

# Understanding Deep Learning Errata

February 15, 2024

Much gratitude to everyone who has pointed out mistakes. If you find a problem not listed here, please contact me via [github](#) or by mailing me at [udlbookmail@gmail.com](mailto:udlbookmail@gmail.com).



# First printing (Dec. 2023)

## 1.1 Errors

These are things that might genuinely confuse you:

- Figure 4.7b had the wrong calculated numbers in it (but pattern is same). Correct version is in figure 1.1 of this document.
- Equation 6.12 (Second printing)

$$\begin{aligned}\mathbf{m}_{t+1} &\leftarrow \beta \cdot \mathbf{m}_t + (1 - \beta) \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\phi_t - \alpha \beta \cdot \mathbf{m}_t]}{\partial \phi} \\ \phi_{t+1} &\leftarrow \phi_t - \alpha \cdot \mathbf{m}_{t+1},\end{aligned}\tag{1.1}$$

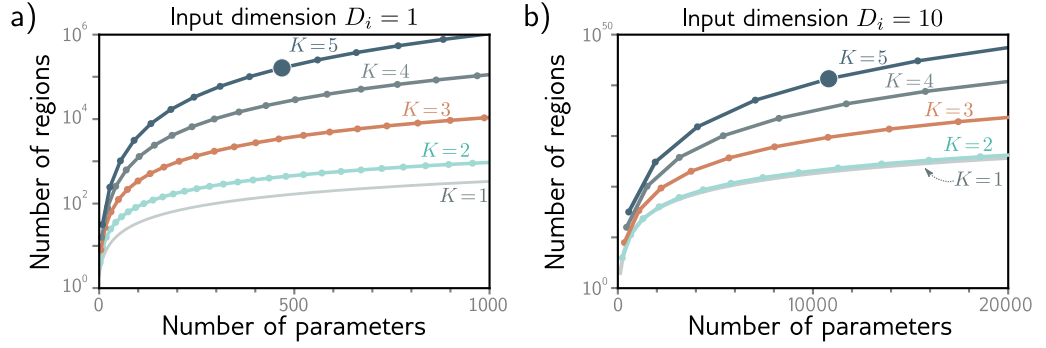
- Section 6.3.1 where now the gradients are evaluated at  $\phi_t - \alpha \beta \cdot \mathbf{m}_t$ .
- Section 7.5.1 The expectation (mean)  $\mathbb{E}[f_i']$  of the intermediate values  $f_i'$  is:
- Equation 15.7 The optimal discriminator **for an example  $\tilde{\mathbf{x}}$**  depends on the underlying probabilities:

$$Pr(\text{real}|\tilde{\mathbf{x}}) = \text{sig}[f[\tilde{\mathbf{x}}, \phi]] = \frac{Pr(\tilde{\mathbf{x}}|\text{real})}{Pr(\tilde{\mathbf{x}}|\text{generated}) + Pr(\tilde{\mathbf{x}}|\text{real})} = \frac{Pr(\mathbf{x})}{Pr(\mathbf{x}^*) + Pr(\mathbf{x})}.\tag{1.2}$$

where on the right hand side, we evaluate  $\tilde{\mathbf{x}}$  against the generated distribution  $Pr(\mathbf{x}^*)$  and the real distribution  $Pr(\mathbf{x})$ .

- Equation 15.9. First integrand should be with respect to  $\mathbf{x}^*$ . Correct version is:

$$\begin{aligned}D_{JS}[Pr(\mathbf{x}^*) || Pr(\mathbf{x})] \\ &= \frac{1}{2} D_{KL} \left[ Pr(\mathbf{x}^*) \left\| \frac{Pr(\mathbf{x}^*) + Pr(\mathbf{x})}{2} \right\| \right] + \frac{1}{2} D_{KL} \left[ Pr(\mathbf{x}) \left\| \frac{Pr(\mathbf{x}^*) + Pr(\mathbf{x})}{2} \right\| \right] \\ &= \frac{1}{2} \int \underbrace{Pr(\mathbf{x}^*) \log \left[ \frac{2Pr(\mathbf{x}^*)}{Pr(\mathbf{x}^*) + Pr(\mathbf{x})} \right]}_{\text{quality}} d\mathbf{x}^* + \frac{1}{2} \int \underbrace{Pr(\mathbf{x}) \log \left[ \frac{2Pr(\mathbf{x})}{Pr(\mathbf{x}^*) + Pr(\mathbf{x})} \right]}_{\text{coverage}} d\mathbf{x}.\end{aligned}$$



