An argument in favor of strong scaling for deep neural networks with small datasets

Renato L. de F. Cunha, Eduardo R. Rodrigues, Matheus Viana, Dario Borges

<renatoc@br.ibm.com> | renatocunha | renatolfc | renatolfc

HPML 2018 - 2018-09-24

Where is our work positioned

- Deep Learning
- Parallel implementations

Key takeaway

> *A **strong scaling** approach can both be **faster** than weak scaling alternatives while **reliably converging to a solution**.*

Why are we doing this?

Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour

Priya Goyal Piotr Dollár Ross Girshick Pieter Noordhuis
Lukasz Wesolowski Aapo Kyrola Andrew Tulloch Yangqing Jia Kaiming He

Facebook

Scaling SGD Batch Size to 32K for ImageNet Training

Yang You [◇] **Igor Gitman** [*] **Boris Ginsburg**[†]
[◇] UC Berkeley [*] CMU [†] NVIDIA
youyang@cs.berkeley.edu igitman@andrew.cmu.edu bginsburg@nvidia.com

PowerAI DDL

Minsik Cho, Ulrich Finkler, Sameer Kumar, David Kung, Vaibhav Saxena, Dheeraj Sreedhar

IBM Research

DON'T DECAY THE LEARNING RATE, INCREASE THE BATCH SIZE

Samuel L. Smith*, **Pieter-Jan Kindermans***, **Chris Ying & Quoc V. Le**
Google Brain
{slsmith, pikinder, chrisying, qvl}@google.com

- These are all very good papers
- With relatively large datasets


Harvard Business Review
GETTING CONTROL OF BIG DATA
BIG DATA
nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT
HTTP://WWW.ANALYTICS-MAGAZINE.ORG
Analytics
DRIVING BETTER BUSINESS DECISIONS
NOVEMBER/DECEMBER 2012
DRILLING

Incentives

- It makes sense to study large models with large machines
  - Interesting challenges
  - Has greater appeal
    - Is good marketing

Natural questions

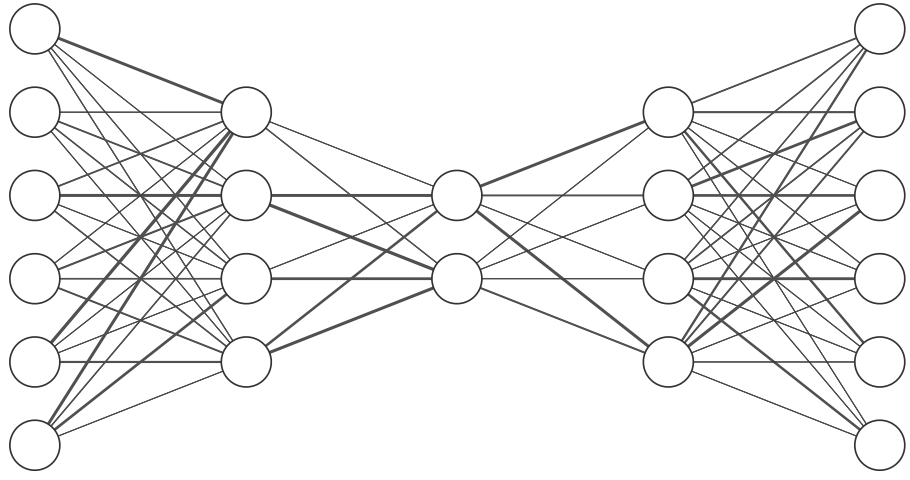
- What if you're working with smaller datasets?
- Do the techniques shared translate to them?
- Are we, as a community, "overfitting" to large datasets?

