General DL Training Techniques

- Dead ReLUs
- Vanishing gradients and skip connections
- Batch normalization
- Transfer learning
- Data Augmentation
- Hyperparameter optimization
- If input to last ReLU layer is 0,
  \[ 0 \geq w_{(j_1,j_2),i} \ast z^{4,(j_1,j_2),i} + b \]

- Then we have that
  \[ 0 = \nabla \sigma(w_{(j_1,j_2),i} \ast z^{4,(j_1,j_2),i} + b) \]

- So all back propagated gradients = 0!
Vanishing gradients

- Consider a pipelined DNN

- Associated back-propagation network
Some Useful Facts

- Operator norm of a matrix
  \[ \|A\| = \max_x \{\|Ax\| : \|x\| \leq 1\} \]

- or equivalently
  \[ \|A\| = \max_x \frac{\|Ax\|}{\|x\|} \]

- Singular value decomposition (SVD)
  \[ A = U \Sigma V^t \]
  where \( \Sigma \) is diagonal, \( U \) and \( V \) have orthogonal columns, and \( \sigma_A^* = \max_i \Sigma_{i,i} \).

- Relationship of SVD to matrix norm
  \[ \|A\| = \sigma_A^* \]

- Bound on norm of product of matrices
  \[ A = A_0 A_1 A_2 \cdots A_{M-1} \]
  Then it’s easily shown that
  \[ \|A\| \leq \|A_0\| \|A_1\| \cdots \|A_{M-1}\| \]

- Finally we have that
  \[ \|A\| \leq \sigma_{A_0}^* \sigma_{A_1}^* \cdots \sigma_{A_{M-1}}^* \]
  And that
  \[ \|Ax\| \leq \sigma_{A_0}^* \sigma_{A_1}^* \cdots \sigma_{A_{M-1}}^* \|x\| \]
Vanishing gradients: Analysis

- Associated back-propagation network

\[
g_0 = B_0^t A_1^t A_2^t A_3^t \cdots A_{M-1}^t \epsilon
\]

- Using the relationships from the previous slide, we can bound the gradients norm by

\[
\|g_0\| \leq \sigma_{B_0^*}^* \sigma_{A_1^*}^* \sigma_{A_2^*}^* \cdots \sigma_{A_{M-1}^*}^* \|\epsilon\|
\]

So if \(\sigma_{A_m^*} \leq 1\), then as the network becomes deeper, we have that

\[
\lim_{M \to \infty} \|g_0\| = 0
\]

- Partial Solutions: ResNet, skip connection, LSTM
Skipped Connections

- Skip connection concept

\[ y \rightarrow f_\theta \rightarrow + \rightarrow \hat{x} \]

- Adjoint gradient

\[ (I + A^t) \rightarrow B^t \rightarrow \epsilon \]

\[ g_2 \]
Deep Skipped Connections

- Skip connection concept

![Diagram showing skip connections with functions $f_{1,\theta}$, $f_{2,\theta}$, ..., $f_{M,\theta}$ and input $y$ and output $\hat{x}$]

- Back propagation

![Diagram showing backpropagation with terms $I + A_1^t$, $I + A_2^t$, ..., $I + A_M^t$ and error $\epsilon$]

- Adjoint gradient

$$g_1 = B_1^t(I + A_2^t)\cdots(I + A_M^t)\epsilon$$

since $\sigma^*_{(I-A)} = 1$, we have that $\|g_0\| \leq \sigma^*_{B_0}\|\epsilon\|$ ⇒ fixes vanishing gradient
Denoising ResNet (DN-ResNet)*

- Typical DN-ResNet denoising network

Merge Connections

- **Problem with skip connection**
  - Some information may be lost by simply adding $y$ and $f_\theta(y)$
  - Example, $f_\theta(y) = -y$

- **Merge Connection**
  - Concatenate $y$ and $f_\theta(y)$

*As far as I know, I made up this term.*
U-Net*


Figure taken from https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/
**Batch normalization**

- Renormalize data in batches according to the following rules:

\[
\text{Input: } \text{Values of } x \text{ over a mini-batch: } \mathcal{B} = \{x_1, \ldots, x_m\}; \\
\text{Parameters to be learned: } \gamma, \beta \\
\text{Output: } \{y_i = \text{BN}_{\gamma, \beta}(x_i)\}
\]

\[
\begin{align*}
\mu_B & \leftarrow \frac{1}{m} \sum_{i=1}^{m} x_i & \quad \text{\(\text{// mini-batch mean}\)} \\
\sigma_B^2 & \leftarrow \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_B)^2 & \quad \text{\(\text{// mini-batch variance}\)} \\
\hat{x}_i & \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} & \quad \text{\(\text{// normalize}\)} \\
y_i & \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) & \quad \text{\(\text{// scale and shift}\)}
\end{align*}
\]

**Algorithm 1:** Batch Normalizing Transform, applied to activation \(x\) over a mini-batch.
Weight initialization and regularization
Transfer learning
Data Augmentation
Hyperparameter optimization