**Figure 1.1** Corrected version of figure 4.7: The maximum number of linear regions for neural networks increases rapidly with the network depth. a) Network with  $D_i = 1$  input. Each curve represents a fixed number of hidden layers  $K$ , as we vary the number of hidden units  $D$  per layer. For a fixed parameter budget (horizontal position), deeper networks produce more linear regions than shallower ones. A network with  $K = 5$  layers and  $D = 10$  hidden units per layer has 471 parameters (highlighted point) and can produce 161,051 regions. b) Network with  $D_i = 10$  inputs. Each subsequent point along a curve represents ten hidden units. Here, a model with  $K = 5$  layers and  $D = 50$  hidden units per layer has 10,801 parameters (highlighted point) and can create more than  $10^{40}$  linear regions.

- Equation 15.12.

$$D_w[Pr(x)||q(x)] = \max_{\mathbf{f}} \left[ \sum_i Pr(x=i)f_i - \sum_j q(x=j)f_j \right], \quad (1.3)$$

- Section 15.2.4 Consider distributions  $Pr(x=i)$  and  $q(x=j)$  defined over  $K$  bins. Assume there is a cost  $C_{ij}$  associated with moving one unit of mass from bin  $i$  in the first distribution to bin  $j$  in the second;
- Equation 15.14. Missing bracket and we don't need to use  $\mathbf{x}^*$  notation here. Correct version is:

$$D_w[Pr(\mathbf{x}), q(\mathbf{x})] = \min_{\pi[\bullet, \bullet]} \left[ \iint \pi(\mathbf{x}_1, \mathbf{x}_2) \cdot \|\mathbf{x}_1 - \mathbf{x}_2\| d\mathbf{x}_1 d\mathbf{x}_2 \right].$$

- Equation 15.15. Don't need to use  $\mathbf{x}^*$  notation here, and second term on right hand side should have  $q[\mathbf{x}]$  term not  $Pr(\mathbf{x})$ . Correct version is:

$$D_w[Pr(\mathbf{x}), q(\mathbf{x})] = \max_{f[\mathbf{x}]} \left[ \int Pr(\mathbf{x})f[\mathbf{x}]d\mathbf{x} - \int q(\mathbf{x})f[\mathbf{x}]d\mathbf{x} \right].$$

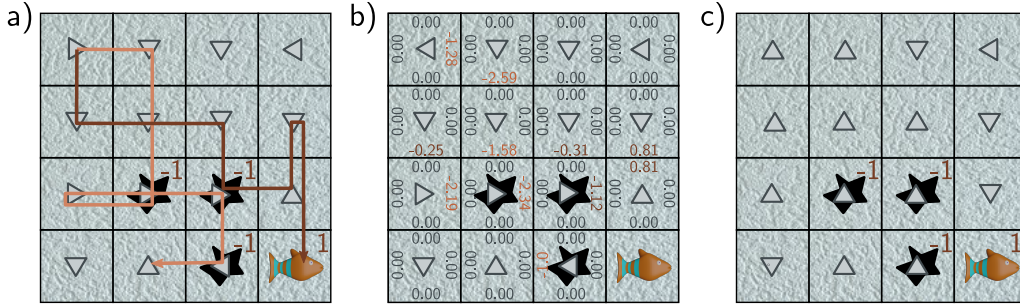


Figure 1.2 Corrected version of figure 19.11

- Equation 16.12 has a mistake in the second term. It should be:

$$f[h_d, \phi] = \left( \sum_{k=1}^{b-1} \phi_k \right) + (hK - b)\phi_b.$$

- Equation 17.34.