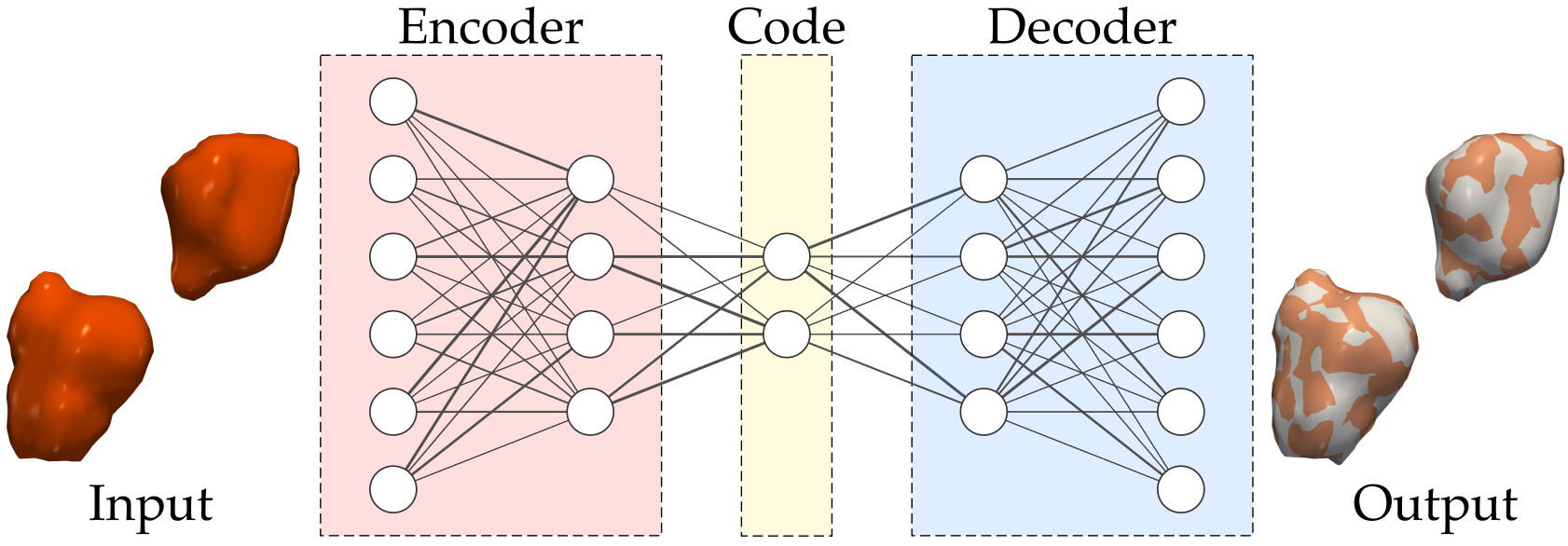
- Does it even make sense to work with small datasets?
- Yes! Many applications rely on hard-to-obtain data
  - Regulatory issues
  - Natural resources
  - Medical imaging

