w_{t+k} = w_t - \eta \sum_{j<k} \nabla L(B_j, w_{t+j})$$

$$\hat{w}_{t+1} = w_t - \hat{\eta} \frac{1}{k} \sum_{j<k} \nabla L(B_j, w_t)$$

Мы сохраняем размер выборки для каждого воркера, когда меняем количество воркеров k

$$L(B, w) = \frac{1}{n} \sum_{x \in B} l_B(x, w) \to L(w) = \frac{1}{\|X^n\|} \sum_{B \in X^n} L(B, w)$$

$$w_{t+k} = w_t - \eta \sum_{j<k} \nabla L(B_j, w_{t+j})$$

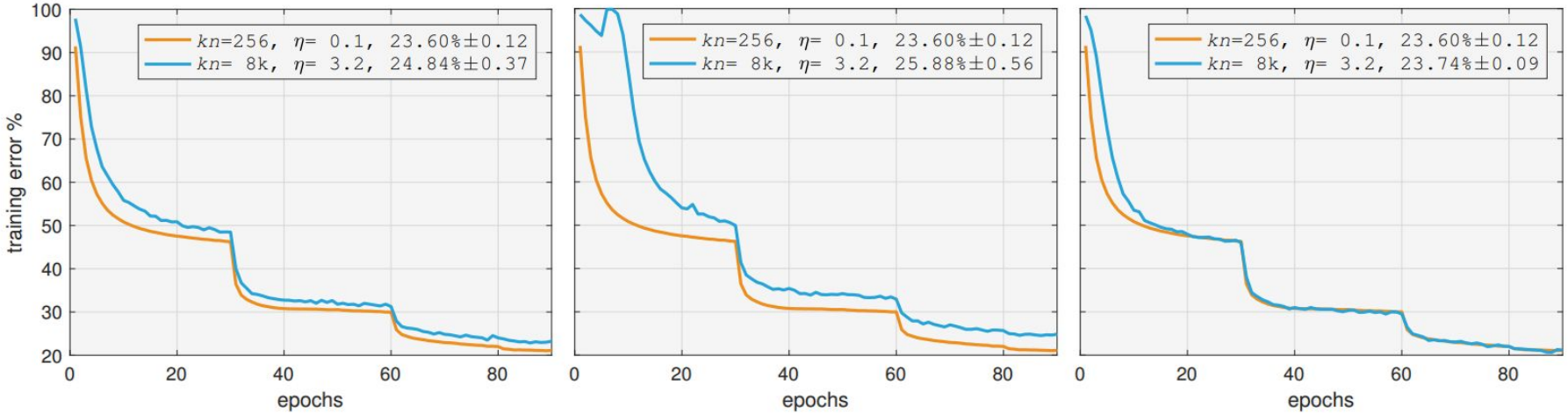
$$\hat{w}_{t+1} = w_t - \hat{\eta} \frac{1}{k} \sum_{j<k} \nabla L(B_j, w_t)$$

Мы сохраняем размер выборки для каждого воркера, когда меняем количество воркеров k

In this work, we use n = 32

	k	n	kn	η	top-1 error (%)
baseline (single server)	8	32	256	0.1	23.60 $\pm$0.12
no warmup, Figure 2a	256	32	8k	3.2	24.84 $\pm$0.37
constant warmup, Figure 2b	256	32	8k	3.2	25.88 $\pm$0.56
gradual warmup, Figure 2c	256	32	8k	3.2	23.74 $\pm$0.09

Table 1. **Validation error on ImageNet using ResNet-50** (mean and std computed over 5 trials). We compare the small minibatch model (kn=256) with large minibatch models (kn=8k) with various warmup strategies. Observe that the top-1 validation error for small and large minibatch training (with gradual warmup) is quite close: 23.60% $\pm$0.12 *vs*. 23.74% $\pm$0.09, respectively.



(a) no warmup (b) constant warmup (c) gradual warmup

Figure 2. **Warmup.** Training error curves for minibatch size 8192 using various warmup strategies compared to minibatch size 256. *Validation* error (mean±std of 5 runs) is shown in the legend, along with minibatch size kn and reference learning rate η.

