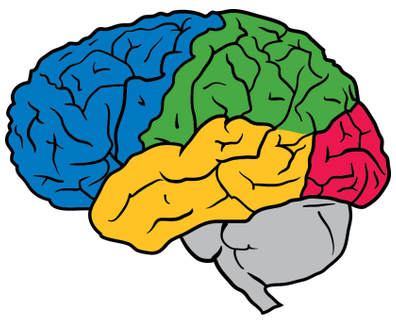
Targeted Dropout

Aidan Gomez

In collaboration with:
Ivan Zhang, Kevin Swersky, Yarin Gal, Geoff Hinton

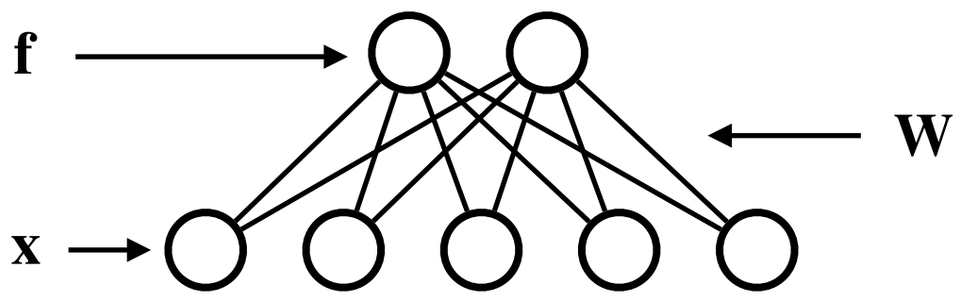


Neural Networks

Briefly:

A neural network is a composition of functions (layers) similar to **f**(**x**) below.

$$f_j(\mathbf{x}) = \phi(\sum_i W_{j,i} \cdot x_i)$$



Targeted Dropout

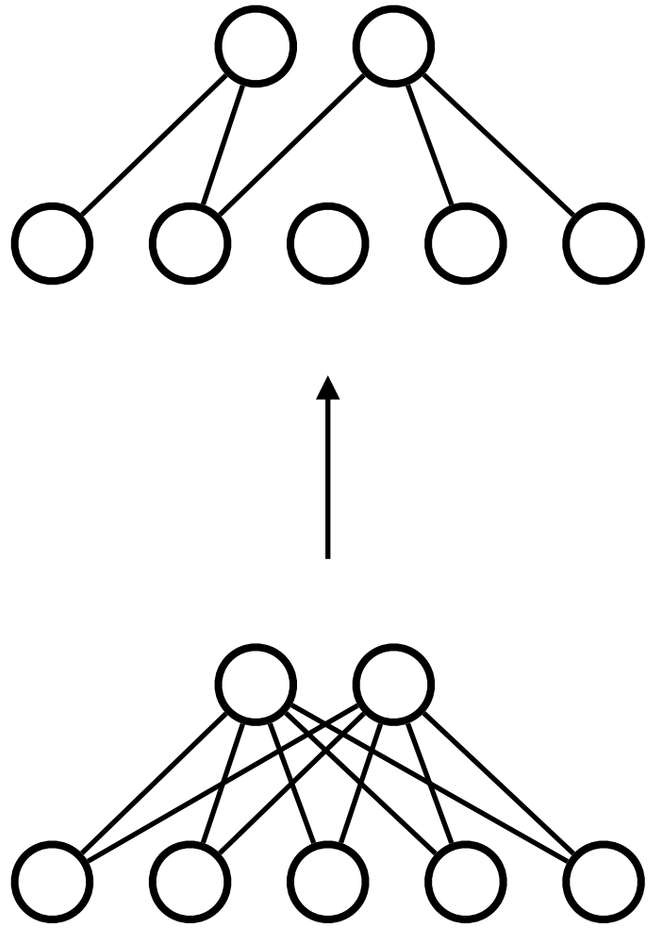
Sparsification

Immense Redundancy

- Neural networks are typically drastically over-parameterized.

- Yet, this seems necessary to facilitate training.

- Ideally, we'd like to be able to remove as much of this redundancy as possible. How?



Identifying Important Subnetworks

- We want to remove portions of our networks; But, which portion?

- We need some measure for determining importance

- Judging weight importance may appear to be a complicated matter:

  - **many** subsets of our network to choose from (combinatorial explosion)

  - measuring subset importance can be **extremely** expensive

Evaluating Importance

Idea:

Look at a Taylor expansion of the change in loss after deleting a subset "**d**" of the network

$$\left| \mathscr{L}(\Theta) - \mathscr{L}(\Theta - \mathbf{d}) \right| \approx \left| -\nabla_{\Theta} \mathscr{L}^{\top} \cdot \mathbf{d} + 0.5\, \mathbf{d}^{\top} \mathbf{H} \mathbf{d} + O(\|\mathbf{d}\|^3) \right|$$

Θ Network parameter vector

$\mathbf{d}$ Deletion vector ($\mathbf{d}_i = \Theta_i$ for deletion, 0s elsewhere)

$\mathbf{H}$ Hessian of loss

LeCun et al. (1990)

Evaluating Importance

Idea:

Look at a Taylor expansion of the change in loss after deleting a subset "**d**" of the network

$$\left| \mathscr{L}(\Theta) - \mathscr{L}(\Theta - \mathbf{d}) \right| \approx \left| -\nabla_{\Theta} \mathscr{L}^{\top} \cdot \mathbf{d} + 0.5\, \mathbf{d}^{\top} \mathbf{H} \mathbf{d} + O(\|\mathbf{d}\|^3) \right|$$

vanishes

Θ Network parameter vector

$\mathbf{d}$ Deletion vector ($\mathbf{d}_i = \Theta_i$ for deletion, 0s elsewhere)

$\mathbf{H}$ Hessian of loss

LeCun et al. (1990)

Evaluating Importance

Idea:

Look at a Taylor expansion of the change in loss after deleting a subset "**d**" of the network

$$\left| \mathscr{L}(\Theta) - \mathscr{L}(\Theta - \mathbf{d}) \right| \approx \left| -\nabla_{\Theta} \mathscr{L}^{\top} \cdot \mathbf{d} + 0.5\, \mathbf{d}^{\top} \mathbf{H} \mathbf{d} + O(\|\mathbf{d}\|^3) \right|$$

Θ Network parameter vector

$\mathbf{d}$ Deletion vector ($\mathbf{d}_i = \Theta_i$ for deletion, 0s elsewhere)

$\underbrace{\mathbf{H}}$ Hessian of loss

enormous

LeCun et al. (1990)

Evaluating Importance

$$\left| \mathscr{L}(\Theta) - \mathscr{L}(\Theta - \mathbf{d}) \right| \approx \left| -\nabla_{\Theta} \mathscr{L}^{\top} \cdot \mathbf{d} + 0.5\, \mathbf{d}^{\top} \mathbf{H} \mathbf{d} + O(\|\mathbf{d}\|^3) \right|$$

$$\mathbf{H} = \begin{bmatrix} \frac{\partial \mathscr{L}}{\partial \Theta_1 \Theta_1} & \frac{\partial \mathscr{L}}{\partial \Theta_1 \Theta_2} & \cdots \\ \frac{\partial \mathscr{L}}{\partial \Theta_2 \Theta_1} & \frac{\partial \mathscr{L}}{\partial \Theta_2 \Theta_2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \qquad \mathbf{d} = \begin{bmatrix} 0 & 0 & \Theta_3 & 0 & \Theta_5 & \dots \end{bmatrix}$$

Evaluating Importance

$$\left| \mathscr{L}(\Theta) - \mathscr{L}(\Theta - \mathbf{d}) \right| \approx \left| -\nabla_{\Theta} \mathscr{L}^{\top} \cdot \mathbf{d} + 0.5\, \mathbf{d}^{\top} \mathbf{H} \mathbf{d} + O(\|\mathbf{d}\|^3) \right|$$

$$\begin{array}{c} \\ \Theta_1 \\ \Theta_2 \end{array} \begin{array}{c} \begin{array}{cccc} \Theta_1 & \Theta_2 & \end{array} \\ \begin{bmatrix} \dfrac{\partial \mathscr{L}}{\partial \Theta_1 \Theta_1} & \dfrac{\partial \mathscr{L}}{\partial \Theta_1 \Theta_2} & \cdots \\ \dfrac{\partial \mathscr{L}}{\partial \Theta_2 \Theta_1} & \dfrac{\partial \mathscr{L}}{\partial \Theta_2 \Theta_2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \end{array}$$