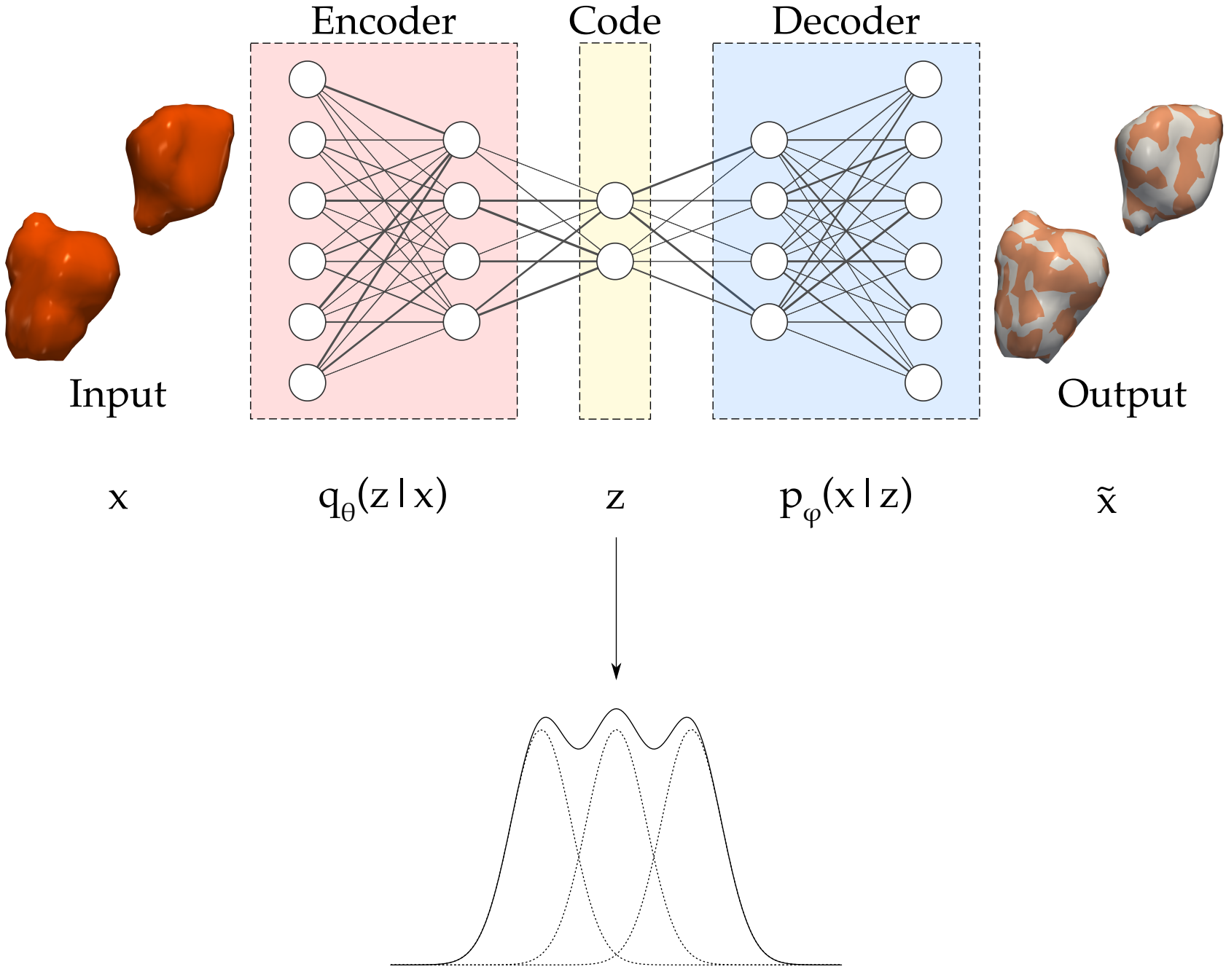
- Isn't Deep Learning data hungry?
  - Yes, it is
- Aren't models that operate on small data fast?
  - Not necessarily
  - Also depends on the complexity of the model