$$\begin{aligned} \frac{\partial}{\partial \phi} \mathbb{E}_{Pr(x|\phi)} [f[x]] &= \mathbb{E}_{Pr(x|\phi)} \left[ f[x] \frac{\partial}{\partial \phi} \log [Pr(x|\phi)] \right] \\ &\approx \frac{1}{I} \sum_{i=1}^I f[x_i] \frac{\partial}{\partial \phi} \log [Pr(x_i|\phi)]. \end{aligned}$$

- Figure 19.11 is wrong in that only the state-action values corresponding to the current state-action pair should be moderated. Correct version above.
- Equation B.4. Square root sign should cover  $x$ . Correct version is:

$$x! \approx \sqrt{2\pi x} \left( \frac{x}{e} \right)^x.$$

- Appendix B.3.6. Consider a matrix  $\mathbf{A} \in \mathbb{R}^{D_1 \times D_2}$ . If the number of columns  $D_2$  of the matrix is fewer than the number of rows  $D_1$  (i.e., the matrix is “portrait”),
- Equation C.20. Erroneous minus sign on covariance matrix. Correct version is:

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \mathbf{z}.$$

### 1.1.1 Minor fixes

These are things that are wrong and need to be fixed, but that will probably not affect your understanding (e.g., math symbols that are in bold but should not be).

- Multiple places in Chapters 2-9 (Second printing). Loss functions  $L[\phi]$  sometimes written as  $L[\phi]$ . Have now all been converted to italic for consistency.
- Section 1.1: ...and what is meant by “**training**” a model.
- Figure 1.13: ~~Adapted from Pablok (2017).~~
- Figure 2.3 legend: Each combination of parameters  $\phi = [\phi_0, \phi_1]^T$ .
- Section 2.3: **1D** linear regression has the obvious drawback
- Figure 3.5 legend: The universal approximation theorem proves that, with enough hidden units, there exists a shallow neural network that can describe any given continuous function defined on a compact subset of  $\mathbb{R}^{D_i}$  to arbitrary precision.
- Notes page 38 Most of these are attempts to avoid the dying **ReLU** problem while limiting the gradient for negative values.
- Figure 4.1 legend: The first network maps inputs  $x \in [-1, 1]$  to outputs  $y \in [-1, 1]$  using a function **comprising** three linear regions that are chosen so that they alternate the sign of their slope (**fourth linear region is outside range of graph**).
- Figure 4.2: Colors changed to avoid ambiguity
- Equation 4.13 is missing a prime sign:

$$\begin{aligned}\mathbf{h} &= \mathbf{a}[\boldsymbol{\theta}_0 + \boldsymbol{\theta}x] \\ \mathbf{h}' &= \mathbf{a}[\boldsymbol{\psi}_0 + \boldsymbol{\Psi}\mathbf{h}] \\ y' &= \phi'_0 + \phi'\mathbf{h}',\end{aligned}$$

- Equation 4.14:  $\phi'_0$  should not be bold.

$$y = \phi'_0 + \phi'\mathbf{h}'$$

- Equation 4.17 is not technically wrong, but the product is unnecessary and it's unclear if the last term should be included in it (no). Better written as:

$$N_r = \left(\frac{D}{D_i} + 1\right)^{D_i(K-1)} \cdot \sum_{j=0}^{D_i} \binom{D}{j}.$$

- Equation 5.10. Second line is disambiguated by adding brackets:

$$\begin{aligned}
\hat{\phi} &= \operatorname{argmin}_{\phi} \left[ - \sum_{i=1}^I \log \left[ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ - \frac{(y_i - f[\mathbf{x}_i, \phi])^2}{2\sigma^2} \right] \right] \right] \\
&= \operatorname{argmin}_{\phi} \left[ - \sum_{i=1}^I \left( \log \left[ \frac{1}{\sqrt{2\pi\sigma^2}} \right] - \frac{(y_i - f[\mathbf{x}_i, \phi])^2}{2\sigma^2} \right) \right] \\
&= \operatorname{argmin}_{\phi} \left[ - \sum_{i=1}^I - \frac{(y_i - f[\mathbf{x}_i, \phi])^2}{2\sigma^2} \right] \\
&= \operatorname{argmin}_{\phi} \left[ \sum_{i=1}^I (y_i - f[\mathbf{x}_i, \phi])^2 \right],
\end{aligned}$$

