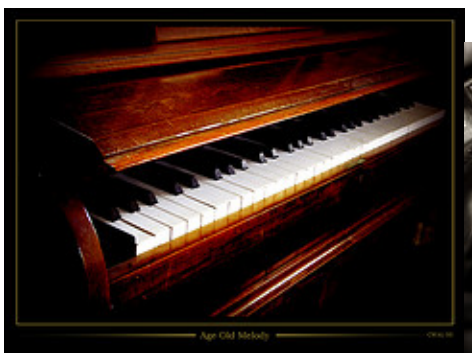
Best Practices for Convolutional Networks

based on *ImageNet Classification with Deep Convolutional Neural Networks* by Krizhevsky, Sutskever and Hinton

Simon Bartels

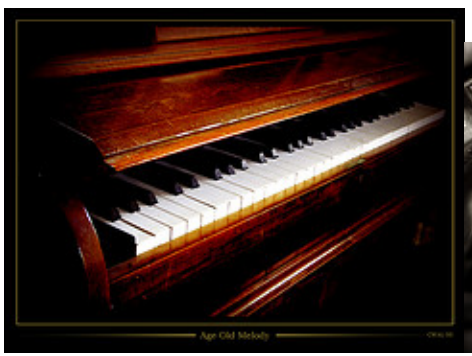
20th November 2013

Motivation



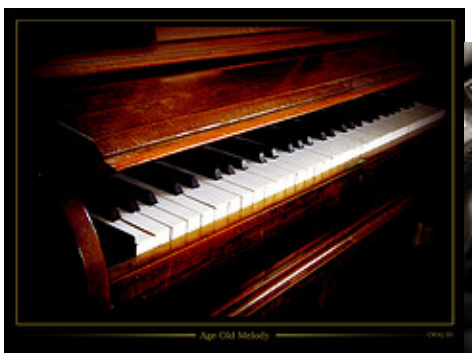

Motivation

- 14 mil. 256×256 coloured images



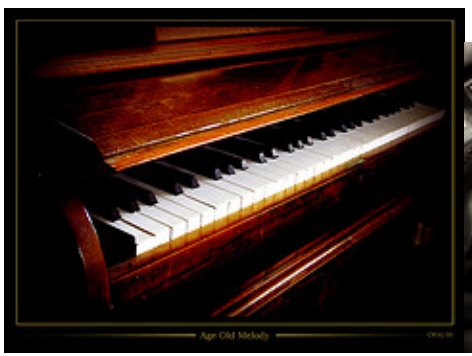
Motivation

- 14 mil. 256×256 coloured images
- 1000 classes



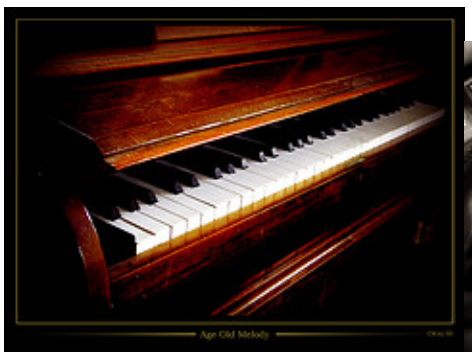
Motivation

- 14 mil. 256×256 coloured images
- 1000 classes
- Large-Scale Visual Recognition Challenge-2010



Motivation

- 14 mil. 256×256 coloured images
- 1000 classes
- Large-Scale Visual Recognition Challenge-2010
- 17% top five error rate

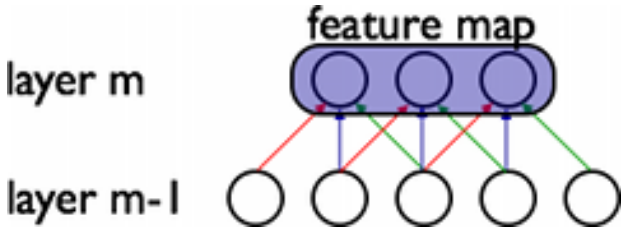


Motivation

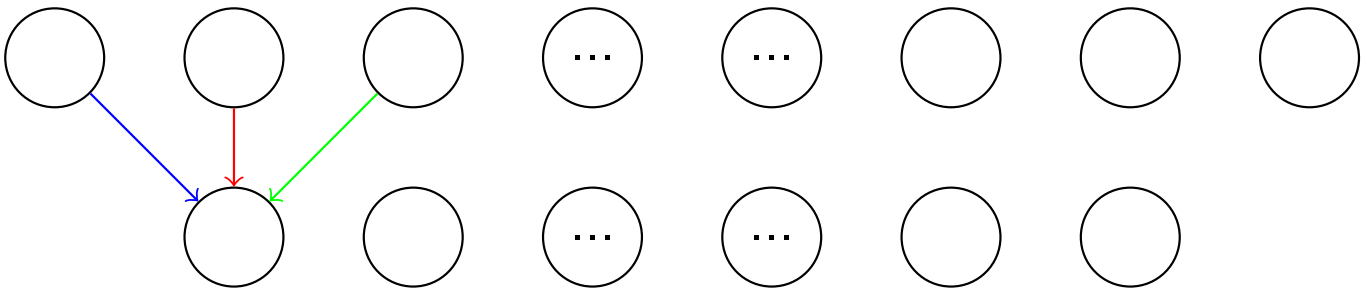
- 14 mil. 256×256 coloured images
- 1000 classes
- Large-Scale Visual Recognition Challenge-2010
- 17% top five error rate
- potential: 0.4% MNIST test error

