Alleviating the data sparsity issue in deep discriminative models

HUNIMAT Research Proposal

Dániel Varga

December 20, 2017

Hungarian Academy of Sciences
Alfréd Rényi Institute of Mathematics
Deep Learning Group

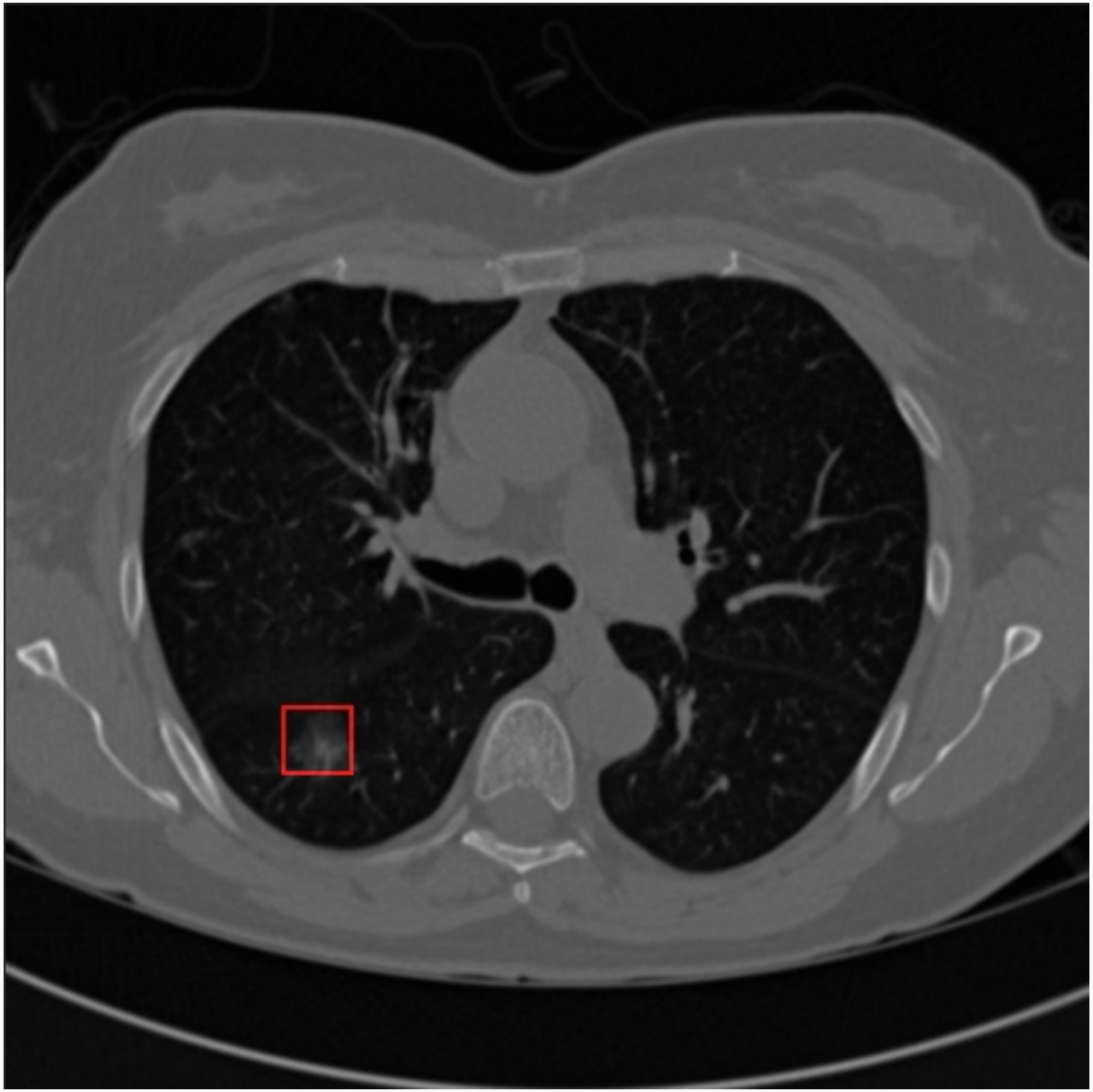
Team

Rényi Deep Learning Group

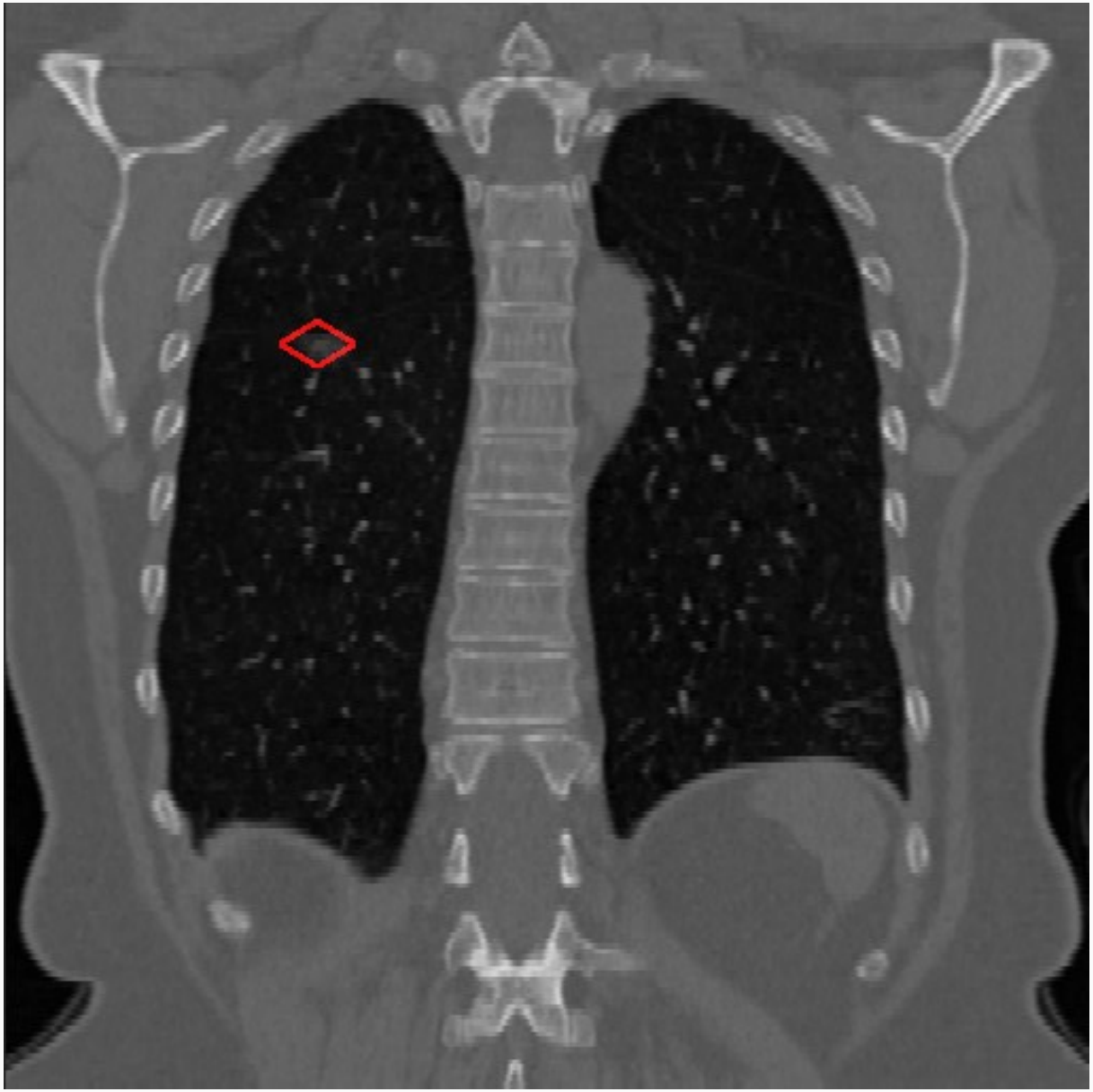
- Balázs Szegedy PhD, Lendület Grantee, ERC grantee
- Dániel Varga PhD
- Zsolt Zombori PhD, junior researcher
- Adrián Csiszárik PhD Student, junior researcher

Background

Data can be very expensive



Data can be very expensive



Data can be very expensive

- Reliably detecting early stage lung cancer (<4mm) requires CAT scanning, X-ray is not sufficient.
- 300-500 slices per scan, very time consuming manual process, requires high level of expertise.
- US National Lung Screening Trial had 450 false negatives in 45000 samples.

Data can be very expensive

- This laborious, expensive and error-prone process is what our industry partner MedInnoScan plans to replace with **deep learning technology**, saving human lives.
- But to train their deep learning systems, laborious and expensive **manual annotation** is required.
- Saving on the required amount of annotated training data can make or break the project.