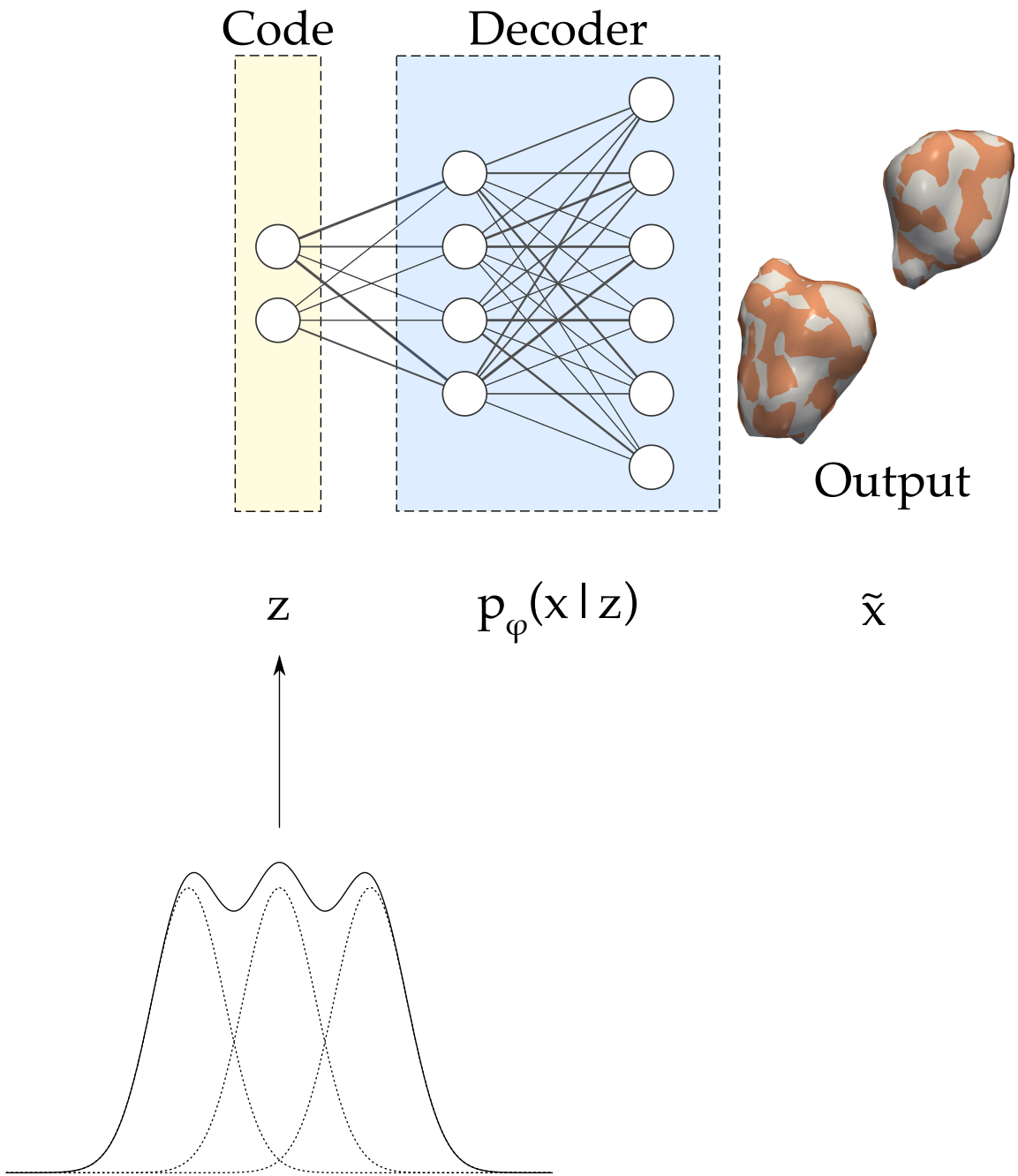
Let's investigate one such model

- Learns the dataset characteristics
  - 888 annotated CT scans
  - Non-nodules (< 3 mm) and nodules (≥ 3mm)
- Takes a week to train on a single GPU (k80)
- Why this model?
  - 3D convnets provide state of the art results
  - But the data in this dataset is not enough



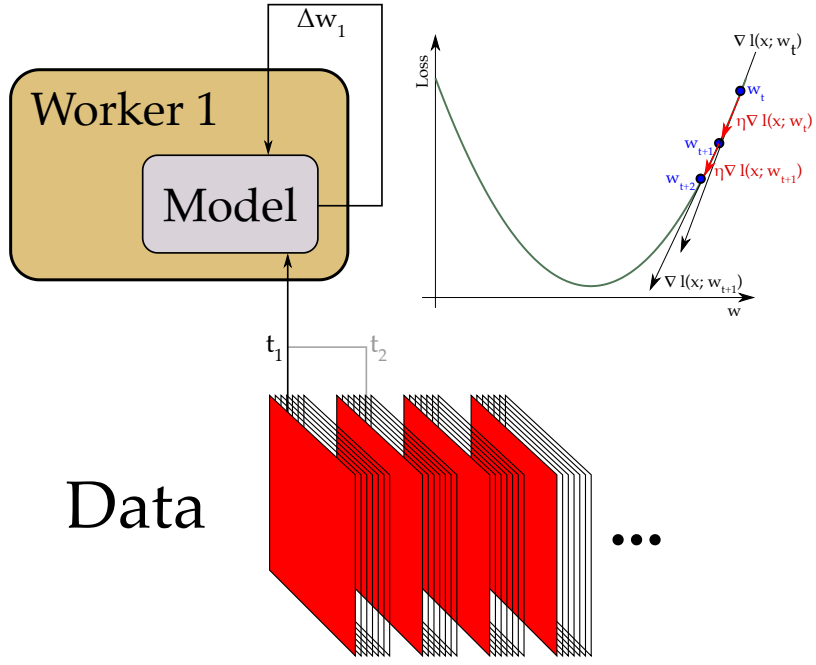

Encoder
Code
Decoder
Input
Output
x
$q_\theta(z \mid x)$
z
$p_\varphi(x \mid z)$
$\tilde{x}$


Encoder
Code
Decoder
Input
Output
x
$q_\theta(z \mid x)$
z
$p_\varphi(x \mid z)$
$\tilde{x}$


Code
Decoder
Output
z
$p_\varphi(x \mid z)$
$\tilde{x}$

We train these networks exactly the same way we train "regular" ones

- Update weights w_t *iteratively* by
  - Sampling entries x from a mini-batch $\mathcal{B}$
  - Computing the loss and the derivatives for each layer
  - Taking a small step towards $\eta \nabla l(x, w_t)$