Road Map

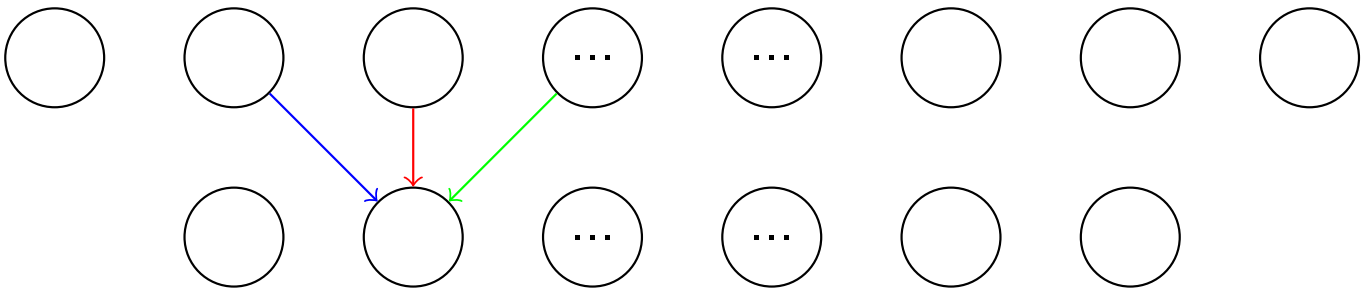
- Convolutional Networks
- Techniques
- Critique