Regularization

Regularization: Any change to a machine learning model that makes it generalize better to unseen data. In the context of artificial neural networks:

- L2 weight decay (Plaut et al 1986)
- dropout (Srivastava et al 2014)
- batch normalization (Ioffe-Szegedy 2015)
- label smoothing (Szegedy et al 2015)
- ...and hundreds of other techniques

Smoothness prior

- The core idea behind our project is to enforce a **smoothness** property of the neural network via a well-chosen regularization term.
- Tiny perturbations of the network input should not lead to large changes in network output.
- This idea was independently rediscovered several times, most recently by our team. The earliest known reference is Drucker and LeCun's double backpropagation from 1991.
- But to the best of our current knowledge, we are the first to report that significant accuracy improvements can be achieved on vision tasks when the amount of data is restricted.

Smoothness prior

We can formalize local smoothness in different ways:

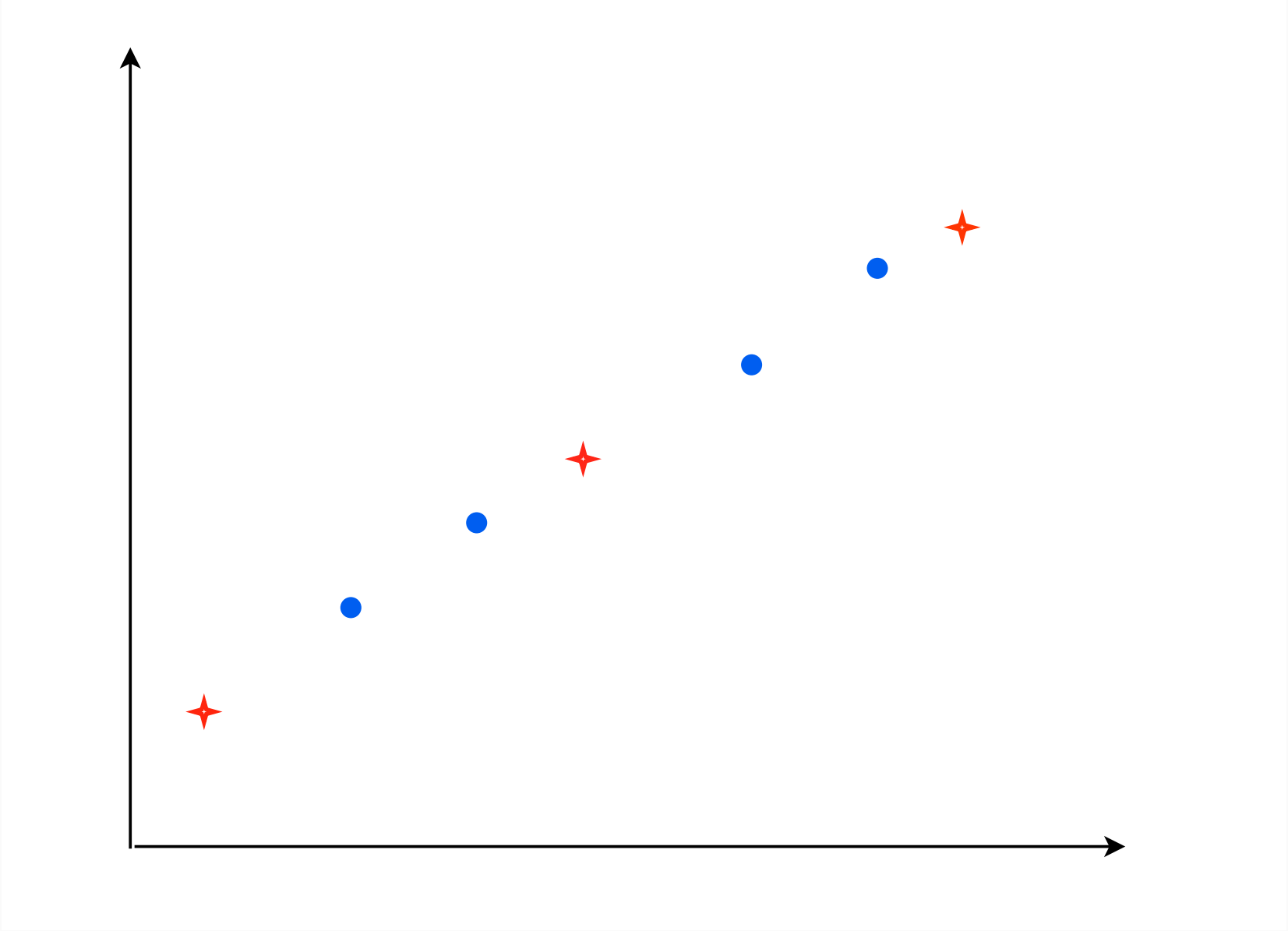
- Our *Spectral Regularizer* (SpectReg) approximates the Frobenius norm of the Jacobian of the input-logit mapping at the training examples:

$$L_{SpectReg}(x, \Theta) = \| \frac{\partial}{\partial x} f_\Theta(x) \|_F^2 = \mathbb{E}_{r \sim \mathcal{N}(0, I^m)} [\| \frac{1}{\sqrt{m}} r^T \frac{\partial}{\partial x} f_\Theta(x) \|_2^2]$$

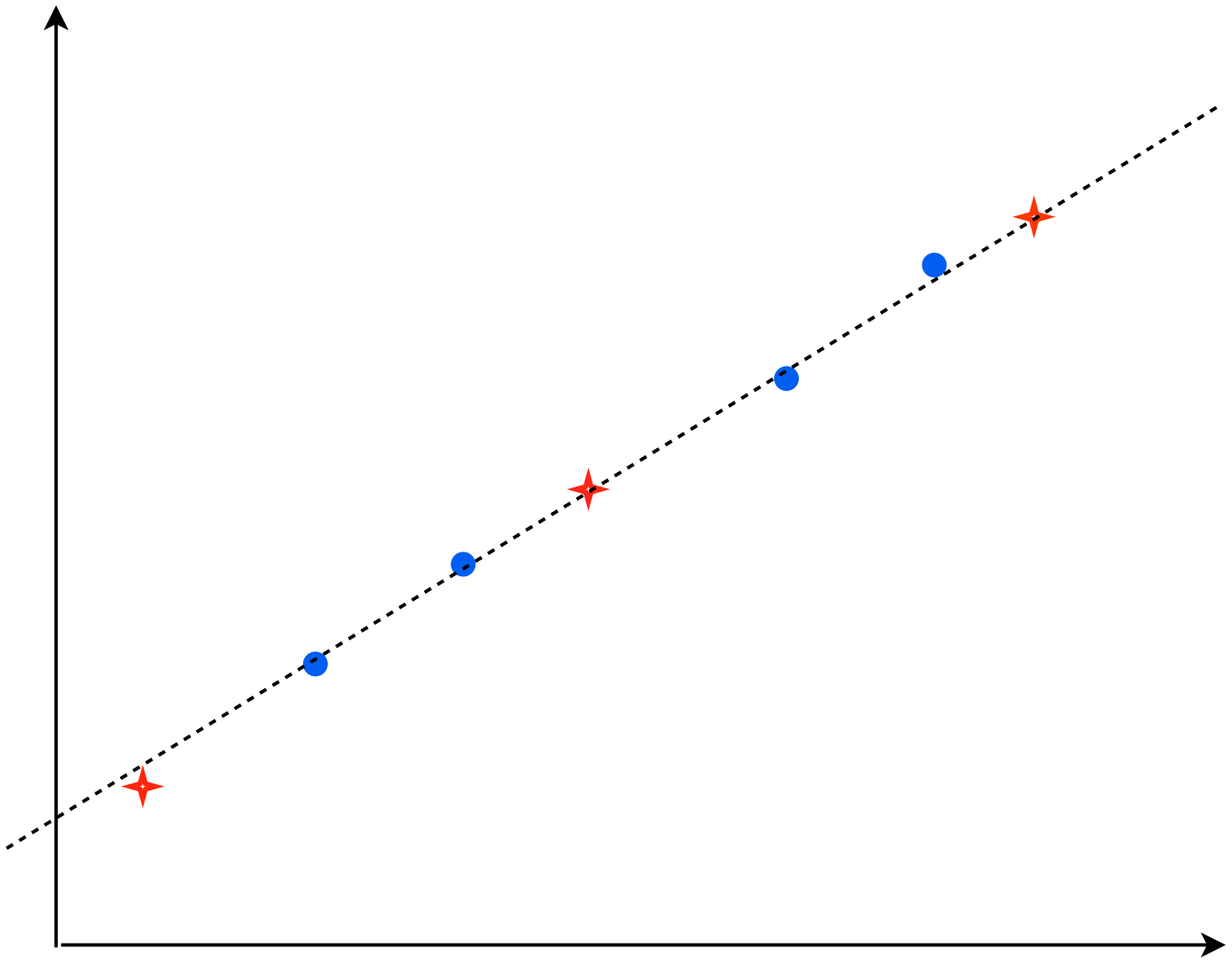
- The *DataGrad* regularizer penalizes large changes of the loss function at the training examples:

$$L_{DataGrad}(x, y, \Theta) = \| \frac{\partial}{\partial x} L(f_\Theta(x), y) \|_2^2$$