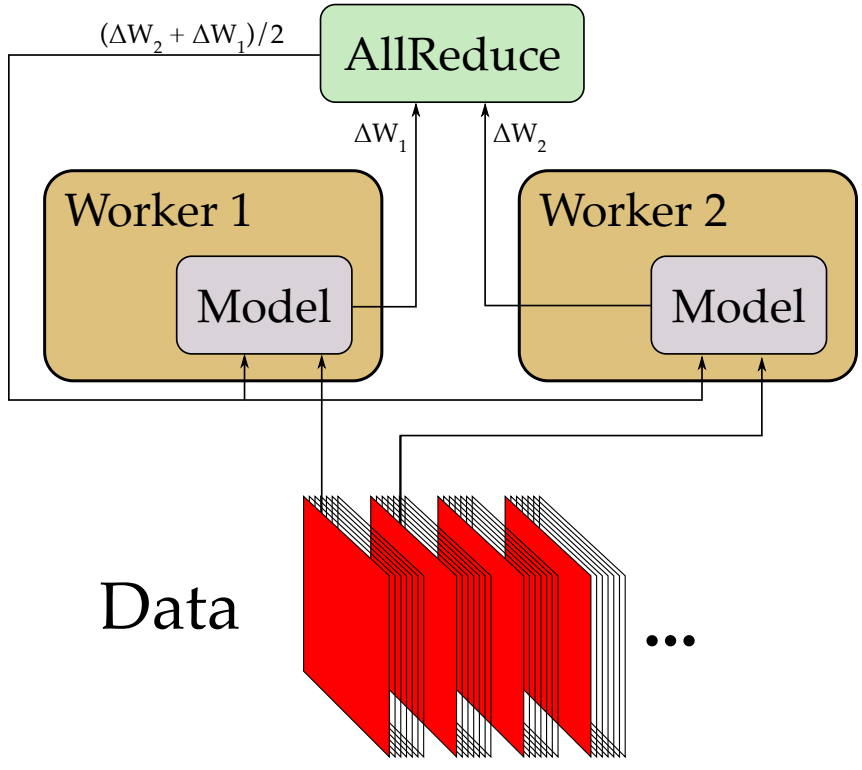


$$w_{t+1} = w_t - \eta \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \nabla l(x, w_t)$$

This is easily parallelizable, because $x \in \mathcal{B}$ are (assumed) *i.i.d.*

All we need is:

- Broadcast
- AllReduce


$(\Delta W_2 + \Delta W_1)/2$
AllReduce
ΔW_1
ΔW_2
Worker 1
Model
Worker 2
Model
Data
...

How we split our work is important

Many papers and frameworks suggest:

- Increase learning rate by k whenever batch size is increased by k
- Optimization *will* take different paths in the loss landscape
  - But can only follow the same path if $k\eta$ is used

$$w_{t+1} = w_t - k\eta \frac{1}{|\mathcal{B}'|} \sum_{j<k} \sum_{x \in B_j} \nabla l(x, w_t)$$

$$\mathcal{B}' = \cup_j B_j$$

$$|B_j| = |\mathcal{B}|$$

$$w_{t+1} = w_t - \alpha \frac{1}{|\mathcal{B}'|} \sum_{j<k} \sum_{x \in B_j} \nabla l(x, w_t)$$

Obviously, the term α directs optimization

- When $\alpha = k\eta$: linear scaling rule

```
# Horovod: adjust learning rate based on number
# of GPUs.
opt = keras.optimizers.Adadelta(1.0 * hvd.size())
```

- When $\alpha \rightarrow k\eta$: linear scaling rule with warmup

```
# Horovod: using `lr = 1.0 * hvd.size()` from the very
# beginning leads to worse final accuracy
hvd.callbacks.LearningRateWarmupCallback(warmup_epochs=5, verbose=1),
```

What *we* do is keep $|\mathcal{B}|$ fixed and $\alpha = \eta$:

- If we have two workers, each gets half the size of the original minibatch
- If we have three, each gets one third
- And so on...

Since batch size is an integer, this limits the number of parallel workers to $|\mathcal{B}|$

This gives us the same convergence properties of the single worker case

Our models run faster, since work is divided between workers

- Our approach: strong scaling
  - Work per iteration is the same no matter number of workers
- Approaches in the literature: weak scaling
  - Each worker performs the same amount of work
  - Per iteration, data processed increases with number of workers