Evaluating Importance

$$\left| \mathscr{L}(\Theta) - \mathscr{L}(\Theta - \mathbf{d}) \right| \approx \left| -\nabla_{\Theta} \mathscr{L}^{\top} \cdot \mathbf{d} + 0.5\, \mathbf{d}^{\top} \mathbf{H} \mathbf{d} + O(\|\mathbf{d}\|^3) \right|$$

$$\begin{array}{c} \\ \Theta_2 \\ \Theta_1 \end{array} \begin{array}{c} \begin{array}{cccc} \Theta_2 & & \Theta_1 & \end{array} \\ \begin{bmatrix} \dfrac{\partial \mathscr{L}}{\partial \Theta_2 \Theta_2} & \dfrac{\partial \mathscr{L}}{\partial \Theta_2 \Theta_1} & \cdots \\ \dfrac{\partial \mathscr{L}}{\partial \Theta_1 \Theta_2} & \dfrac{\partial \mathscr{L}}{\partial \Theta_1 \Theta_1} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \end{array}$$

Evaluating Importance

$$\left| \mathscr{L}(\Theta) - \mathscr{L}(\Theta - \mathbf{d}) \right| \approx \left| -\nabla_{\Theta} \mathscr{L}^{\top} \cdot \mathbf{d} + 0.5\, \mathbf{d}^{\top} \mathbf{H} \mathbf{d} + O(\|\mathbf{d}\|^3) \right|$$

	Θ_1	Θ_2	Θ_3	Θ_4
Θ_1	$\frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_1}$	$\frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_2}$	$\frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_3}$	$\frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_4}$
Θ_2	$\frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_1}$	$\frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_2}$	$\frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_3}$	$\frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_4}$
Θ_3	$\frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_1}$	$\frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_2}$	$\frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_3}$	$\frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_4}$
Θ_4	$\frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_1}$	$\frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_2}$	$\frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_3}$	$\frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_4}$

Evaluating Importance

$$\left| \mathscr{L}(\Theta) - \mathscr{L}(\Theta - \mathbf{d}) \right| \approx \left| -\nabla_{\Theta}\mathscr{L}^{\top} \cdot \mathbf{d} + 0.5\, \mathbf{d}^{\top}\mathbf{H}\mathbf{d} + O(\|\mathbf{d}\|^{3}) \right|$$

$$\mathbf{d}^{\top}\mathbf{H}\mathbf{d}$$

$$\begin{bmatrix} 0 & 0 & \Theta_3 & \Theta_4 \end{bmatrix} \bullet \begin{bmatrix} \frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_1} & \frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_2} & \frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_3} & \frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_4} \\ \frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_1} & \frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_2} & \frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_3} & \frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_4} \\ \frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_1} & \frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_2} & \frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_3} & \frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_4} \\ \frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_1} & \frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_2} & \frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_3} & \frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_4} \end{bmatrix} \bullet \begin{bmatrix} 0 \\ 0 \\ \Theta_3 \\ \Theta_4 \end{bmatrix}$$

Evaluating Importance

$$\left| \mathscr{L}(\Theta) - \mathscr{L}(\Theta - \mathbf{d}) \right| \approx \left| -\nabla_{\Theta}\mathscr{L}^{\top} \cdot \mathbf{d} + 0.5\,\mathbf{d}^{\top}\mathbf{H}\mathbf{d} + O(\|\mathbf{d}\|^{3}) \right|$$

$$\mathbf{b}^{\top}(\mathbf{d}^{\top} \odot \mathbf{H} \odot \mathbf{d})\mathbf{b}$$

0	0	1	1

$\bullet$

$0\frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_1}0$	$0\frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_2}0$	$\Theta_3\frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_3}0$	$\Theta_4\frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_4}0$
$0\frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_1}0$	$0\frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_2}0$	$\Theta_3\frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_3}0$	$\Theta_4\frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_4}0$
$0\frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_1}\Theta_3$	$0\frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_2}\Theta_3$	$\Theta_3\frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_3}\Theta_3$	$\Theta_4\frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_4}\Theta_3$
$0\frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_1}\Theta_4$	$0\frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_2}\Theta_4$	$\Theta_3\frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_3}\Theta_4$	$\Theta_4\frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_4}\Theta_4$

$\bullet$

0
0
1
1

Evaluating Importance

$$\left|\mathscr{L}(\Theta) - \mathscr{L}(\Theta - \mathbf{d})\right| \approx \left|-\nabla_{\Theta}\mathscr{L}^{\top} \cdot \mathbf{d} + 0.5\,\mathbf{d}^{\top}\mathbf{H}\mathbf{d} + O(\|\mathbf{d}\|^{3})\right|$$

$$\mathbf{b}^{\top}(\mathbf{d}^{\top} \odot \mathbf{H} \odot \mathbf{d})\mathbf{b}$$

0	0	1	1

$\bullet$

$0\frac{\partial\mathscr{L}}{\partial\Theta_1\partial\Theta_1}0$	$0\frac{\partial\mathscr{L}}{\partial\Theta_1\partial\Theta_2}0$	$\Theta_3\frac{\partial\mathscr{L}}{\partial\Theta_1\partial\Theta_3}0$	$\Theta_4\frac{\partial\mathscr{L}}{\partial\Theta_1\partial\Theta_4}0$
$0\frac{\partial\mathscr{L}}{\partial\Theta_2\partial\Theta_1}0$	$0\frac{\partial\mathscr{L}}{\partial\Theta_2\partial\Theta_2}0$	$\Theta_3\frac{\partial\mathscr{L}}{\partial\Theta_2\partial\Theta_3}0$	$\Theta_4\frac{\partial\mathscr{L}}{\partial\Theta_2\partial\Theta_4}0$
$0\frac{\partial\mathscr{L}}{\partial\Theta_3\partial\Theta_1}\Theta_3$	$0\frac{\partial\mathscr{L}}{\partial\Theta_3\partial\Theta_2}\Theta_3$	$\Theta_3\frac{\partial\mathscr{L}}{\partial\Theta_3\partial\Theta_3}\Theta_3$	$\Theta_4\frac{\partial\mathscr{L}}{\partial\Theta_3\partial\Theta_4}\Theta_3$
$0\frac{\partial\mathscr{L}}{\partial\Theta_4\partial\Theta_1}\Theta_4$	$0\frac{\partial\mathscr{L}}{\partial\Theta_4\partial\Theta_2}\Theta_4$	$\Theta_3\frac{\partial\mathscr{L}}{\partial\Theta_4\partial\Theta_3}\Theta_4$	$\Theta_4\frac{\partial\mathscr{L}}{\partial\Theta_4\partial\Theta_4}\Theta_4$

$\bullet$

0
0
1
1

Evaluating Importance

$$\left| \mathscr{L}(\Theta) - \mathscr{L}(\Theta - \mathbf{d}) \right| \approx \left| -\nabla_{\Theta}\mathscr{L}^{\top} \cdot \mathbf{d} + 0.5\, \mathbf{d}^{\top}\mathbf{H}\mathbf{d} + O(\|\mathbf{d}\|^{3}) \right|$$

$$\mathbf{b}^{\top}(\Theta^{\top} \odot \mathbf{H} \odot \Theta)\mathbf{b}$$

0	0	1	1

$\bullet$

$\Theta_1 \frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_1} \Theta_1$	$\Theta_2 \frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_2} \Theta_1$	$\Theta_3 \frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_3} \Theta_1$	$\Theta_4 \frac{\partial \mathscr{L}}{\partial \Theta_1 \partial \Theta_4} \Theta_1$
$\Theta_1 \frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_1} \Theta_2$	$\Theta_2 \frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_2} \Theta_2$	$\Theta_3 \frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_3} \Theta_2$	$\Theta_4 \frac{\partial \mathscr{L}}{\partial \Theta_2 \partial \Theta_4} \Theta_2$
$\Theta_1 \frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_1} \Theta_3$	$\Theta_2 \frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_2} \Theta_3$	$\Theta_3 \frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_3} \Theta_3$	$\Theta_4 \frac{\partial \mathscr{L}}{\partial \Theta_3 \partial \Theta_4} \Theta_3$
$\Theta_1 \frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_1} \Theta_4$	$\Theta_2 \frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_2} \Theta_4$	$\Theta_3 \frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_3} \Theta_4$	$\Theta_4 \frac{\partial \mathscr{L}}{\partial \Theta_4 \partial \Theta_4} \Theta_4$