- Equation 5.12. More properly written as:

$$\hat{y} = \operatorname{argmax}_y \left[ Pr(y | \mathbf{f}[\mathbf{x}, \hat{\phi}, \sigma^2]) \right]. \quad (1.4)$$

although the value of  $\sigma^2$  does not actually matter or change the position of the maximum.

- Equation 5.15. Disambiguated by adding brackets:

$$\hat{\phi} = \operatorname{argmin}_{\phi} \left[ - \sum_{i=1}^I \left( \log \left[ \frac{1}{\sqrt{2\pi f_2[\mathbf{x}_i, \phi]^2}} \right] - \frac{(y_i - f_1[\mathbf{x}_i, \phi])^2}{2f_2[\mathbf{x}_i, \phi]^2} \right) \right].$$

- Section 5.5 The likelihood that input  $\mathbf{x}$  has label  $y = k$  (figure 5.10) is hence:
- Section 5.6 Removed  $i$  index from this paragraph for consistency. Independence implies that we treat the probability  $Pr(\mathbf{y} | \mathbf{f}[\mathbf{x}, \phi])$  as a product of univariate terms for each element  $y_d \in \mathbf{y}$ :

$$Pr(\mathbf{y} | \mathbf{f}[\mathbf{x}, \phi]) = \prod_d Pr(y_d | \mathbf{f}_d[\mathbf{x}, \phi]),$$

where  $\mathbf{f}_d[\mathbf{x}, \phi]$  is the  $d^{th}$  set of network outputs, which describe the parameters of the distribution over  $y_d$ . For example, to predict multiple continuous variables  $y_d \in \mathbb{R}$ , we use a normal distribution for each  $y_d$ , and the network outputs  $\mathbf{f}_d[\mathbf{x}, \phi]$  predict the means of these distributions. To predict multiple discrete variables  $y_d \in \{1, 2, \dots, K\}$ , we use a categorical distribution for each  $y_d$ . Here, each set of network outputs  $\mathbf{f}_d[\mathbf{x}, \phi]$  predicts the  $K$  values that contribute to the categorical distribution for  $y_d$ .

- Problem 5.8. Construct a loss function for making multivariate predictions  $\mathbf{y} \in \mathbb{R}^{D_i}$  based on independent normal distributions...

- Notes page 94. However, this is strange since SGD is a special case of Adam (when  $\beta = \gamma = 0$ )
- Section 7.3. The final derivatives from the term  $f_0 = \beta_0 + \omega_0 \cdot x_i$  are:
- Section 7.4. Similarly, the derivative for the weights **matrix**  $\Omega_k$ , is given by
- Section 7.5.1 and the second moment  $\mathbb{E}[h_j^2]$  will be half the variance  $\sigma_f^2$
- Figure 7.8. Not wrong, but changed to “nn.init.kaiming\_**normal**\_(layer\_in.weight)” for compatability with text and to avoid deprecated warning.
- Problem 7.13. (Second printing) For the same function as in problem 7.12, compute the derivative...
- Section 8.3.3 (i.e., with four hidden units and four linear regions **in the range of the data**) + minor changes in text to accommodate extra words
- Figure 8.9 number of hidden units / linear regions **in range of data**
- Section 8.4.1 When the number of parameters is very close to the number of training data examples (figure 8.11**b**)
- Equation 9.6 (Second printing) LHS should be total derivative, not partial derivative:

$$\frac{d\phi}{dt} = -\frac{\partial L}{\partial \phi}. \quad (1.5)$$

- Figure 9.4 (Second printing) Added to legend for panel (a) **Blue point represents global minimum.** Added to legend for panel (b) **Blue point represents global minimum which may now be in a different place from panel (a).**
- Figure 9.5 legend: Effect of learning rate (**LR**) and batch size for 4000 training and 4000 test examples from MNIST-1D (see figure 8.1) for a neural network with two hidden layers. a) Performance is better for large learning rates than for intermediate or small ones. In each case, the number of iterations is **6000/LR**, so each solution has the opportunity to move the same distance.
- Figure 9.9 legend (Second printing) Here we are using full-batch gradient descent, and the model (**from figure 8.4**) fits the data as well as possible, so further training won't remove the kink
- Figure 9.9 legend (Second printing) Consider what happens if we remove the **eighth** hidden unit
- Figure 9.11 legend: a-c) Two sets of parameters (cyan **and gray** curves) sampled from the posterior
- Figure 9.11 legend (Second printing) When the prior variance  $\sigma_\phi^2$  is small



- Equation 9.15 (Second printing) Should be total derivative not partial:

$$\frac{d\phi}{dt} = \mathbf{g}[\phi]. \quad (1.6)$$

- Equation 9.16 (Second printing) Should be total derivative not partial:

$$\frac{d\phi}{dt} \approx \mathbf{g}[\phi] + \alpha \mathbf{g}_1[\phi] + \dots, \quad (1.7)$$

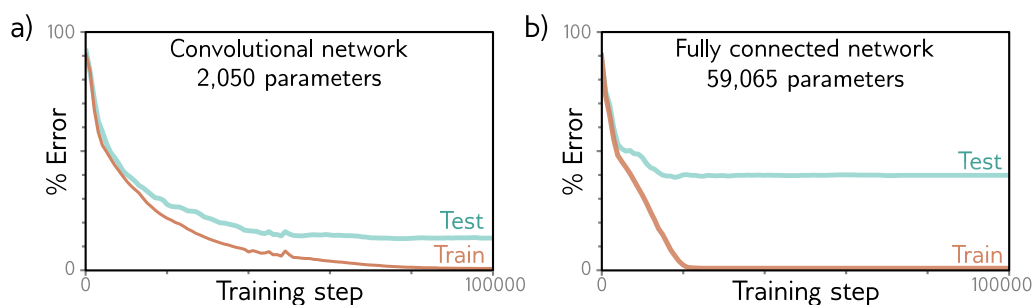
- Equation 9.17 (Second printing) Should be total derivative not partial:

$$\begin{aligned} \phi[\alpha] &\approx \phi + \alpha \frac{d\phi}{dt} + \frac{\alpha^2}{2} \frac{d^2\phi}{dt^2} \Big|_{\phi=\phi_0} \\ &\approx \phi + \alpha (\mathbf{g}[\phi] + \alpha \mathbf{g}_1[\phi]) + \frac{\alpha^2}{2} \left( \frac{\partial \mathbf{g}[\phi]}{\partial \phi} \frac{d\phi}{dt} + \alpha \frac{\partial \mathbf{g}_1[\phi]}{\partial \phi} \frac{d\phi}{dt} \right) \Big|_{\phi=\phi_0} \\ &= \phi + \alpha (\mathbf{g}[\phi] + \alpha \mathbf{g}_1[\phi]) + \frac{\alpha^2}{2} \left( \frac{\partial \mathbf{g}[\phi]}{\partial \phi} \mathbf{g}[\phi] + \alpha \frac{\partial \mathbf{g}_1[\phi]}{\partial \phi} \mathbf{g}[\phi] \right) \Big|_{\phi=\phi_0} \\ &\approx \phi + \alpha \mathbf{g}[\phi] + \alpha^2 \left( \mathbf{g}_1[\phi] + \frac{1}{2} \frac{\partial \mathbf{g}[\phi]}{\partial \phi} \mathbf{g}[\phi] \right) \Big|_{\phi=\phi_0}, \end{aligned} \quad (1.8)$$