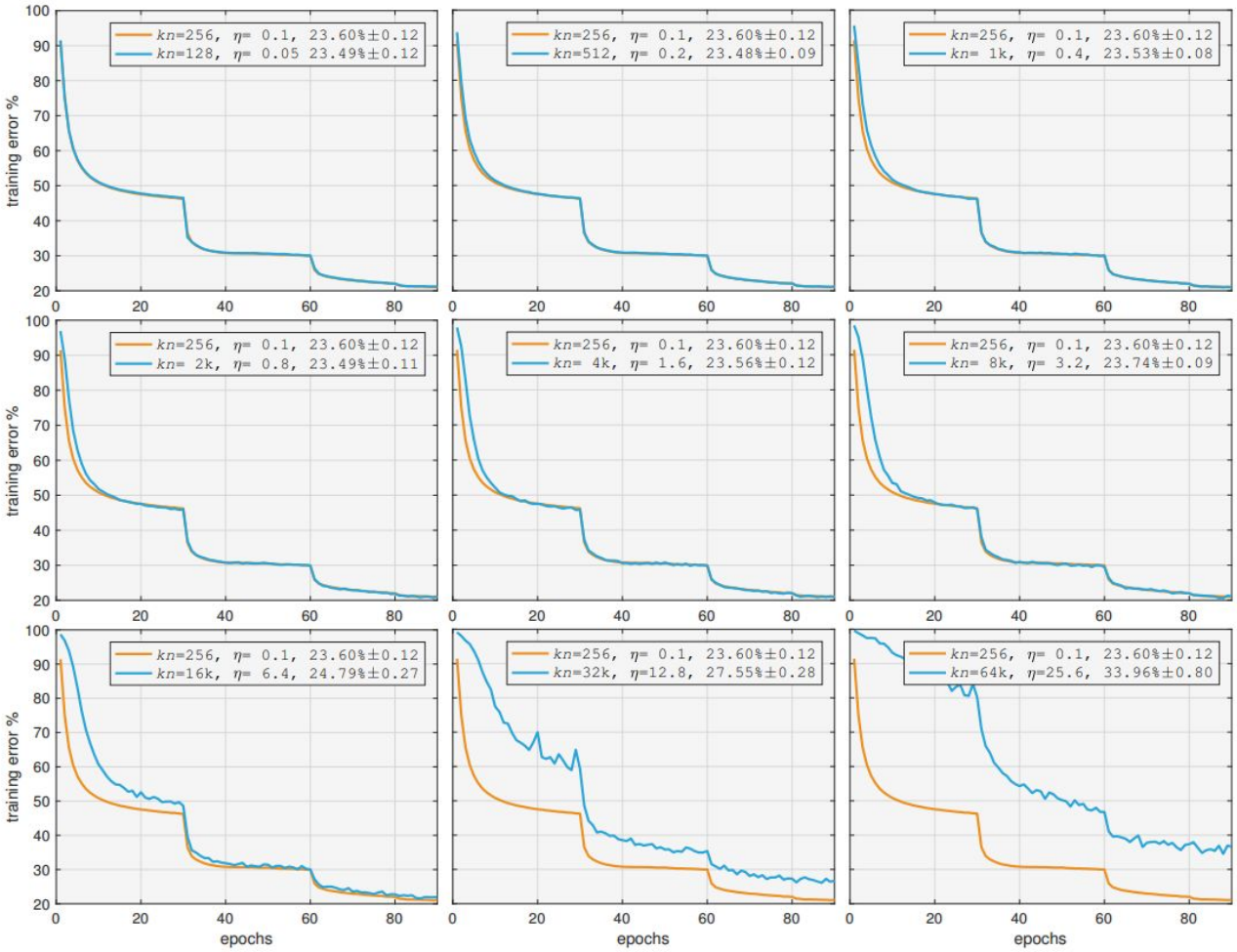


Figure 3. **Training error *vs.* minibatch size.** Training error curves for the 256 minibatch baseline and larger minibatches using gradual warmup and the linear scaling rule. Note how the training curves closely match the baseline (aside from the warmup period) up through 8k minibatches. *Validation* error (mean±std of 5 runs) is shown in the legend, along with minibatch size kn and reference learning rate η.