$\bullet$

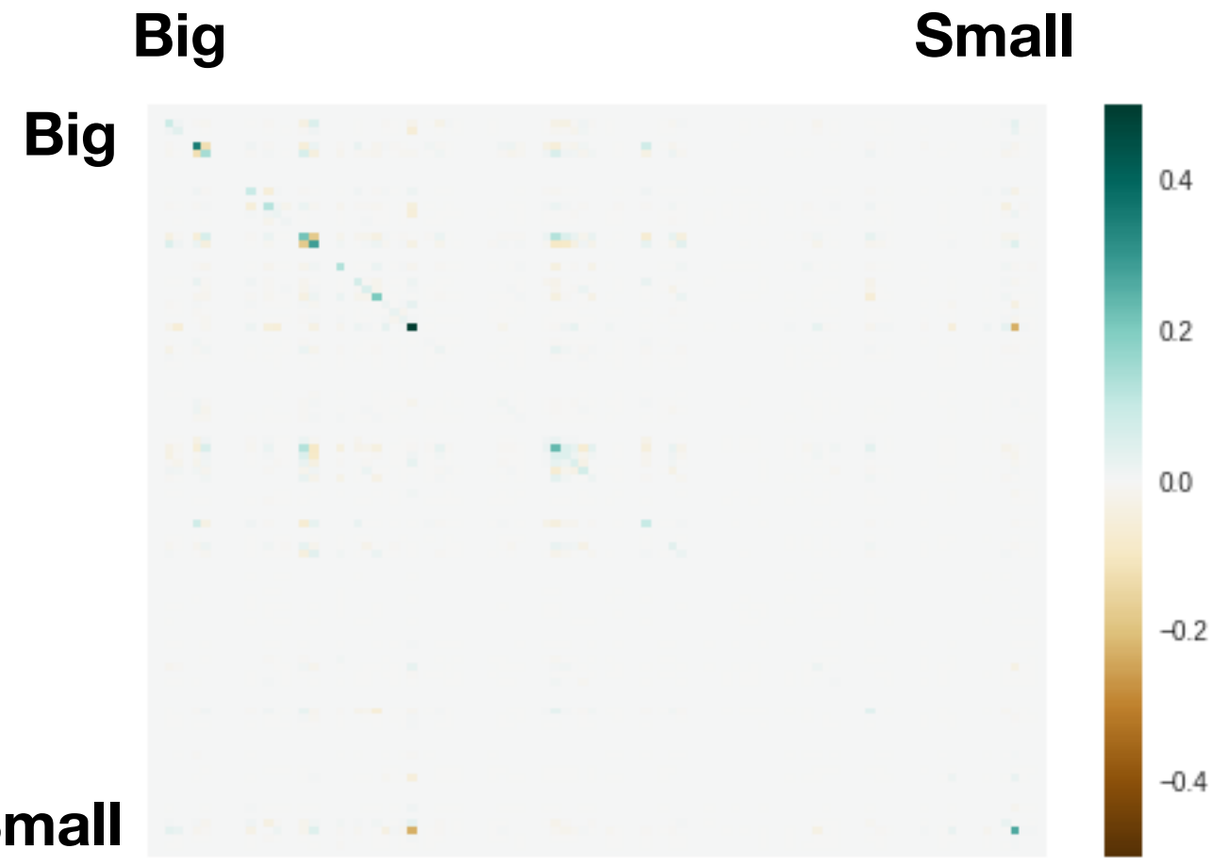
0
0
1
1

Visualizing Importance

- We can test our assumptions by inspecting the coadaptation matrix

- Position (*i*, *j*) of the matrix represents the dependence of parameter *i* on parameter *j*.

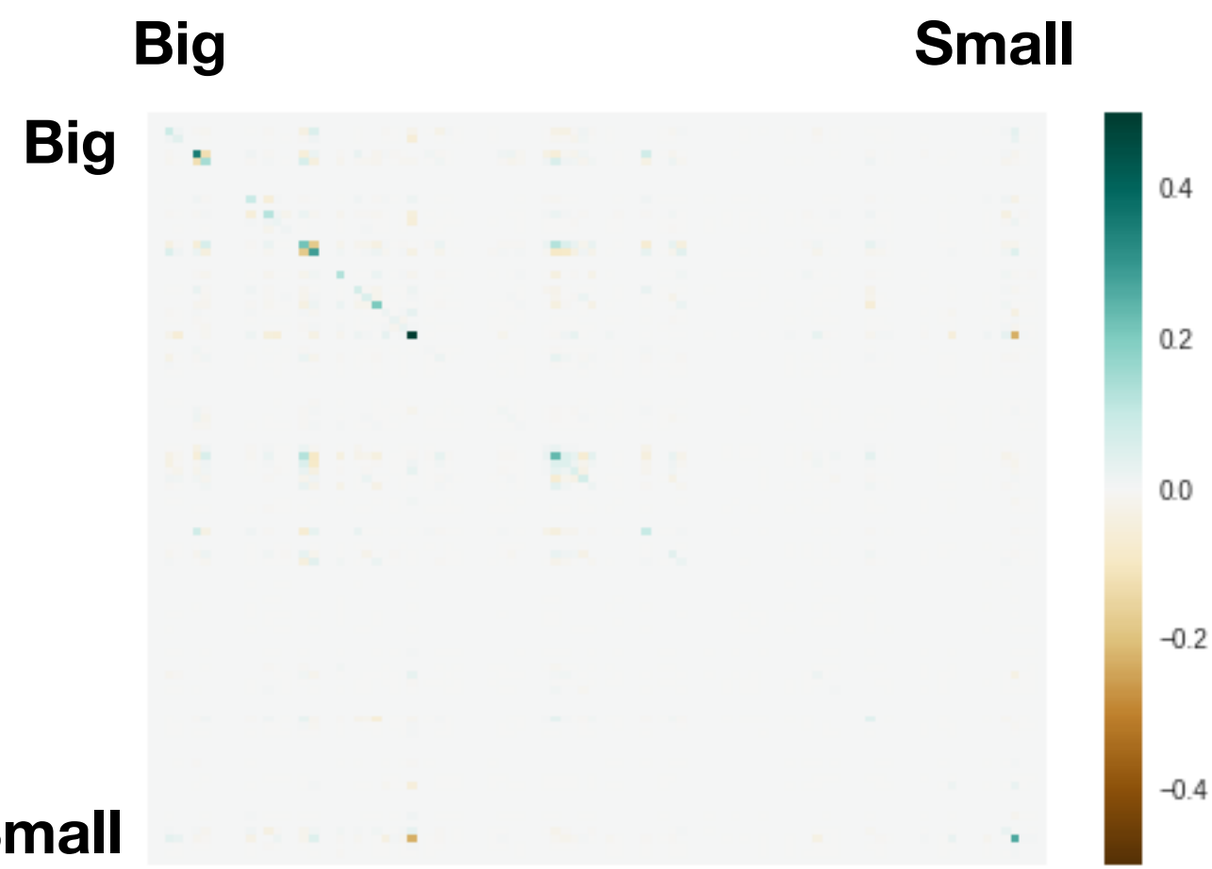
- Pruning a set of weights {*l*, *m*, *n*, …} will change the loss proportional to the sum of the *l*, *m*, and *n*th columns



$$\left[\Theta^{\top} \odot \mathbf{H} \odot \Theta\right]_{i,j} = \Theta_i \odot \mathbf{H}_{i,j} \odot \Theta_j$$