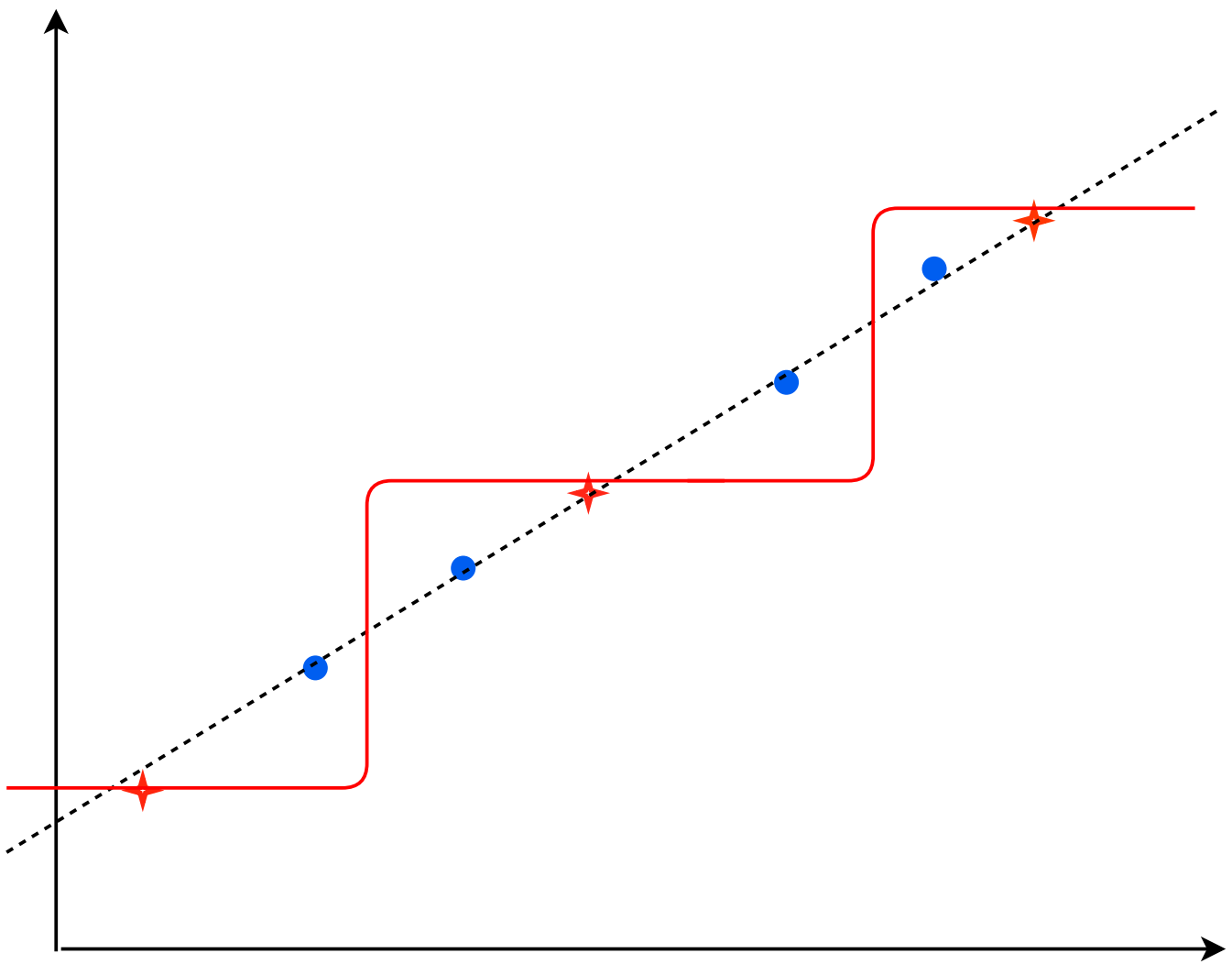
Wait, this is not supposed to work!



Wait, this is not supposed to work!