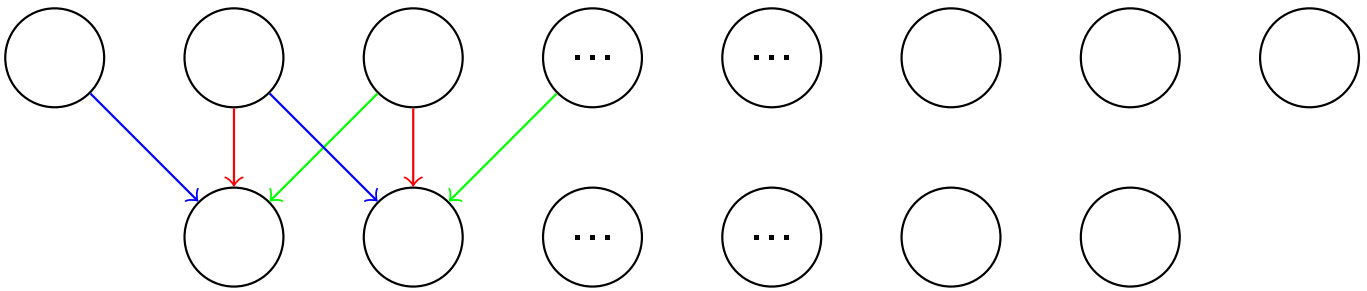
Convolutional Networks



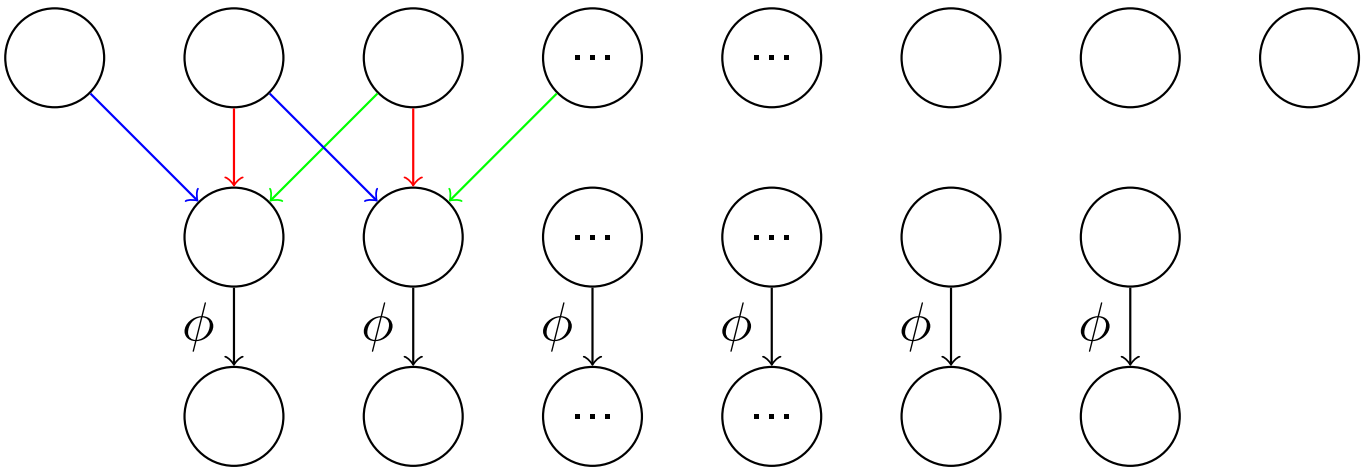
Convolutional Networks



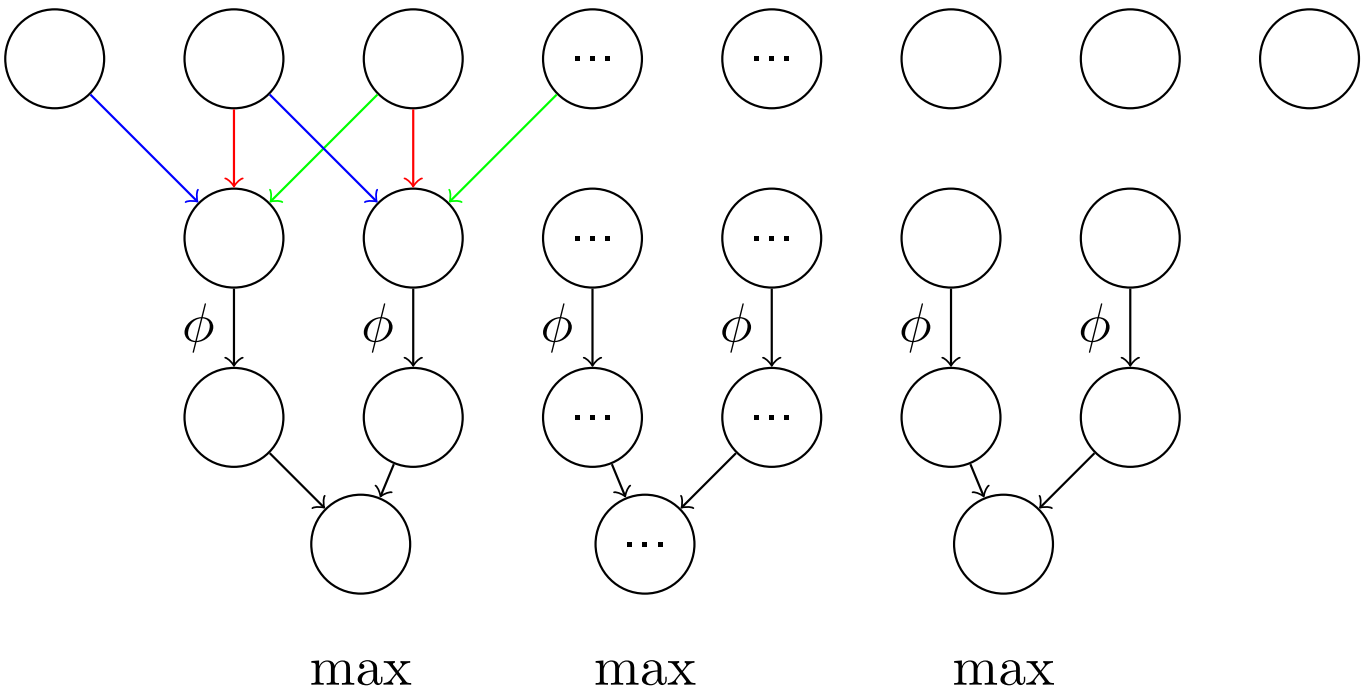
Convolutional Networks



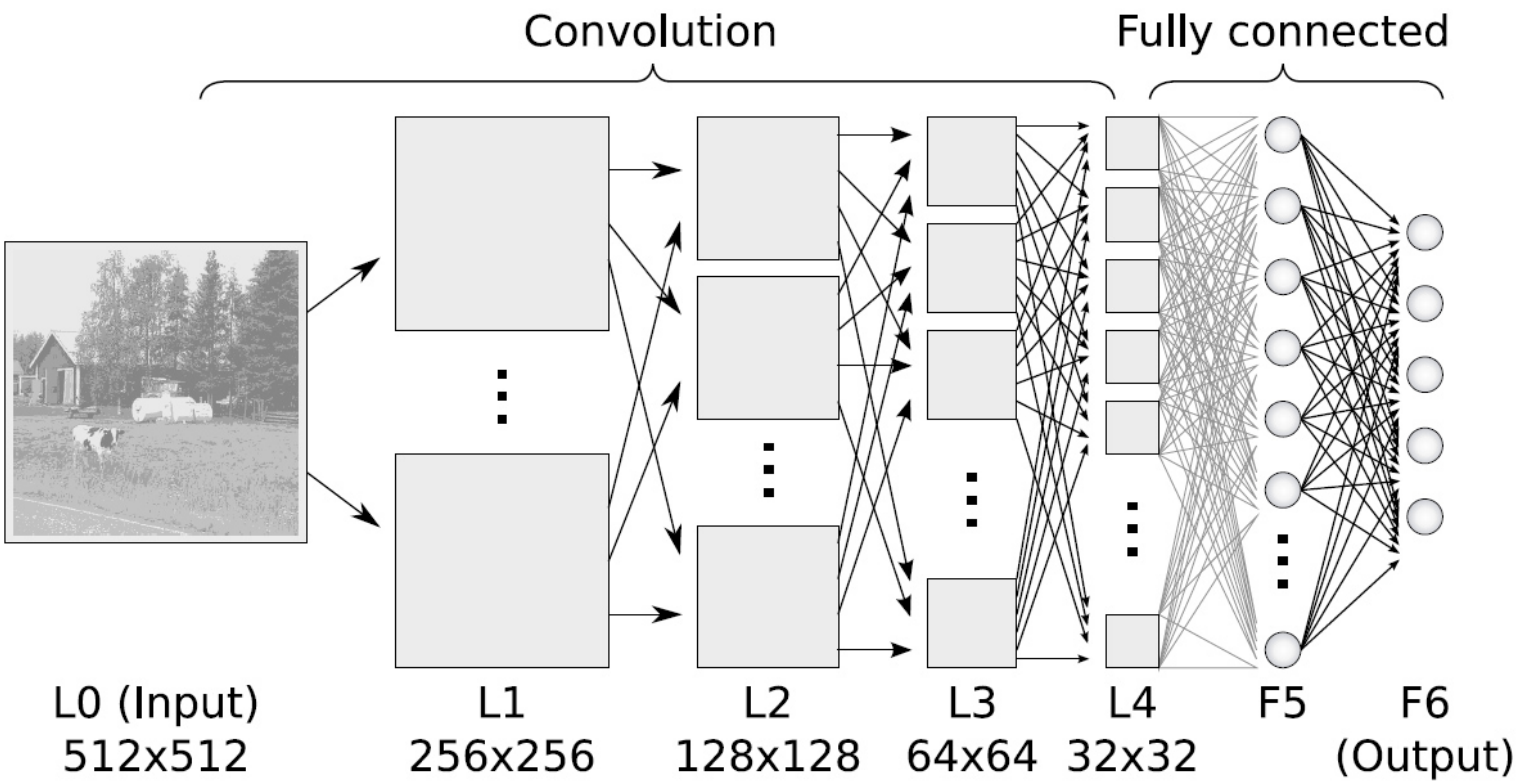