Visualizing Importance

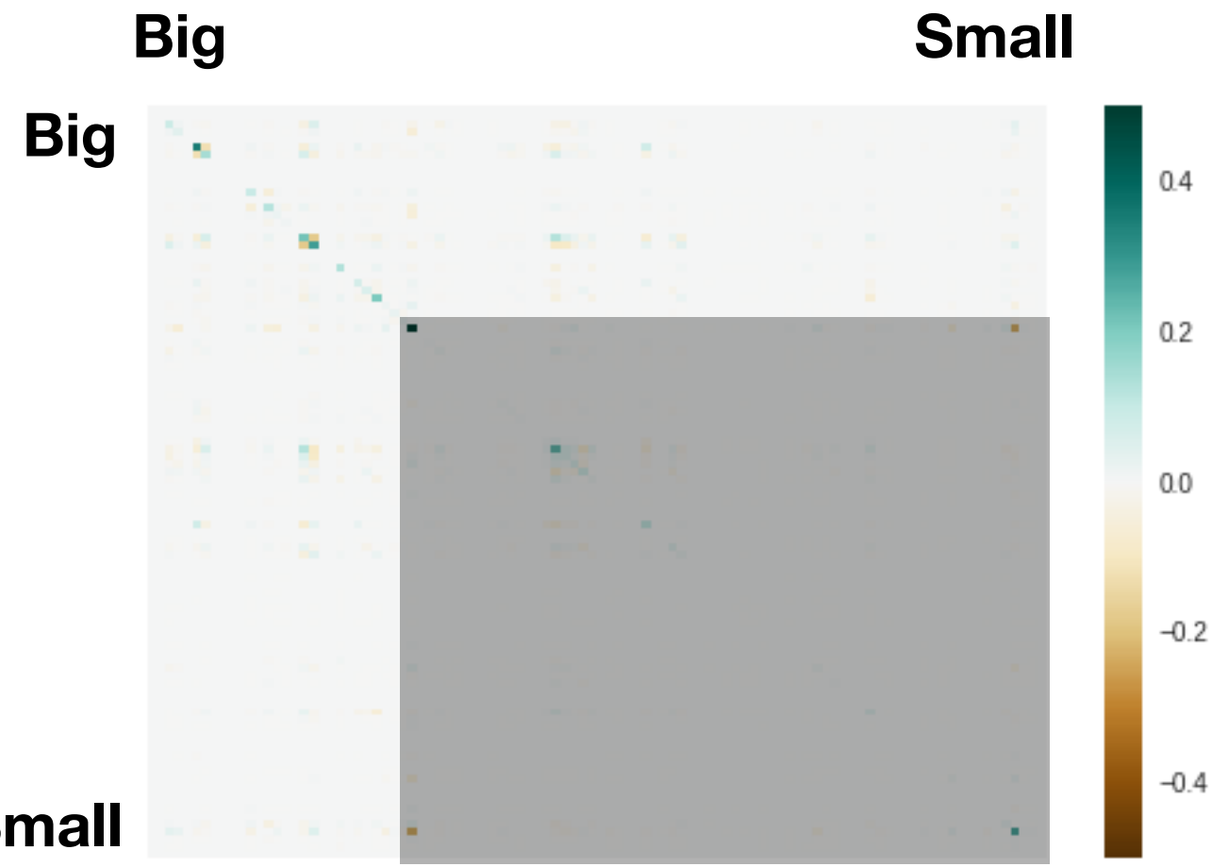
- First 25% of weights are those with the largest magnitude.
- Remaining 75% are those we intend to prune.
- We hope that very little importance is placed on the 75% deemed unimportant.
- Ideally, all importance would be concentrated in the top-left 16th



$$\left[\Theta^{\top} \odot \mathbf{H} \odot \Theta\right]_{i,j} = \Theta_i \odot \mathbf{H}_{i,j} \odot \Theta_j$$

Unregularized Model

- First 25% of weights are those with the largest magnitude.
- Remaining 75% are those we intend to prune.
- We hope that very little importance is placed on the 75% deemed unimportant.
- Ideally, all importance would be concentrated in the top-left 16th



$$\left[\Theta^{\top} \odot \mathbf{H} \odot \Theta\right]_{i,j} = \Theta_i \odot \mathbf{H}_{i,j} \odot \Theta_j$$

Simplifying Assumptions

- It’s clear that some simplifications are necessary:
  - Assume group/weight-wise independence
  - Assume a first order approx. is “good enough”
  - Assume a zeroth order approx. is “good enough”

Simplifying Assumptions

- It’s clear that some simplifications are necessary:
  - Assume group/weight-wise independence
  - Assume a first order approx. is “good enough”
  - Assume a zeroth order approx. is “good enough”

Saves computing the full Hessian, much less the **many** Hessian-vector products originally required.

Now we need only compute a block-diagonal or diagonal approximation.

This technique is known as Optimal Brain Damage (LeCun et al., 1990)

Simplifying Assumptions

- It's clear that some simplifications are necessary:
  - Assume group/weight-wise independence
  - Assume a first order approx. is "good enough"
  - Assume a zeroth order approx. is "good enough"

Despite the claim on a previous slide that $\nabla_\Theta \mathscr{L} \to 0$, Pavlo Molchanov and colleagues suggest that:

- yes, $\nabla_\Theta \mathscr{L} \cdot \mathbf{d}^\top$ is uninformative in expectation, but…
- the variance of this quantity correlates with the local stability of the loss, and
- we have access to this variance since $\mathbb{E}[\,|\nabla_\Theta \mathscr{L} \cdot \mathbf{d}^\top|\,] \propto \sigma$

Molchanov et al. (2017)

Simplifying Assumptions

- It’s clear that some simplifications are necessary:
  - Assume group/weight-wise independence
  - Assume a first order approx. is “good enough”
  - Assume a zeroth order approx. is “good enough”

Assume that one can judge the importance of a weight by its magnitude.

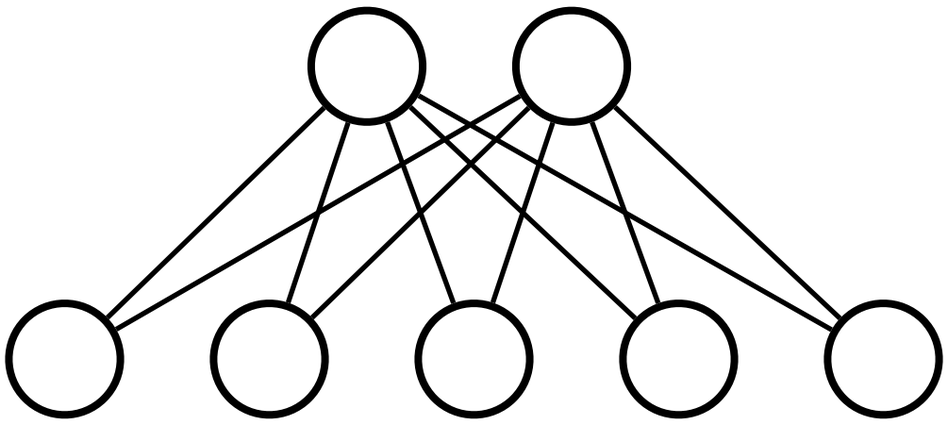
In practice this works quite well.