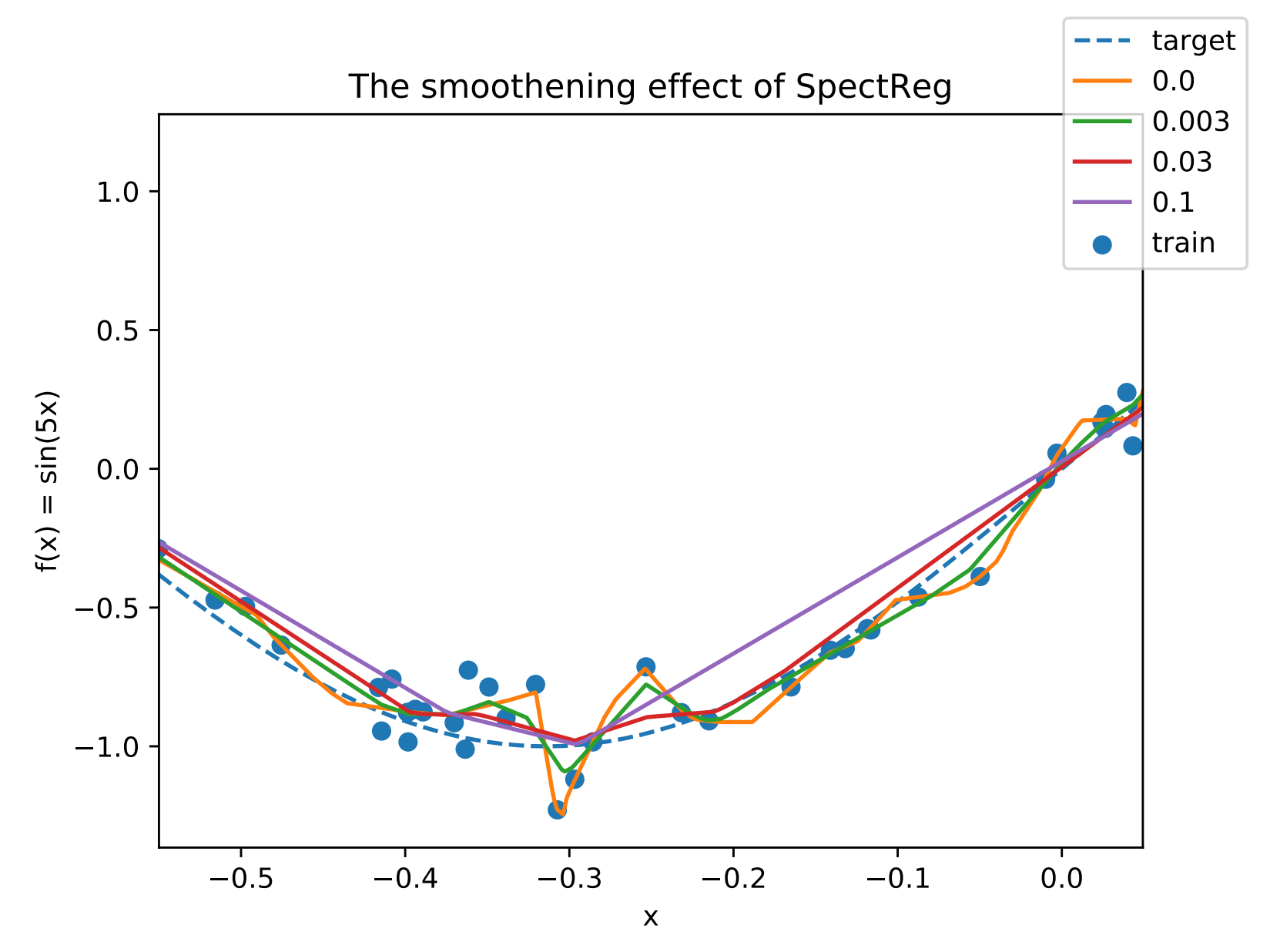
Wait, this is not supposed to work!



...But it does work.

- We are still in the process of understanding the phenomenon better,
- ...but apparently the low-dimensional intuition of the previous slides does not generalize to complex high-dimensional loss surfaces,
- ...especially not when our method of discovery of these surfaces is the gradient descent.
- Gradient descent does not converge to step function-like solutions such as seen on the previous slide.
- It is probably better to think of gradient regularization as "smarter weight decay". It influences the gradient norm at points far away from the training dataset.

...But it does work.