Convolutional Networks



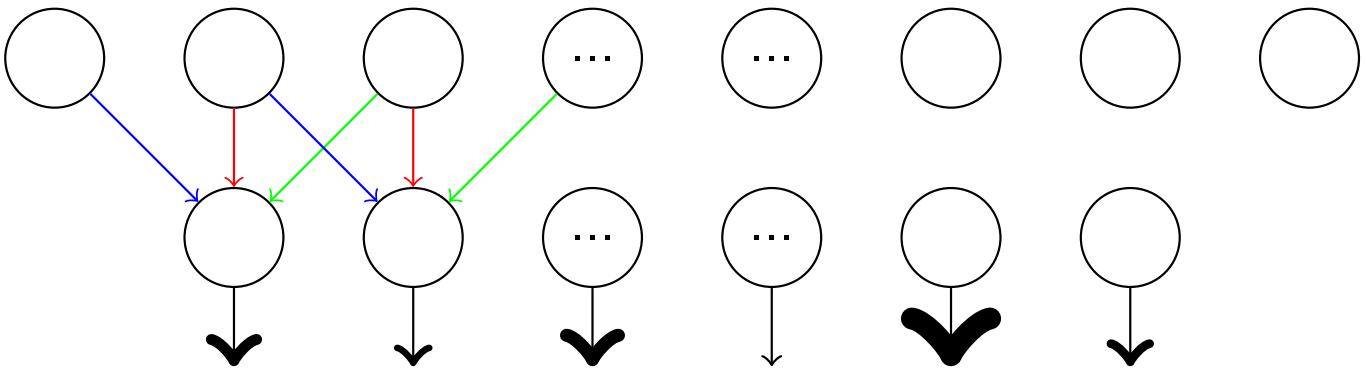
Convolutional Networks



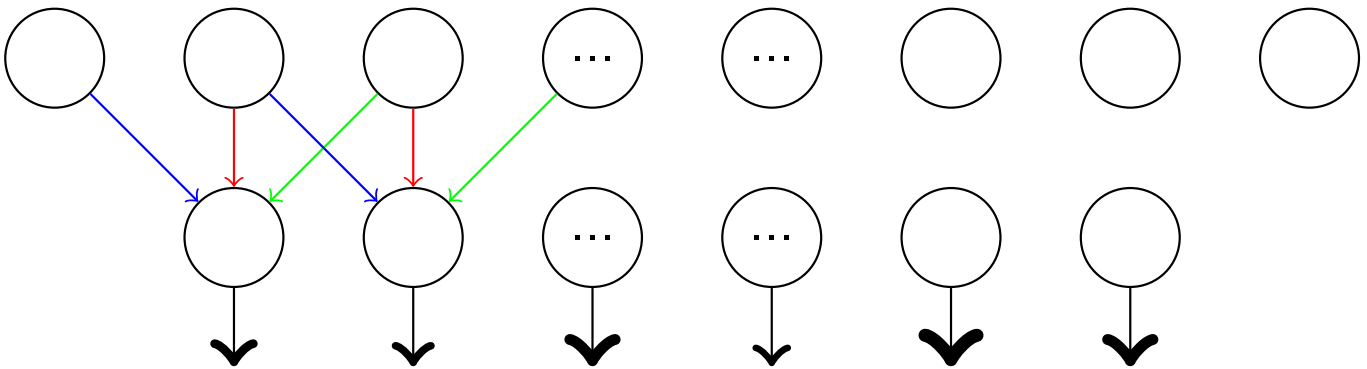
Architecture



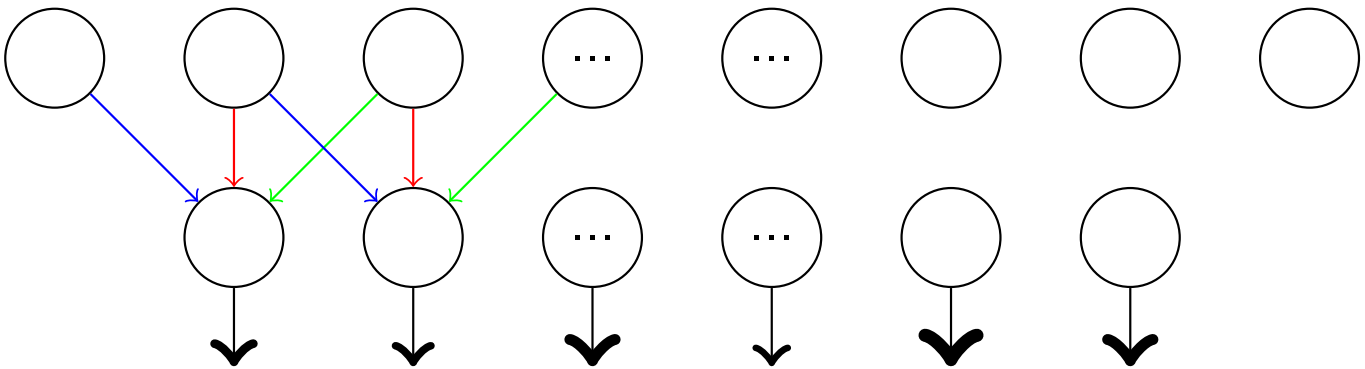
Response Normalization



Response Normalization