But far from perfect…

probably: Rumelhart (1988)

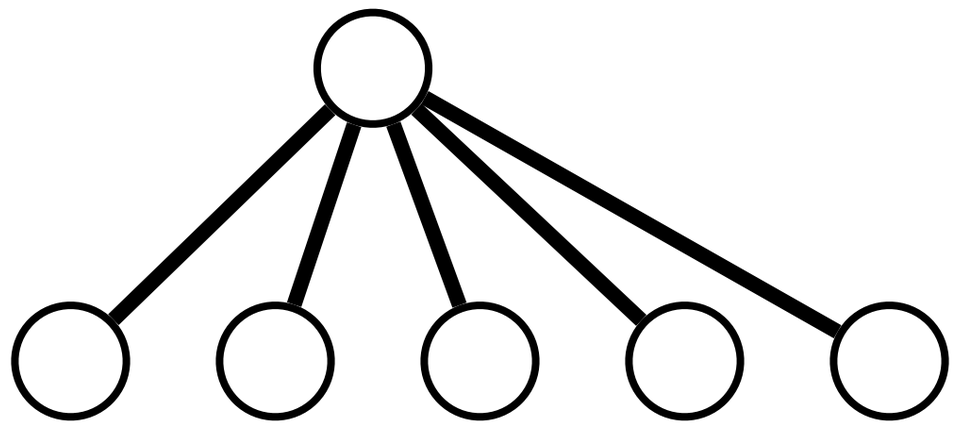
Targeted Dropout

Idea: Inspired by Hinton's description of dropout reducing coadaptation between units, use dropout to **reduce dependance of the identified important subnetwork on its complement** in the network.



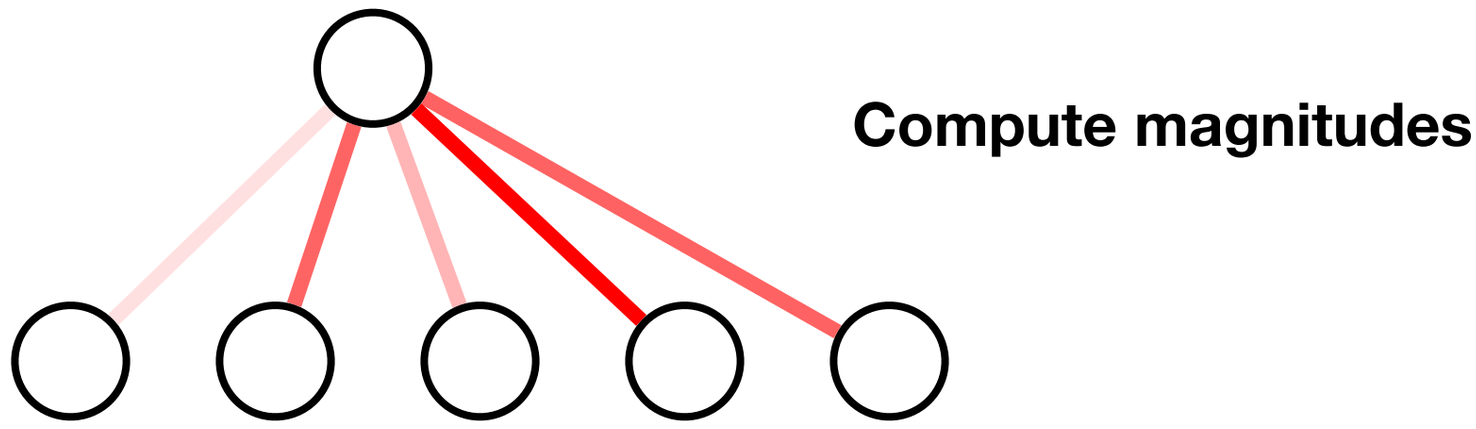
Targeted Dropout

Idea: Inspired by Hinton's description of dropout reducing coadaptation between units, use dropout to **reduce dependance of the identified important subnetwork on its complement** in the network.



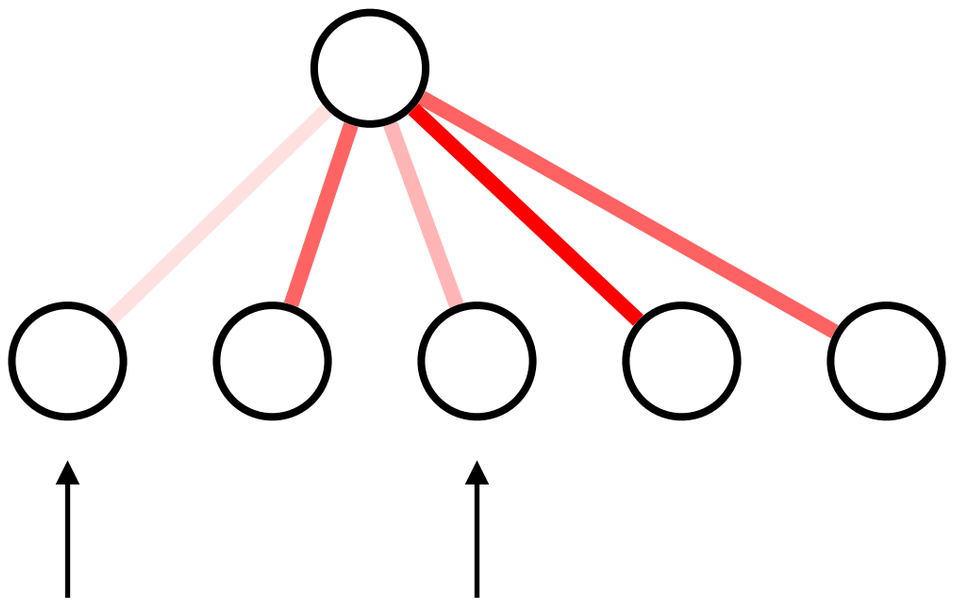
Targeted Dropout

Idea: Inspired by Hinton's description of dropout reducing coadaptation between units, use dropout to **reduce dependance of the identified important subnetwork on its complement** in the network.



Targeted Dropout

Idea: Inspired by Hinton's description of dropout reducing coadaptation between units, use dropout to **reduce dependance of the identified important subnetwork on its complement** in the network.

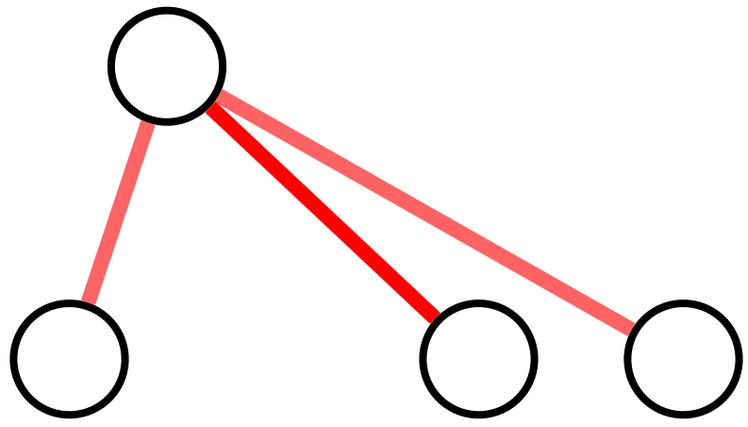


Apply dropout to <u>only</u> the smallest *k*%

Targeted Dropout

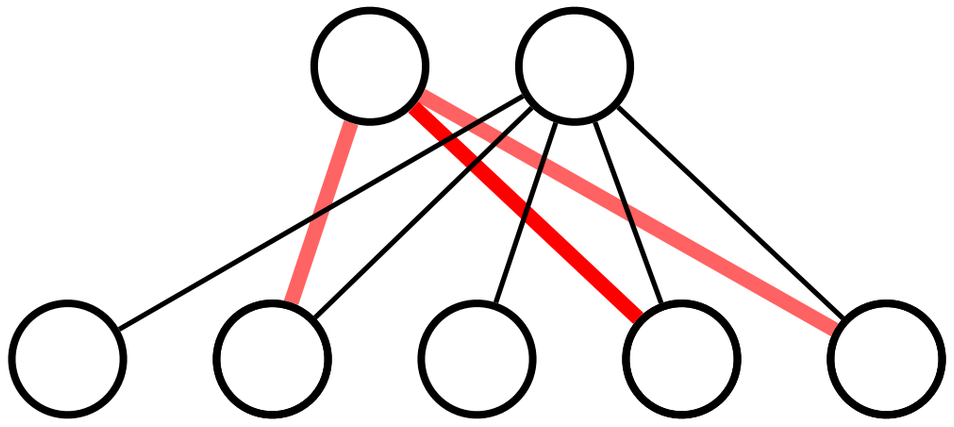
Idea: Inspired by Hinton's description of dropout reducing coadaptation between units, use dropout to **reduce dependance of the identified important subnetwork on its complement** in the network.