- Equation 9.19 (Second printing) Should be total derivative not partial:

$$\begin{aligned} \frac{d\phi}{dt} &\approx \mathbf{g}[\phi] + \alpha \mathbf{g}_1[\phi] \\ &= -\frac{\partial L}{\partial \phi} - \frac{\alpha}{2} \left( \frac{\partial^2 L}{\partial \phi^2} \right) \frac{\partial L}{\partial \phi}. \end{aligned} \quad (1.9)$$

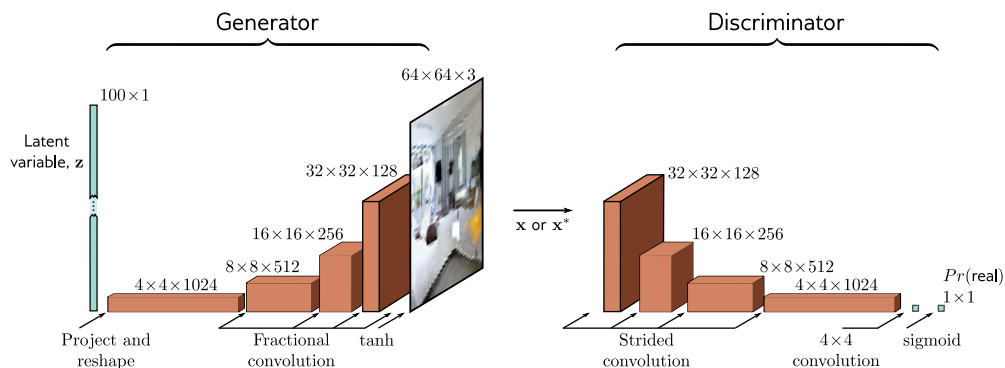
- Page 156 (Second printing) Notes: Wrong marginal reference — Appendix B.3.7 Spectral Norm
- Section 10.2.1 Not wrong, but could be disambiguated: The **size of the** region over which inputs are **combined** is termed the *kernel size*.
- Figure 10.3. The dilation rates are wrong by one, so should be 1,1,1, and 2 in panels a,b,c,d, respectively.
- Figure 10.3 legend (Second printing) – In dilated or atrous convolution (**from the French “à trous” – with holes**), we intersperse zeros in the weight vector...
- Section 10.2.3 The number of zeros we intersperse between the weights **determines** the dilation rate.
- Section 10.2.4 With kernel size three, stride one, and dilation rate **one**.



Corrected version of figure 10.8

- Section 10.2.7 The convolutional network has 2,050 parameters, and the fully connected network has 59,065 parameters.
- Figure 10.8 Number of parameters also wrong in figure 10.8 (correct version in this document). Recalculated curve is slightly different.
- Section 10.5.1 (Second printing) A final max-pooling layer yields a  $6 \times 6$  representation with 256 channels which is resized into a vector of length 9, 216 and passed through three fully connected layers containing 4096, 4096, and 1000 hidden units, respectively.
- Section 10.5.1 (Second printing) The complete network contains  $\sim 60$  million parameters, most of which are in the fully connected layers and the end of the network.
- Section 10.5.3 The first part of the network is a smaller version of VGG (figure 10.17) that contains thirteen rather than sixteen convolutional layers.
- Section 10.6 The weights and the bias are the same at every spatial position, so there are far fewer parameters than in a fully connected network, and the number of parameters doesn't increase with the input image size.
- Problem 10.1 Show that the operation in equation 10.3 is equivariant with respect to translation.
- Problem 10.2 Equation 10.3 defines 1D convolution with a kernel size of three, stride of one, and dilation one.
- Problem 10.3 Write out the equation for the 1D dilated convolution with a kernel size of three and a dilation rate of two.
- Problem 10.4 Write out the equation for a 1D convolution with kernel size of seven, a dilation rate of three, and a stride of three.
- Problem 10.9 A network consists of three 1D convolutional layers. At each layer, a zero-padded convolution with kernel size three, stride one, and dilation one is applied.