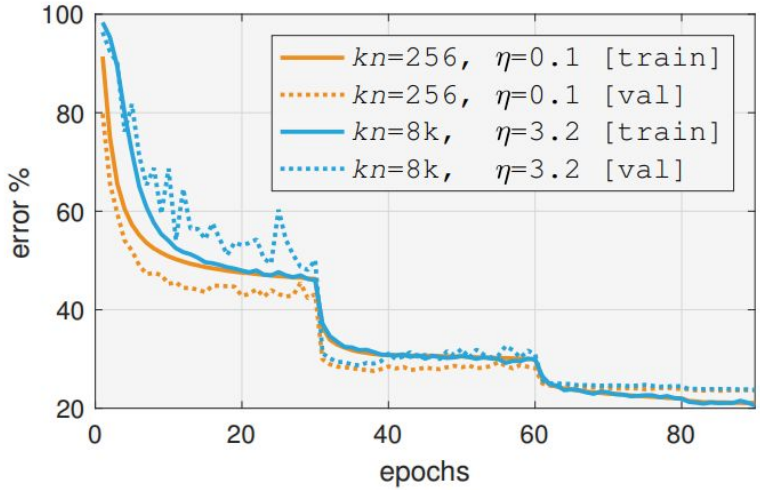


Figure 4. **Training and validation curves** for large minibatch SGD with gradual warmup *vs*. small minibatch SGD. Both sets of curves match closely after training for sufficient epochs. We note that the BN statistics (for inference only) are computed using *running* average, which is updated less frequently with a large minibatch and thus is noisier in early training (this explains the larger variation of the validation error in early epochs).

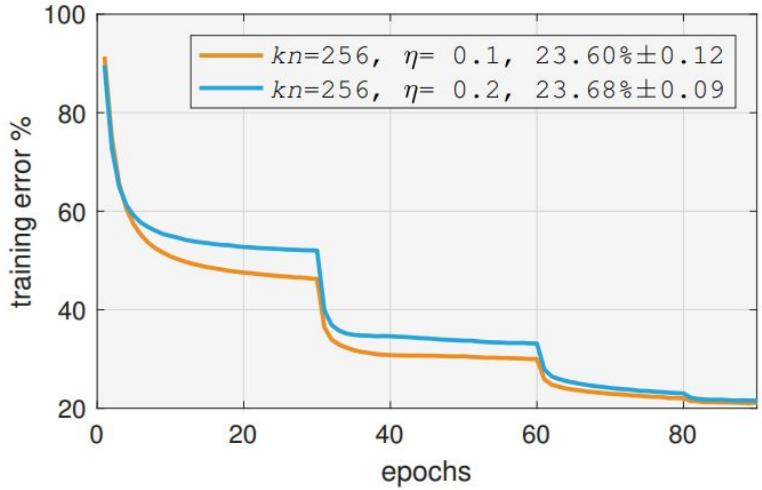


Figure 5. **Training curves for small minibatches with different learning rates η.** As expected, changing η results in curves that *do not match*. This is in contrast to changing batch-size (and linearly scaling η), which results in curves that *do match*, *e.g*. see Figure 3.

kn	η	top-1 error (%)
256	0.05	23.92 ±0.10
256	0.10	23.60 ±0.12
256	0.20	23.68 ±0.09
8k	$0.05 \cdot 32$	24.27 ±0.08
8k	$0.10 \cdot 32$	23.74 ±0.09
8k	$0.20 \cdot 32$	24.05 ±0.18
8k	0.10	41.67 ±0.10
8k	$0.10 \cdot \sqrt{32}$	26.22 ±0.03

(a) **Comparison of learning rate scaling rules.** A reference learning rate of $\eta = 0.1$ works best for $kn = 256$ (23.68% error). The linear scaling rule suggests $\eta = 0.1 \cdot 32$ when $kn = 8$k, which again gives best performance (23.74% error). Other ways of scaling η give worse results.