After training delete those weights



Targeted Dropout

Idea: Inspired by Hinton's description of dropout reducing coadaptation between units, use dropout to **reduce dependance of the identified important subnetwork on its complement** in the network.



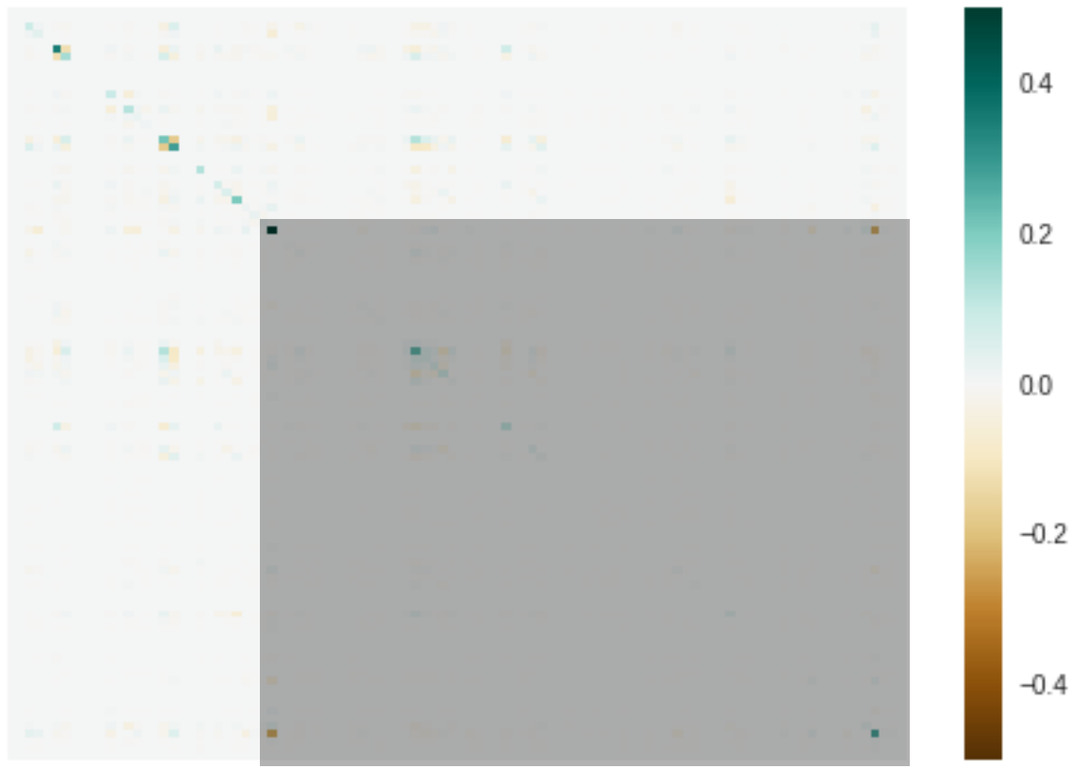
Targeted Dropout

```
def targeting_fn(inputs, k):
  shape = tf.shape(inputs)
  size = tf.to_int32(tf.reduce_prod(shape[:-1]))

  inputs = tf.reshape(inputs, [size, shape[-1]])
  transpose = tf.transpose(inputs)

  thres = tf.contrib.framework.sort(tf.abs(transpose), axis=1)[:, k]
  mask = tf.to_float(tf.abs(inputs) <= thres[None, :])

  return tf.reshape(mask, shape)
```

```
def targeted_dropout(inputs, targ_rate, keep_prob, is_training):
  dim = tf.to_float(tf.shape(inputs)[-1])
  k = tf.round(dim * targ_rate)

  if not is_training and do_prune:
    drop_rate = 1. - keep_prob
    k = tf.round(dim * targ_rate * drop_rate)

  mask = targeting_fn(inputs, k)
  mask = tf.cast(mask, inputs.dtype)

  if is_training:
    return inputs * (1 - mask) + tf.nn.dropout(inputs, keep_prob) * mask
  else:
    return inputs * (1 - mask)
```

Unregularized Model

- First 25% of weights are those with the largest magnitude.
- Remaining 75% are those we intend to prune.
- We hope that very little importance is placed on the 75% deemed unimportant.
- Ideally, all importance would be concentrated in the top-left 16th



$$\left[\Theta^{\top} \odot \mathbf{H} \odot \Theta\right]_{i,j} = \Theta_i \odot \mathbf{H}_{i,j} \odot \Theta_j$$

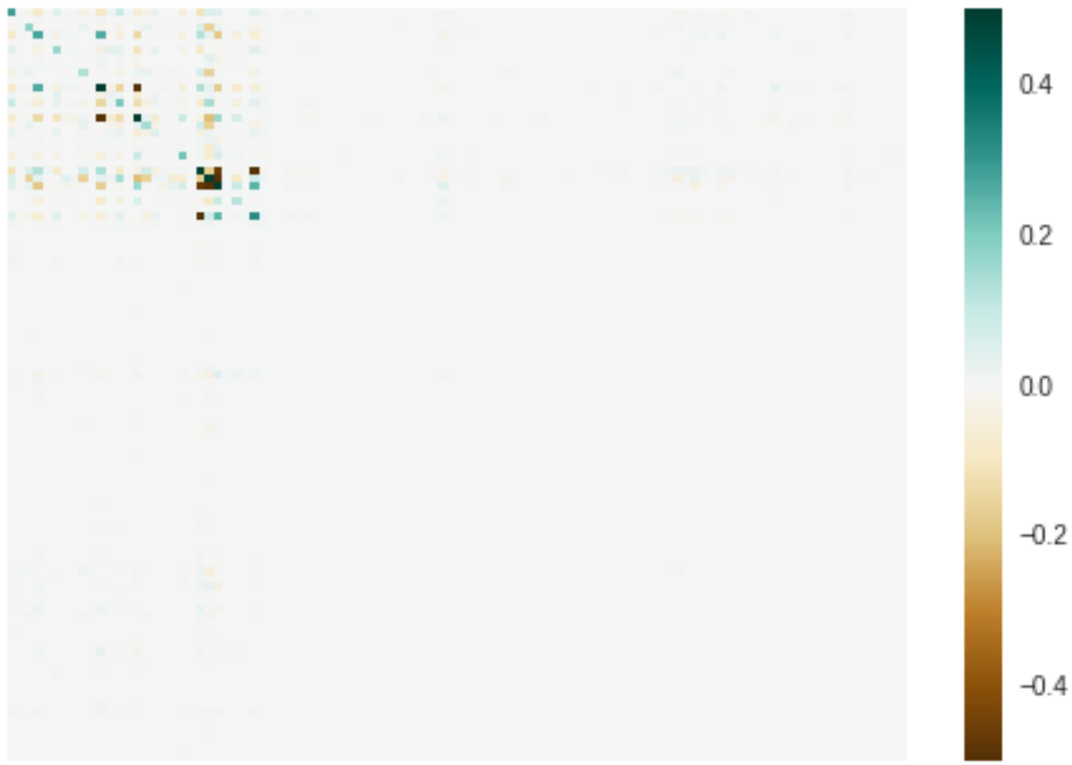
Targeted Dropout

- Targeted Dropout achieves precisely what we had hoped for:
  - decoupling the “important” sub-network from the “unimportant” one.

$$\left[\Theta^{\top} \odot \mathbf{H} \odot \Theta\right]_{i,j} = \Theta_i \odot \mathbf{H}_{i,j} \odot \Theta_j$$

Targeted Dropout

- Targeted Dropout achieves precisely what we had hoped for:
  - decoupling the “important” sub-network from the “unimportant” one.



$$\left[\Theta^{\top} \odot \mathbf{H} \odot \Theta\right]_{i,j} = \Theta_i \odot \mathbf{H}_{i,j} \odot \Theta_j$$

Weight-level Results

prune percentage	none	dropout α=0.25	targeted α=0.66,γ=0.75	targeted α=0.5,γ=0.5
0 %	94.30	94.11	93.87	94.38
10%	94.16	94.05	93.85	94.31
20%	94.19	93.97	93.84	94.27
30%	94.21	93.90	93.89	94.27
40%	93.93	93.59	93.81	94.27
50%	93.31	92.00	93.84	94.33
60%	91.50	88.23	93.89	93.92
70%	82.89	58.07	93.84	92.19
80%	38.35	10.66	92.31	74.31
90%	12.76	9.92	46.57	16.67

- We baseline against three recent techniques: L1 regularisation, variational dropout, and smallify (to come). Results on CIFAR-10 using a ResNet-32.