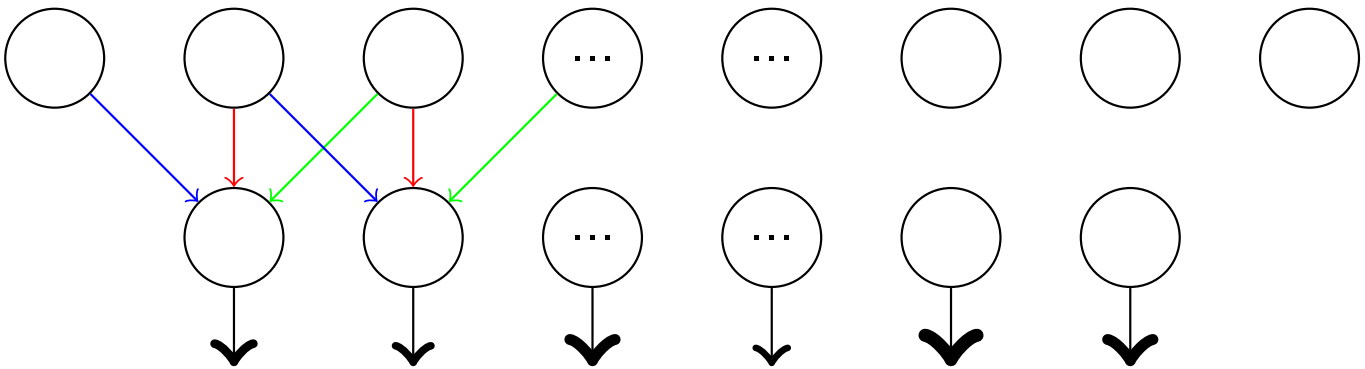


Response Normalization



$$a'_i := \frac{a_i}{\left(c + \sum_{j \in neighbours(a_i)} a_j^2\right)^b}$$

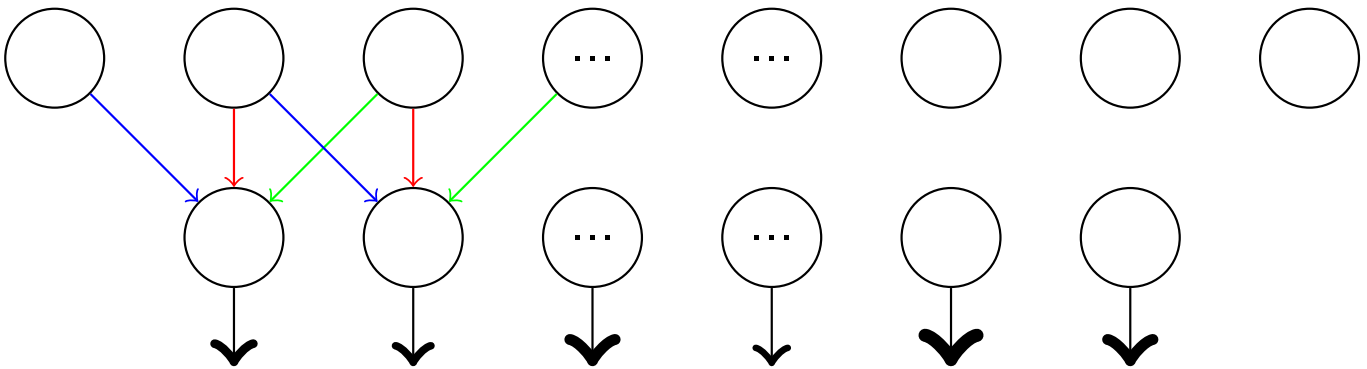
Response Normalization



$$a'_i := \frac{a_i}{\left(c + \sum_{j \in neighbours(a_i)} a_j^2\right)^b}$$