Table 2. **ImageNet classification experiments.** Unless noted all experiments use ResNet-50 and are averaged over 5 trials.

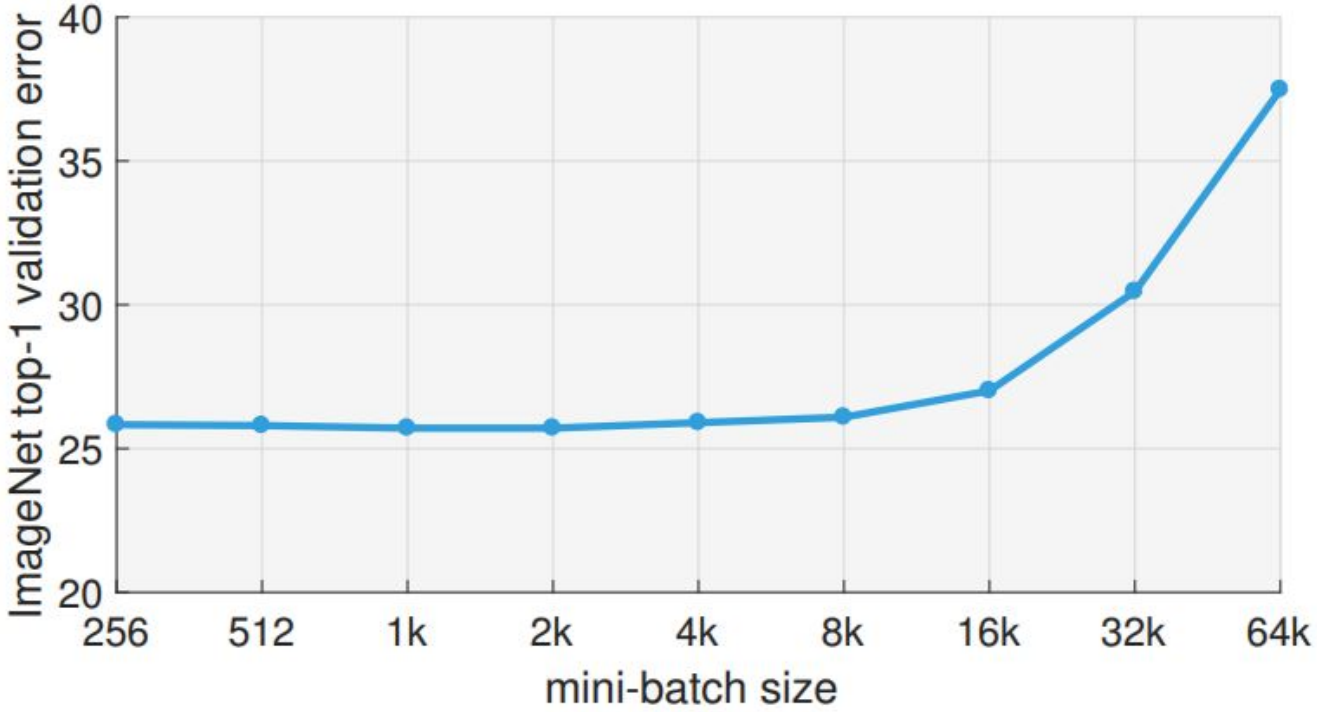


Figure 6. **ImageNet-5k** top-1 validation error *vs*. minibatch size with a fixed 90 epoch training schedule. The curve is qualitatively similar to results on ImageNet-1k (Figure 1) showing that a 5× increase in training data does not lead to a significant change in the maximum effective minibatch size.

- Взяли 256 GPU
- Наобучали разными способами сети с командой из 20 человек
- Вывели свои правила для обучения с large minibatch, которые получили эмпирически. (Простые смертные даже не могут проверить)
- Нашли в них какую-то логику
- Profit: Обучили ResNet-50 за 1 час на ImageNet с ошибкой близкой к baseline из статьи + написали статейку

References:

1. https://arxiv.org/abs/1605.07678
2. https://arxiv.org/abs/1706.02677