Experimental Results

MNIST – training on 2000 randomly chosen samples, interaction with weight decay

Weight decay	NoGR	SpectReg	DataGrad
LeNet			
no WD	97.15	97.55 ($\lambda = 0.03$)	**97.93** ($\lambda = 20$)
WD=0.0005	97.32	97.67 ($\lambda = 0.05$)	**97.93** ($\lambda = 50$)

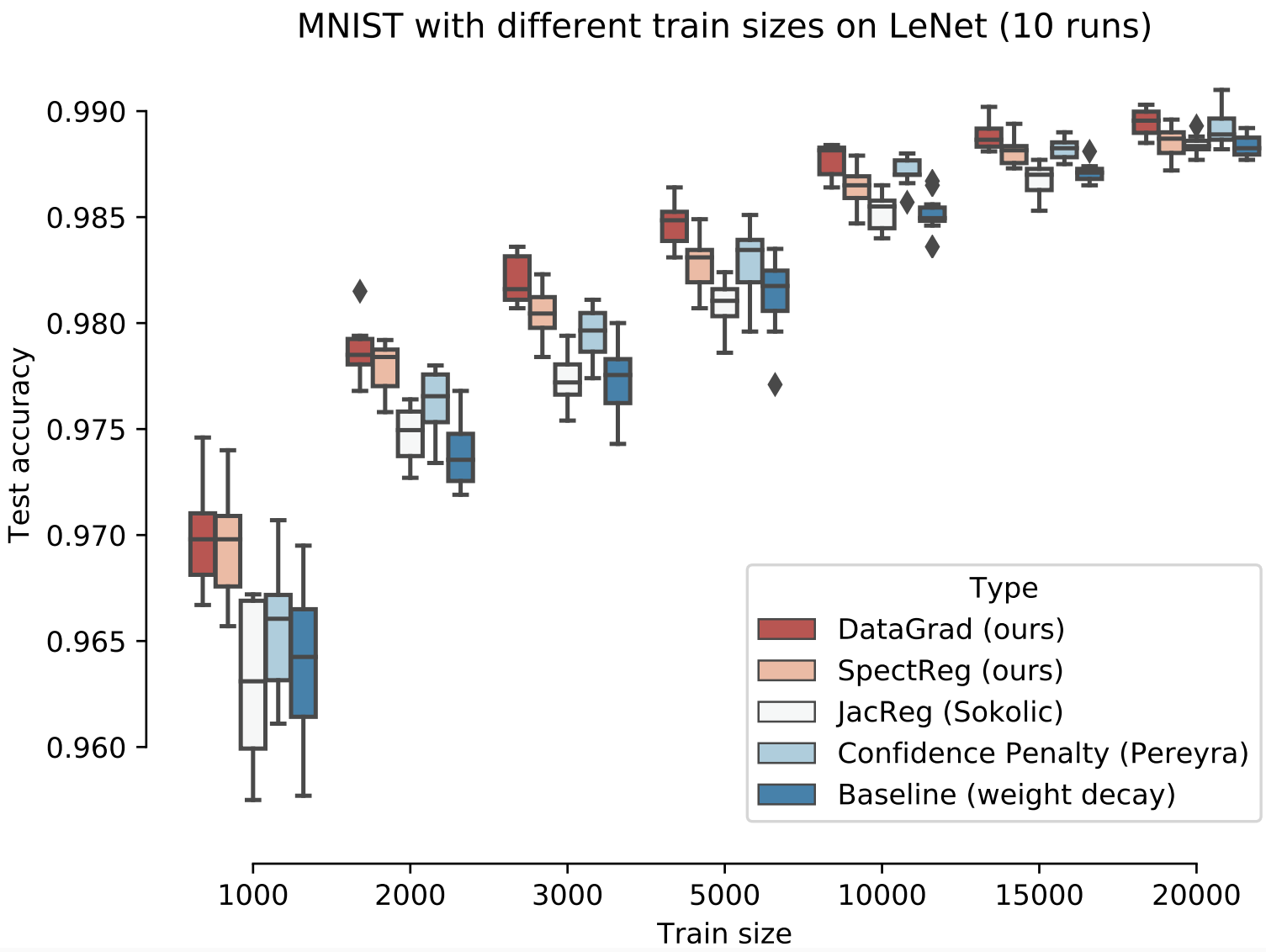
Dropout is on in all of these runs, dropout rate 0.5.