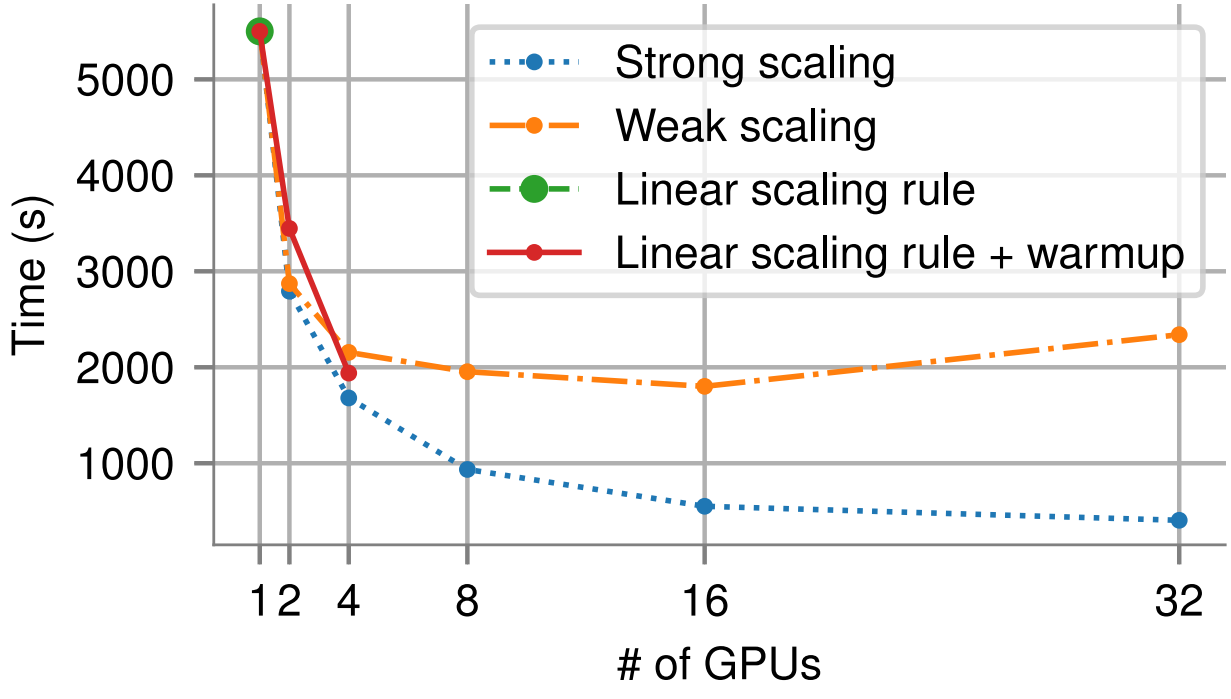
Methodology

- We trained a generative network
- with {1, 2, 4, 8, 16, 32} workers
- on a dataset of 888 annotated CT scans with coordinates of 2100+ nodules > 3mm
- until it reached a **given loss** (MSLE)
- with different scaling rules
  - Strong scaling
  - Weak scaling
  - Weak scaling with linear scaling rule
  - Weak scaling with linear scaling rule + warmup

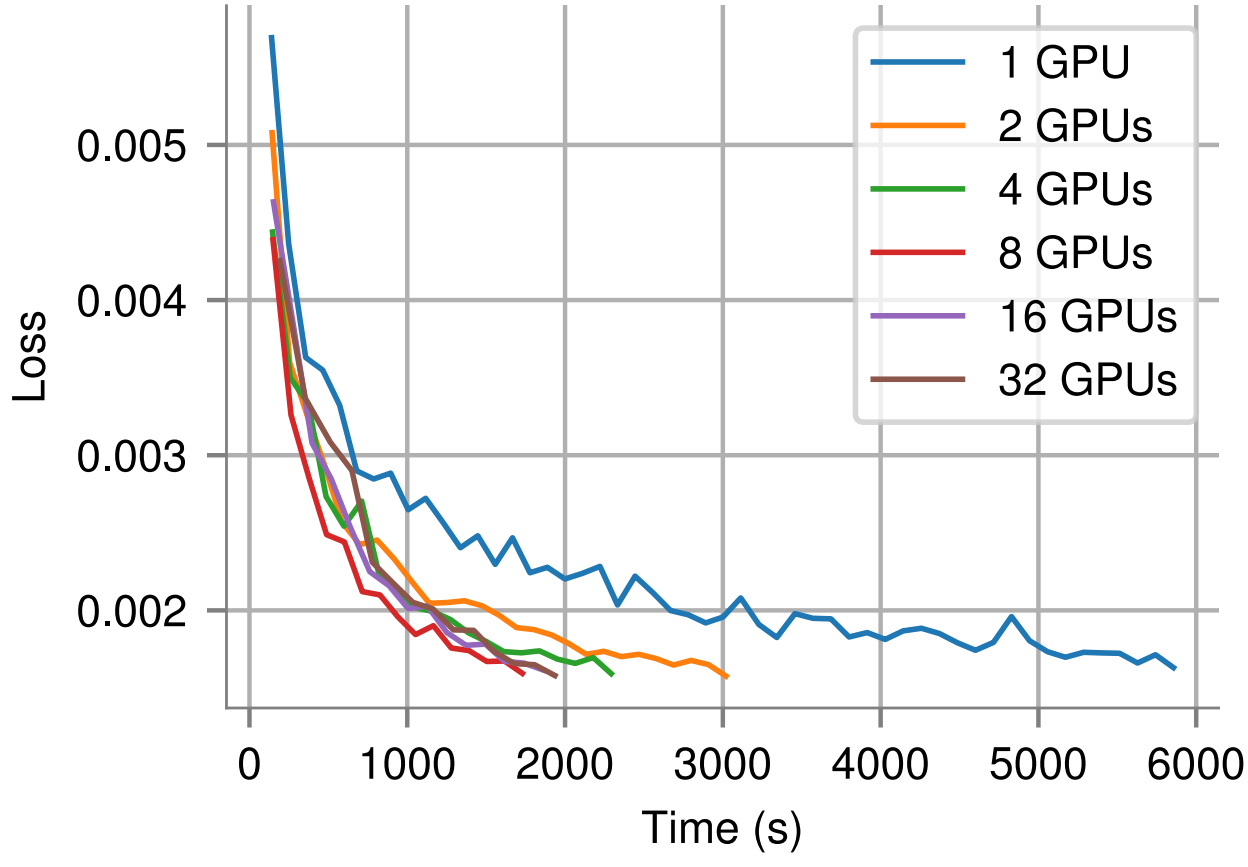
Why time to loss?