- aids generalization

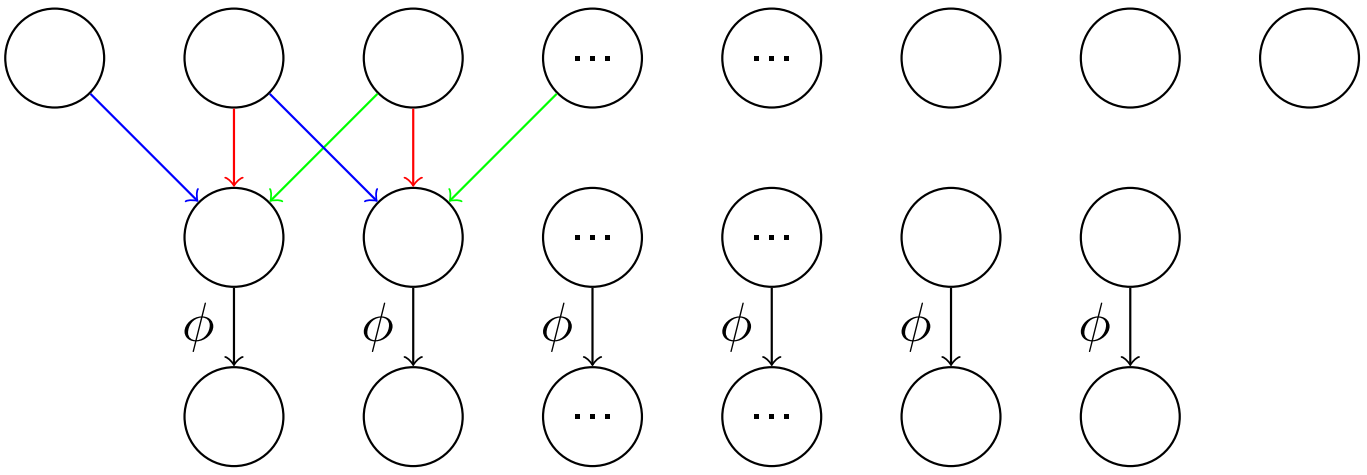
Response Normalization



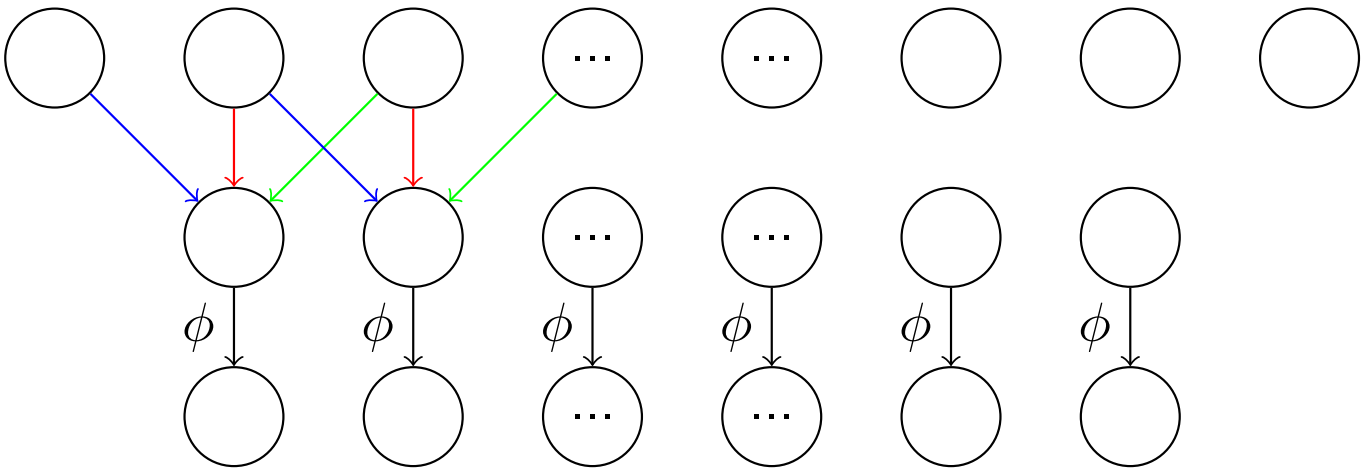
$$a'_i := \frac{a_i}{\left(c + \sum_{j \in neighbours(a_i)} a_j^2\right)^b}$$

