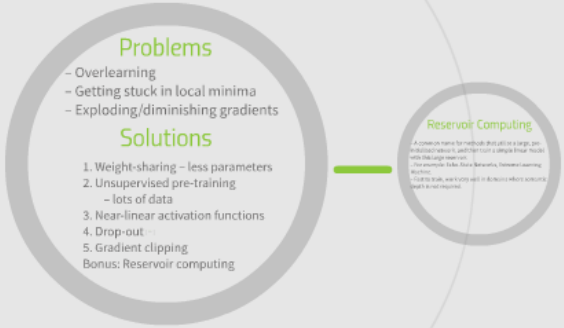
Deep Learning

Learning High-degree Non-linear Models:
Enabling factors, tips and tricks

Deep Learning



Learning High-degree Non-linear Models:
Enabling factors, tips and tricks

Deep Learning

Learning High-degree Non-linear Models: Enabling factors, tips and tricks

Problems

- Overlearning
- Getting stuck in local minima
- Exploding/diminishing gradients

Solutions

1. Weight-sharing – less parameters
2. Unsupervised pre-training
   - lots of data
3. Near-linear activation functions
4. Drop-out [redacted]
5. Gradient clipping

Bonus: Reservoir computing

Weight sharing / Regularization

- Reducing the number of parameters or degrees of freedom of the model is regularization. Regularization prevents overfitting.
- In Convolutional Neural Networks, the weights are identical for each window, and window outputs are max-pooled to the next layer. This exploits the translational symmetry of the domain, and is often used for images.
- Regularization can also be done by introducing a loss term for weights, L2 norm is often used.

Problems

- Overlearning
- Getting stuck in local minima
- Exploding/diminishing gradients

Solutions

1. Weight-sharing – less parameters
2. Unsupervised pre-training
   - lots of data
3. Near-linear activation functions
4. Drop-out [redacted]
5. Gradient clipping

Bonus: Reservoir computing

Unsupervised pre-training

- Unsupervised pre-training makes use of a lot of unlabeled data, learning the overall structure of the domain, which can be fine-tuned with backpropagation for specific labels at a later stage. More data, less labels.
- This could be thought as "pre-training as regularization", or approaching neural reinforcement learning.
- Free pre-trained models are available from Google and Oxford University. Mostly convolutional nets trained on images and video.
- The models are taught using for example greedy DBN Contrastive Divergence or other autoencoding methods.

http://www.jmlr.org/papers/volume11/erhan10a/erhan10a.pdf
http://www.vlfeat.org/matconvnet/pretrained/

Problems

- Overlearning
- Getting stuck in local minima
- Exploding/diminishing gradients

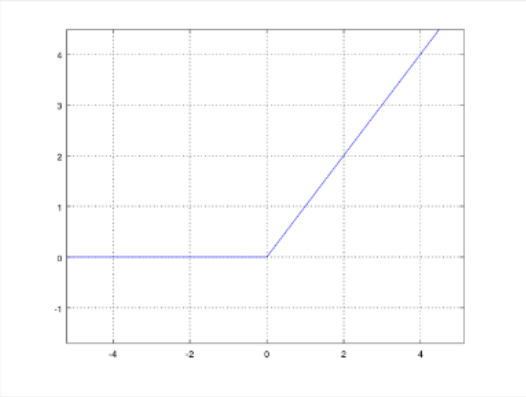
Solutions

1. Weight-sharing – less parameters
2. Unsupervised pre-training
   - lots of data
3. Near-linear activation functions
4. Drop-out [redacted]
5. Gradient clipping

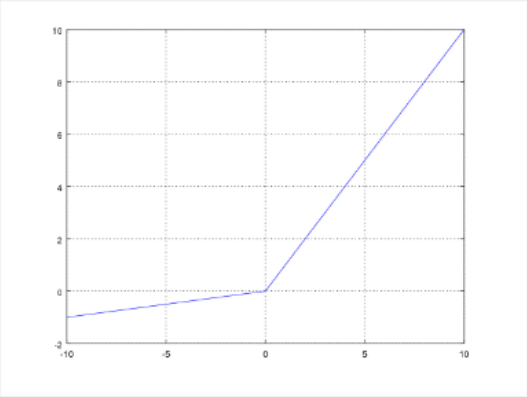
Bonus: Reservoir computing

Near-linear Activation Functions

- Deep neural networks are effective because of non-linear activation functions – Stacking linear layers only reduces to a single linear layer.
- However, non-linearities are challenging to learn. Minimizing the non-linearity has been shown to be really useful. Deep neural networks can be trained without unsupervised pre-training if rectifier activation functions are used. (otherwise cannot)