Unit-level Results

prune percentage	none	dropout α=0.25	targeted α=0.66,γ=0.75	targeted α=0.5,γ=0.5
0 %	94.29	92.93	90.55	93.27
10%	90.38	84.98	90.83	92.87
20%	74.38	21.27	89.88	91.95
30%	36.49	10.32	87.35	89.84
40%	12.64	12.52	85.39	86.41
50%	10.19	10.13	80.84	67.01
60%	11.32	9.95	71.97	11.70
70%	10.14	9.98	55.98	9.98
80%	10.01	9.91	10.02	9.94
90%	10.12	9.95	10.07	9.95

- We baseline against three recent techniques: L1 regularisation, variational dropout, and smallify (to come). Results on CIFAR-10 using a ResNet-32.

Ramping Results

prune percentage

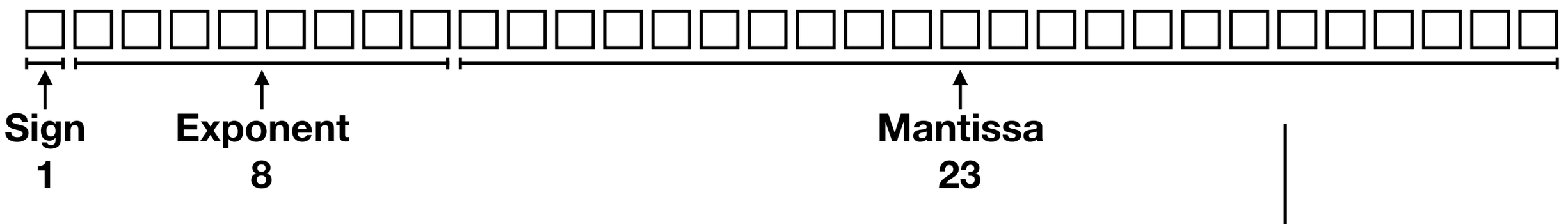
	targeted α=0.66,γ=0.75		ramp targ $alpha$=0.99,γ=0.99		ramp targ $alpha$=0.99,γ=0.99
0 %	93.87	90%	88.70	98.5%	88.74
10%	93.85	91%	88.74	98.6%	88.82
20%	93.84	92%	88.75	98.7%	88.79
30%	93.89	93%	88.75	98.8%	88.74
40%	93.81	94%	88.80	98.9%	88.70
50%	93.84	95%	88.73	99.0%	88.71
60%	93.89	96%	88.74	99.1%	87.63
70%	93.84	97%	88.67	99.2%	67.25
80%	92.31	98%	88.70	99.3%	51.86
90%	46.57	99%	88.75	99.4%	11.59

- **Idea**: slowly ramp up the targeting proportion from 0% to some large percent (99% in our experiments)

Bitrot + Exprot Quantization

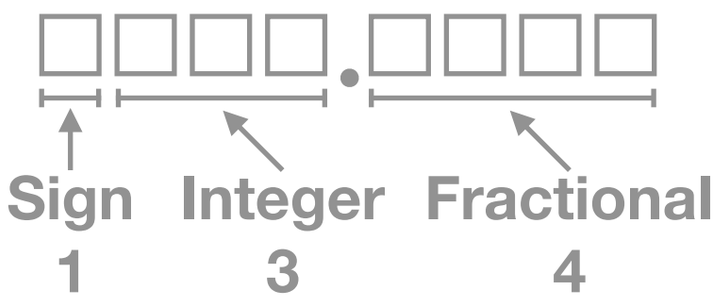
Numeric Representation

Single-precision floating-point (float32)



Truncated half-precision floating-point (bfloat16)

8-bit fixed-point (Q3.4)

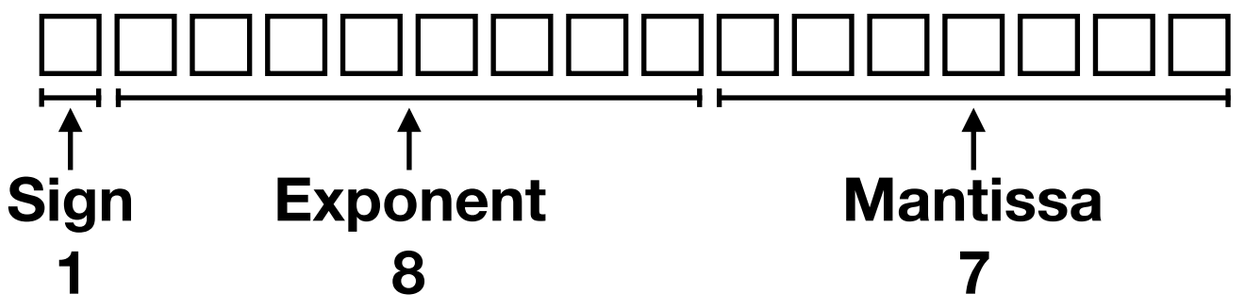


$$-1^{\textbf{sign}} \cdot 2^{\textbf{exp}-127} \cdot 1.\textbf{mant}$$

Numeric Representation

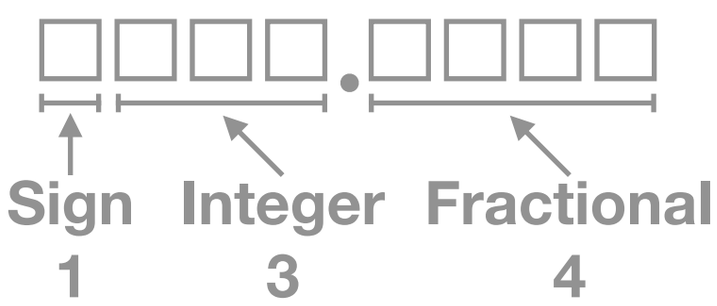
Single-precision floating-point (float32)

Truncated half-precision floating-point (bfloat16)



$$-1^{\textbf{sign}} \cdot 2^{\textbf{exp}-127} \cdot 1.\textbf{mant}$$

8-bit fixed-point (Q3.4)

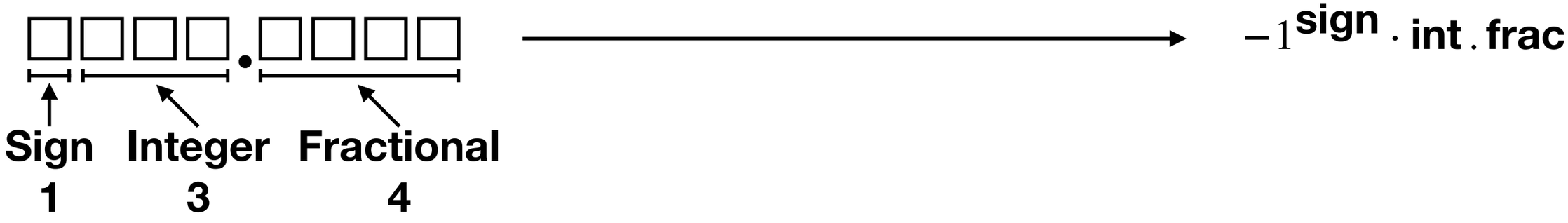


Numeric Representation

Single-precision floating-point (float32)

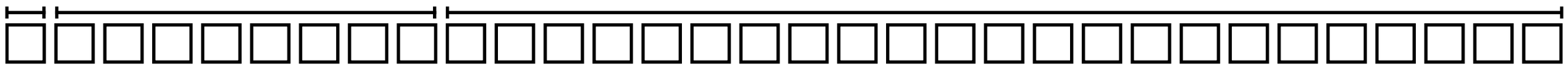
Truncated half-precision floating-point (bfloat16)

8-bit fixed-point (Q3.4)



Bitrot

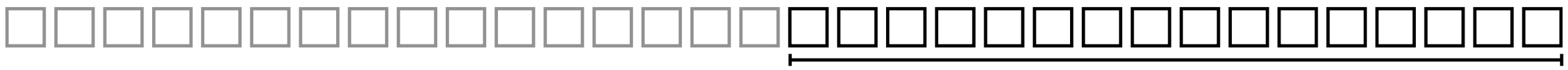
Single-precision floating-point (float32)



Idea: Apply dropout to the bits of the mantissa in order to gradually anneal the network from a float32 to a bfloat16.