- aids generalization
- though negligible: 1.3% less test error

Non-Linearity

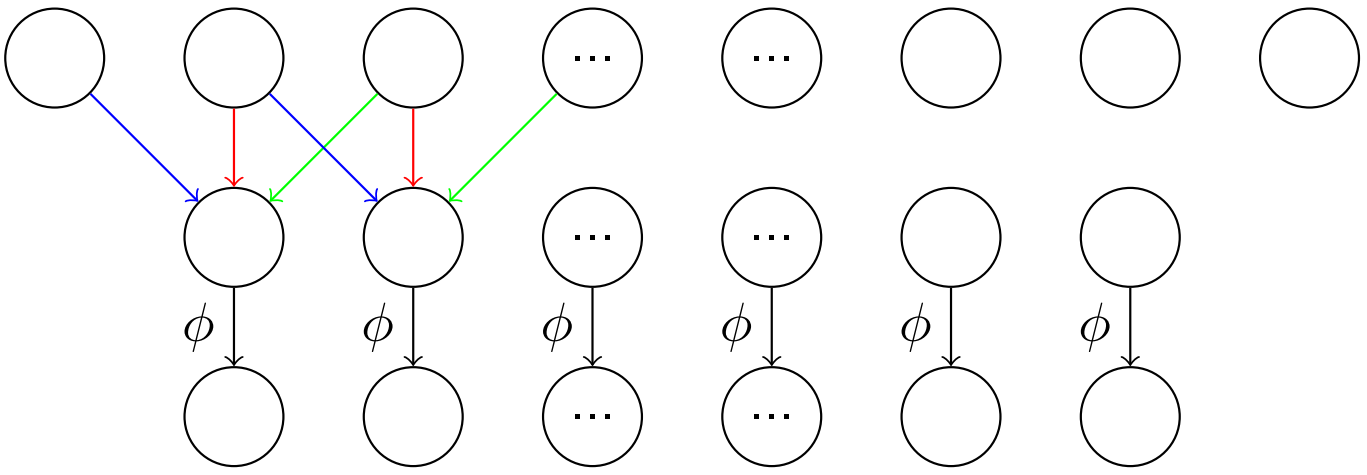


Non-Linearity



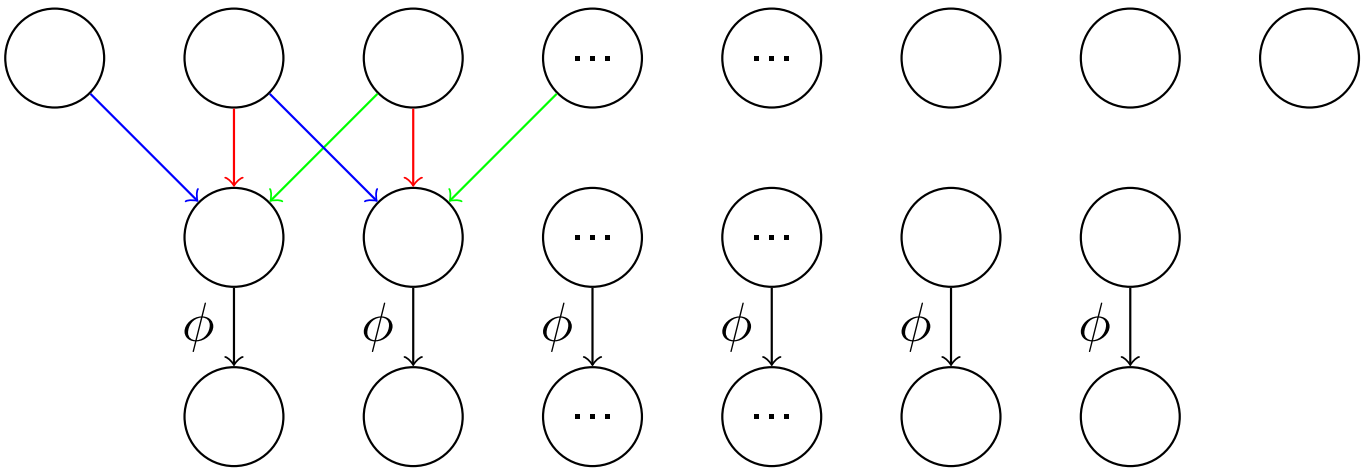
- $\phi(x) := \max(x, 0)$

Non-Linearity



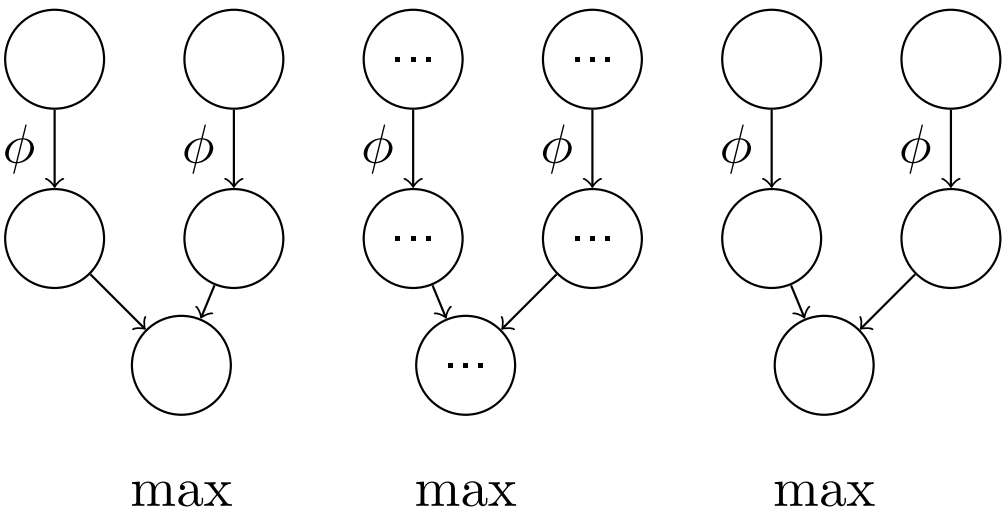
- $\phi(x) := \max(x, 0)$
- speeds up training time