Rectifier: max(0,x)



Leaky rectifier: max(0.1*x,x)

Problems

- Overlearning
- Getting stuck in local minima
- Exploding/diminishing gradients

Solutions

1. Weight-sharing – less parameters
2. Unsupervised pre-training
   - lots of data
3. Near-linear activation functions
4. Drop-out [redacted]
5. Gradient clipping

Bonus: Reservoir computing

Dropout

- Dropout is a very effective strategy for preventing overfitting.
- In practice it means randomly dropping out for example half of the internal representation neurons and their output weights for each training step, and scaling the outputs accordingly.
- This prevents neurons from co-adapting with each other, and learn features more independently. I.e. one neuron learns one feature, not multiple neurons learning one feature.
- It is a very strong method against overfitting, and should be almost always used.
- Analogous to Random Forests.

https://www.cs.toronto.edu/~hinton/absps/JMLRdropout.pdf

Problems

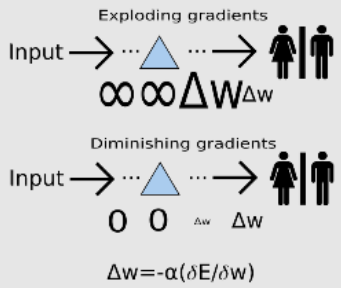
- Overlearning
- Getting stuck in local minima
- Exploding/diminishing gradients

Solutions

1. Weight-sharing – less parameters
2. Unsupervised pre-training
   - lots of data
3. Near-linear activation functions
4. Drop-out [redacted]
5. Gradient clipping