Bitrot

Single-precision floating-point (float32)

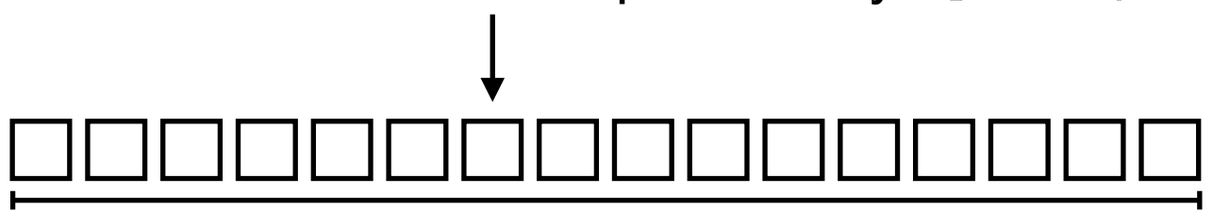


Idea: Apply dropout to the bits of the mantissa in order to gradually anneal the network from a float32 to a bfloat16.

Rippel, Gelbart, and Adams. (2014)

Nested Dropout

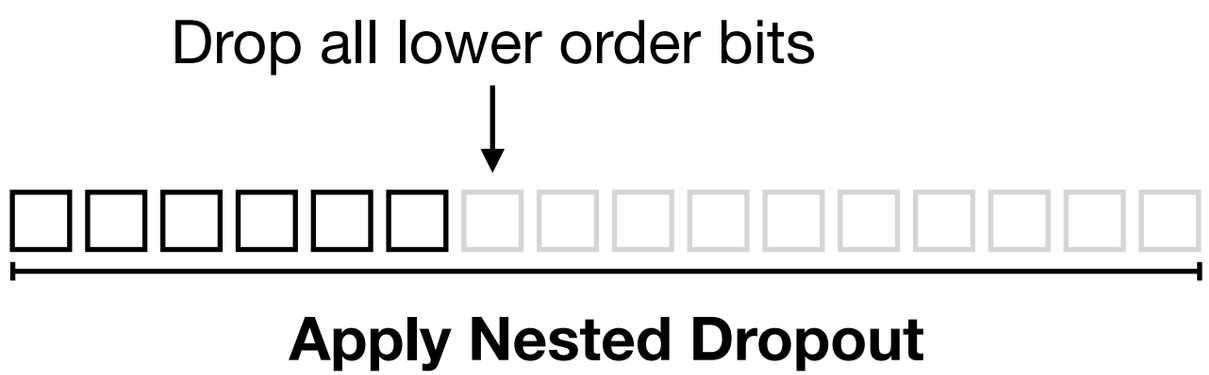
Choose index *i* with probability: $p(i) = \rho^i(1-\rho)$



Apply Nested Dropout

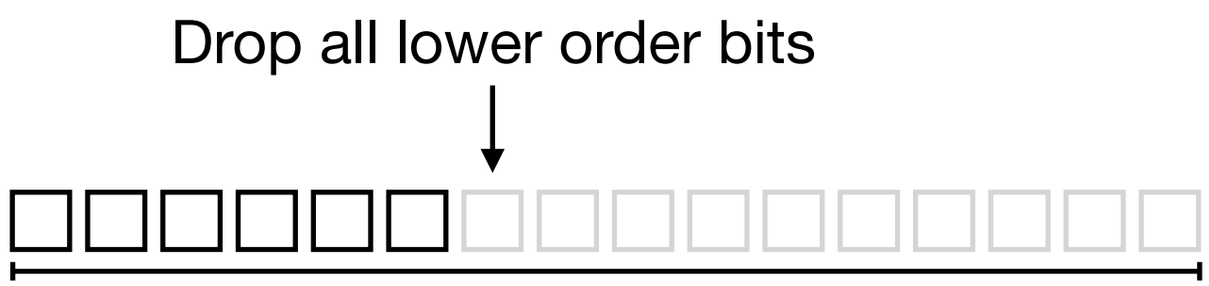
Rippel, Gelbart, and Adams. (2014)

Nested Dropout



Rippel, Gelbart, and Adams. (2014)

Nested Dropout



Borrowing from *ramping targeted dropout*, instead of sampling indices, we simply anneal upwards from zero over the course of training.

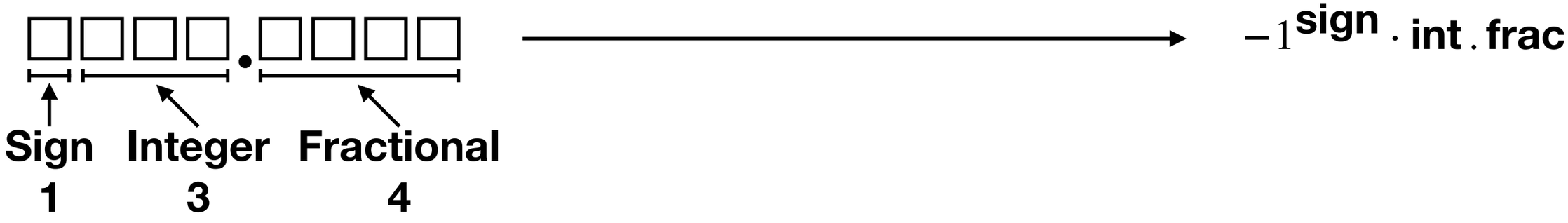
Rippel, Gelbart, and Adams. (2014)

Numeric Representation

Single-precision floating-point (float32)

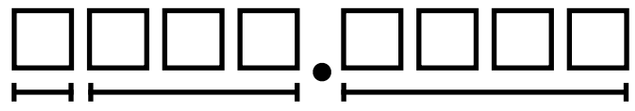
Truncated half-precision floating-point (bfloat16)

8-bit fixed-point (Q3.4)



Exprot

8-bit fixed-point (Q3.4)



Our Q3.4 can’t represent any number with magnitude larger than $8 - 2^{-4}$ or smaller than 2^{-4} (but greater than zero)

Idea: We can gradually constrain the exponents of our numbers to be as close to 0 as possible simply by minimizing $|\log_2(W)|$ and stochastically perturbing the exponent towards zero.

Bitrot

```
def bitrot(inputs, targ_bits, keep_prob, is_training):
  shifted = tf.bitwise.right_shift(inputs, targ_bits)
  rotten = tf.bitwise.left_shift(shifted, targ_bits)

  mask = tf.random_uniform(tf.shape(inputs)) <= keep_prob

  if is_training:
    return inputs * mask + rotten * (1 - mask)
  else:
    return rotten


def exprot(inputs, exp_shift, keep_prob, is_training):
  log2 = tf.log(inputs) / tf.log(2.)
  shift = -tf.sign(log2) * tf.minimum(exp_shift, log2)
  rotten = inputs * 2**(shift)

  mask = tf.random_uniform(tf.shape(inputs)) <= keep_prob

  if is_training:
    return inputs * mask + rotten * (1 - mask), tf.abs(log2)
  else:
    return rotten
```

Bitrot + Exprot Results

	Acc.
Naive (float32)	93.48
Bitrot (4 mant.)	92.19
Bitrot (3 mant.)	87.37
Bitrot + Exprot (Q4.4)	93.36
Bitrot + Exprot (Q3.4)	~20.00

More to come! 🖤💜

More precisely, much more in next couple months.

Still, go give it a try yourself!

Conclusion

- Early pruning
- Better sparsity support

Keep and eye out for the full version of the Targeted Dropout paper in the coming weeks!

Check out for.ai for opportunities to get involved in ML projects!