Non-Linearity

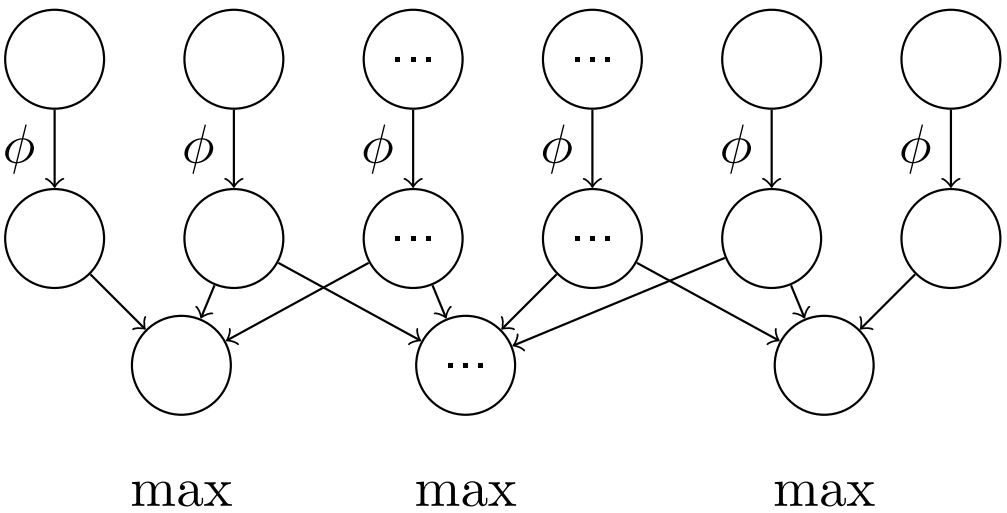


- $\phi(x) := \max(x, 0)$
- speeds up training time
- simpler computation of derivatives

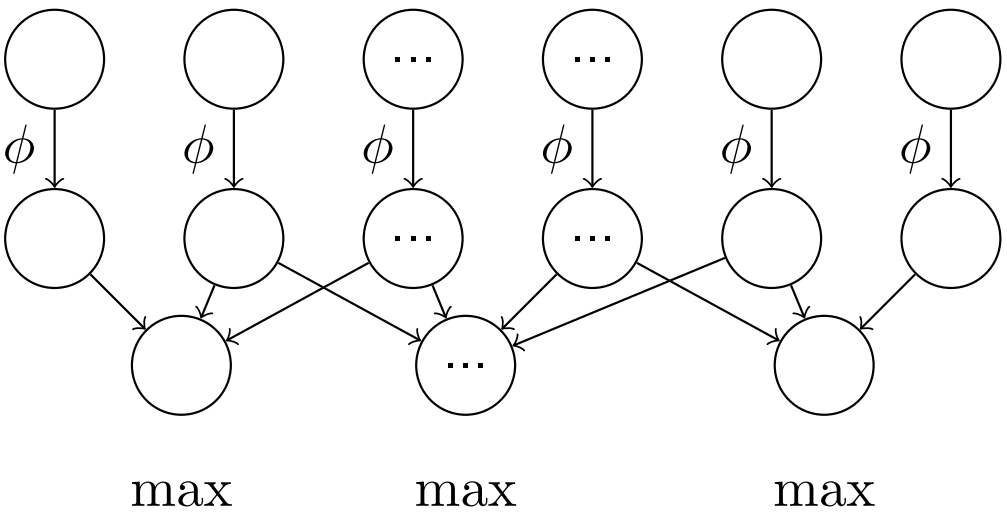
Overlapping Subsampling



Overlapping Subsampling



Overlapping Subsampling



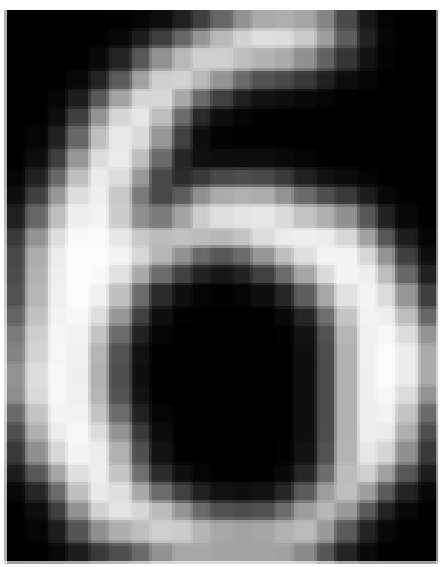
- again negligible: 0.3% less test error

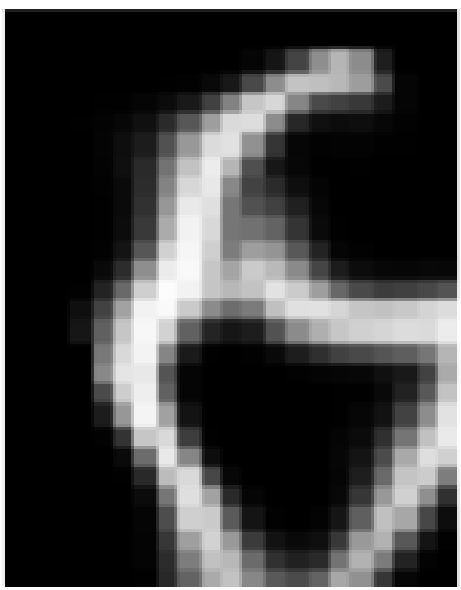
Dropout

- training: mute neuron with 50% chance
- testing: rescale outputs by 0.5
- reduces over-fitting
- intution: different architectures with shared weights

Getting more Data

- create more data
- use label preserving transformations





Critique

Critique

- negligible effects

Critique

- negligible effects
- increases training time

Critique

- negligible effects
- increases training time
- no reports about sensitivity of parameters