Results

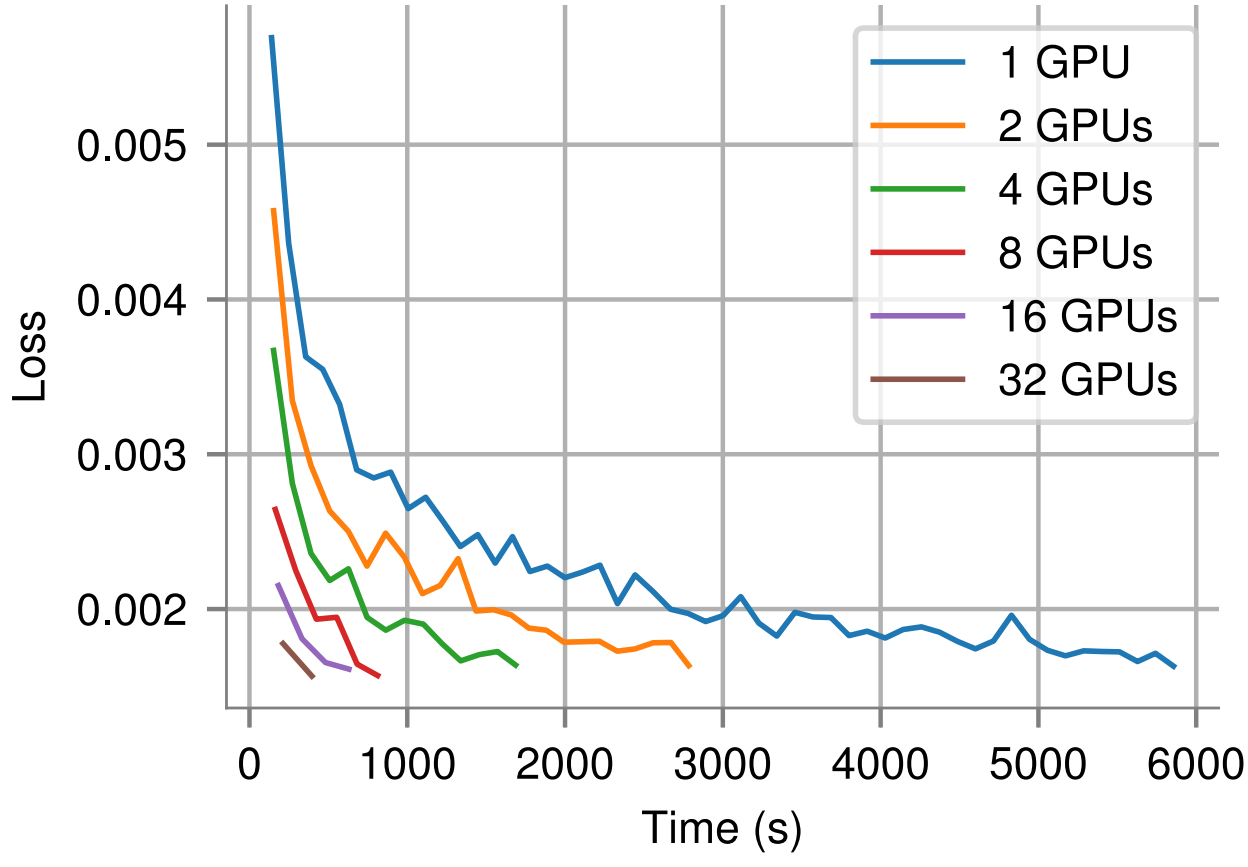
Strong scaling has better performance



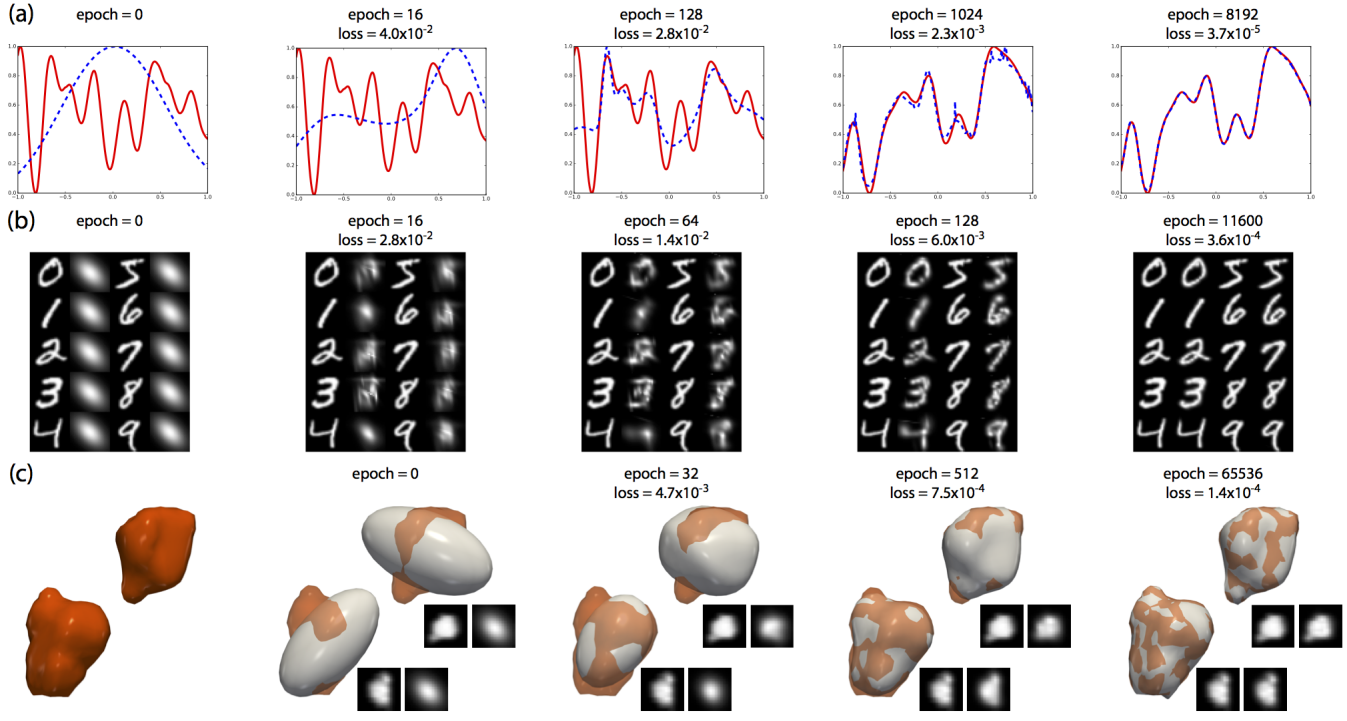
Behavior of the loss with weak scaling



Behavior of the loss with strong scaling



Progress of the model



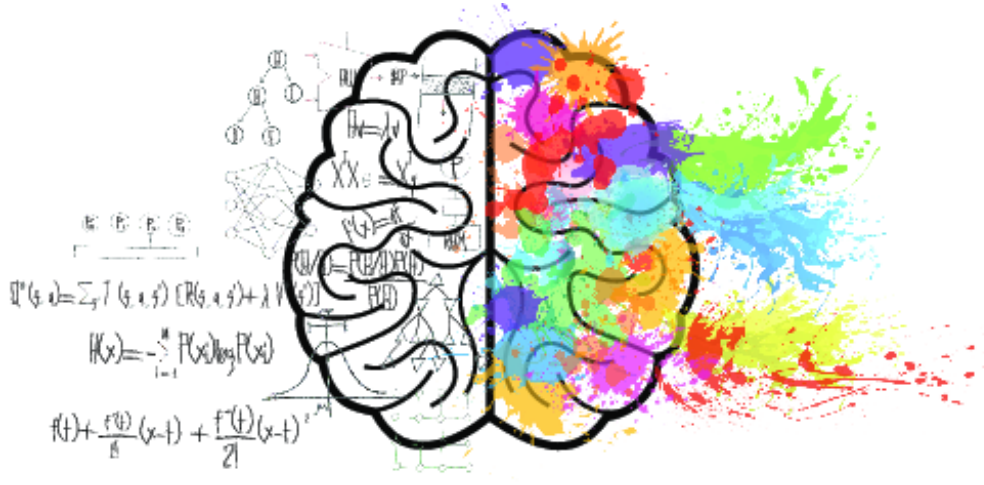
Conclusions

- Insights obtained in large datasets do not necessarily translate to smaller ones
- Models with modest data still can take a long time to train
- Strong scaling outperforms weak scaling in the application tested
- We expect these findings to generalize to other applications
  - More experimentation is needed

Thank you! Questions?

Renato L. de F. Cunha, Eduardo R. Rodrigues, Matheus Viana, Dario Borges

<renatoc@br.ibm.com> | renatocunha | renatolfc | renatolfc



HPML 2018 - 2018-09-24