Gradient regularization's interaction with DO and BN

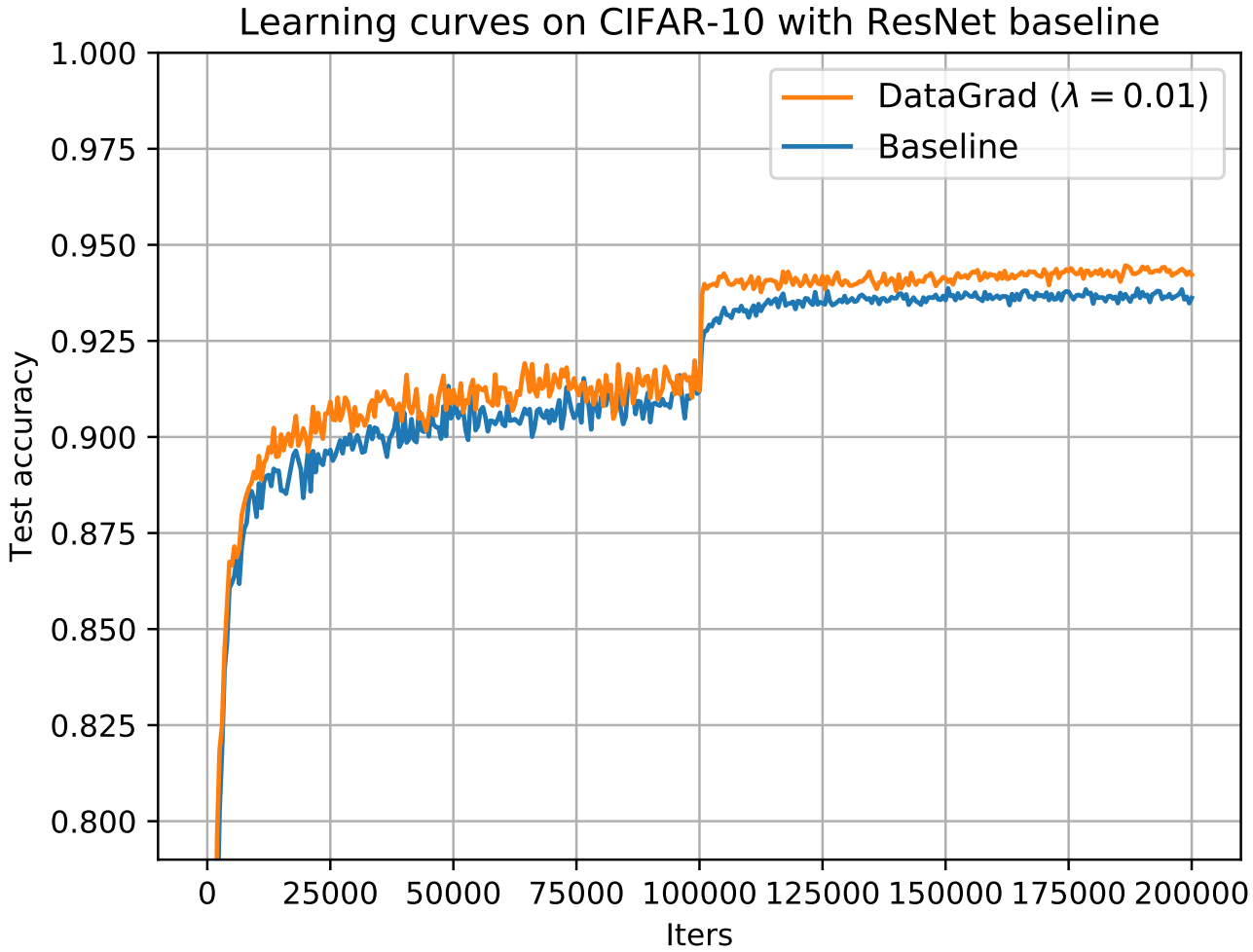
	NoGR	SpectReg	DataGrad
LeNet unreg	96.99	*97.59*	97.56
LeNet BatchNorm	96.89	*96.94*	96.89
LeNet Dropout	97.29	97.67	**97.93**

Comparison of dropout, batch normalization and two variants of gradient regularization: symbolic DataGrad and SpectReg. Train size was set to 2000. Each hyperparameter was tuned individually on a development set.

Comparison of various regularization methods on MNIST



DataGrad learning curve on full CIFAR-10



Improvements even on full CIFAR-10 with data augmentation.

Research plan

- Better understanding the behavior of these regularizers is a promising research project.
- Scaling up these results to high resolution images is in the works.
- We hope to be able to reduce the necessary amount of training data for annotation-heavy tasks like **lung cancer detection**.

Deliverables

- We commit to publish two papers on our methods as conference papers at prestigious deep learning conferences. (This is the preferred method of publication in the field of deep learning.)
- We open source all our neural network code and experiments.
- We commit to evaluate the performance of our method class on the problem of lung cancer detection.