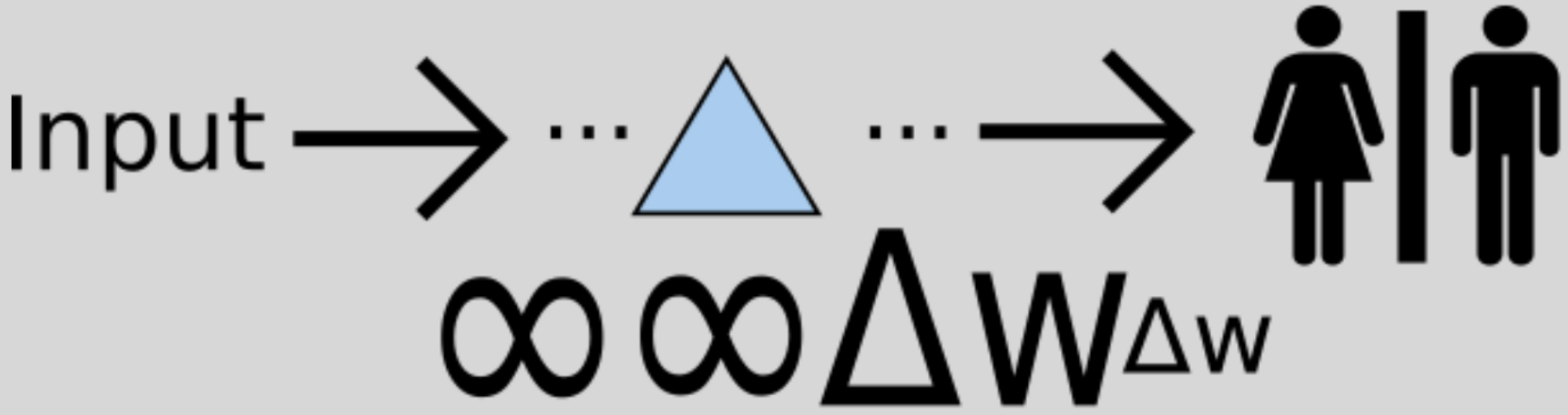
Bonus: Reservoir computing

Gradient Clipping

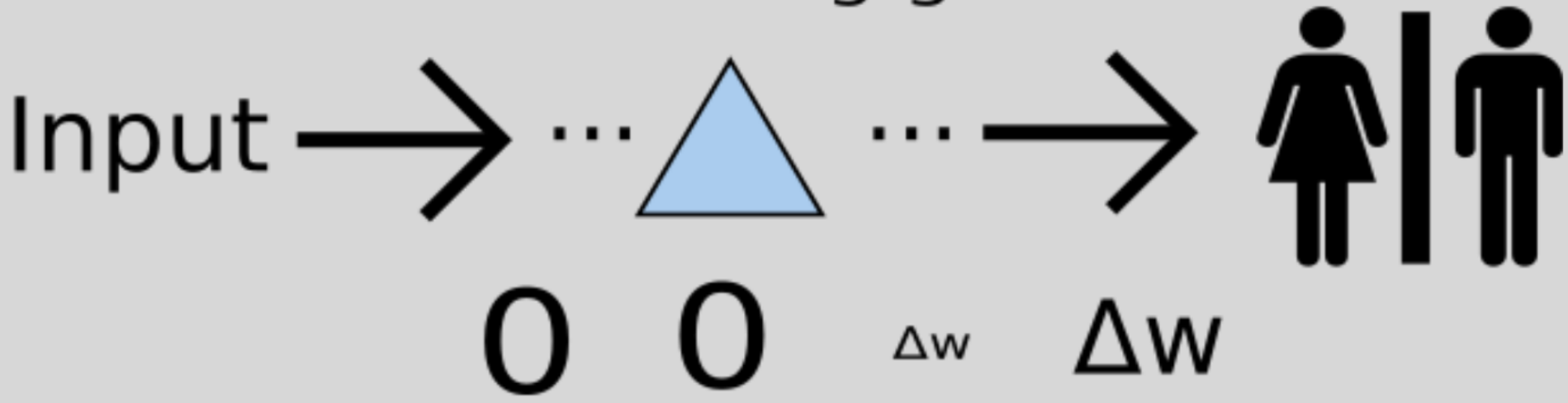


- Gradient explosion happens in backpropagation when the squared error gradients are amplified through activation functions of multiple layers. This typically causes infinities and divergence. Limited activation functions (sigmoid, tanh) are prone to this behavior.
- Gradients are often clipped to some maximum absolute value to prevent infinities.
- Gradients can also diminish to zero, when a small change in internal representation means a too large change in output. If a gradient goes to zero, it means it continues to be zero in backpropagation towards input. Adding noise might help the system in escaping local plateaus.
- Microsoft uses special one-bit gradients only signifying the direction of change, which is always non-zero.

Exploding gradients

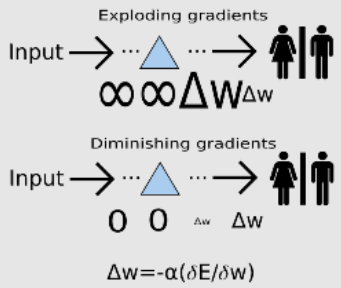


Diminishing gradients



$$\Delta w = -\alpha(\delta E/\delta w)$$

Gradient Clipping

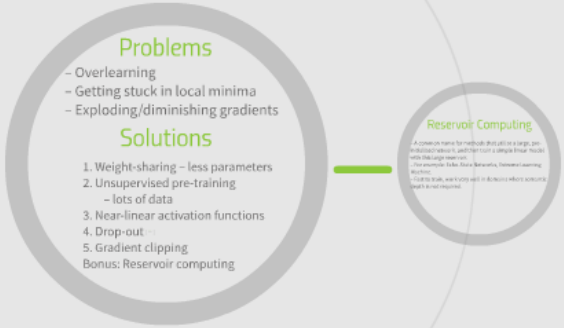


- Gradient explosion happens in backpropagation when the squared error gradients are amplified through activation functions of multiple layers. This typically causes infinities and divergence. Limited activation functions (sigmoid, tanh) are prone to this behavior.
- Gradients are often clipped to some maximum absolute value to prevent infinities.
- Gradients can also diminish to zero, when a small change in internal representation means a too large change in output. If a gradient goes to zero, it means it continues to be zero in backpropagation towards input. Adding noise might help the system in escaping local plateaus.
- Microsoft uses special one-bit gradients only signifying the direction of change, which is always non-zero.

Reservoir Computing

– A common name for methods that utilize a large, pre-initialized network, and then train a simple linear model with this large reservoir.
– For example: Echo-State Networks, Extreme Learning Machine.
– Fast to train, work very well in domains where semantic depth is not required.

Deep Learning



Learning High-degree Non-linear Models:
Enabling factors, tips and tricks