- Problem 10.10 A network consists of three 1D convolutional layers. At each layer, a zero-padded convolution with kernel size seven, stride one, and dilation **one** is applied.
- Problem 10.11 Consider a convolutional network with 1D input  $\mathbf{x}$ . The first hidden layer  $\mathbf{H}_1$  is computed using a convolution with kernel size five, stride two, and a dilation rate of **one**. The second hidden layer  $\mathbf{H}_2$  is computed using a convolution with kernel size three, stride one, and a dilation rate of **one**. The third hidden layer  $\mathbf{H}_3$  is computed using a convolution with kernel size five, stride one, and a dilation rate of **two**. What are the receptive field sizes at each hidden layer?
- Legend to figure 11.15. Computational graph for batch normalization (see problem **11.5**).
- Section 12.2: Not a mistake, but this is clearer: where  $\beta_v \in \mathbb{R}^D$  and  $\Omega_v \in \mathbb{R}^{D \times D}$  represent biases and weights, respectively.
- Section 12.3.3 to make **self-attention** work well
- Section 12.4 Title changed to **Transformer layers**
- Section 12.4 a larger transformer
- Section 12.4 a series of these transformer **layers** ...
- Section 12.5 The previous section described the transformer **layer**... a series of transformer **layers**...
- Figure 12.8 legend: The transformer  $\rightarrow$  **Transformer layer**...The transformer **layer** consists
- Figure 12.8 has some minor mistakes in the calculation. The corrected version is shown at the end of this document.
- Figure 12.8 legend. At each iteration, the sub-word tokenizer looks for the most commonly occurring adjacent pair of **tokens**
- Section 12.5.3 a series of  $K$  transformer **layers**
- Section 12.6 through 24 **transformer layers**
- Section 12.6 in the fully connected networks ~~in the transformer~~ is 4096
- Figure 12.10 legend: a series of transformer **layers**
- Figure 12.10 legend (second printing): A small fraction of the input tokens **are** randomly replaced with...
- Section 12.7.2 the **transformer layers** use masked...
- Figure 12.12 are passed through a series of **transformer layers**... and **those** of tokens earlier



Corrected version of figure 15.3

- Section 12.7.4 There are 96 **transformer layers**
- Section 12.7 comprises a series of transformer **layers**
- Section 12.8 **Originally, these**
- Section 12.8 a series of transformer **layers**... a series of transformer **layers**
- Section 13.5.1 Given  $I$  training graphs  $\{\mathbf{X}_i, \mathbf{A}_i\}$  and their labels  $y_i$ , the parameters  $\Phi = \{\beta_k, \Omega_k\}_{k=0}^K$  can be learned using SGD...
- Figure 13.10 (Second printing) Right hand column should be labelled as "**output**" not "hidden layer two".
- Figure 15.3 legend: At the end is a **tan h** function that maps the...
- Figure 15.3  $\arctan \rightarrow \tanh$ . Corrected version nearby in this document.
- Section 15.1.3: At the final layer, the  $64 \times 64 \times 3$  signal is passed through a **tan h** function to generate an image  $\mathbf{x}^*$
- Equation 15.6. Minor problems with brackets in this equation. Should be:

$$\begin{aligned}
 L[\phi] &= \frac{1}{J} \sum_{j=1}^J \left( \log \left[ 1 - \text{sig}[f[\mathbf{x}_j^*, \phi]] \right] \right) + \frac{1}{I} \sum_{i=1}^I \left( \log \left[ \text{sig}[f[\mathbf{x}_i, \phi]] \right] \right) \\
 &\approx \mathbb{E}_{\mathbf{x}^*} \left[ \log \left[ 1 - \text{sig}[f[\mathbf{x}^*, \phi]] \right] \right] + \mathbb{E}_{\mathbf{x}} \left[ \log \left[ \text{sig}[f[\mathbf{x}, \phi]] \right] \right] \\
 &= \int Pr(\mathbf{x}^*) \log \left[ 1 - \text{sig}[f[\mathbf{x}^*, \phi]] \right] d\mathbf{x}^* + \int Pr(\mathbf{x}) \log \left[ \text{sig}[f[\mathbf{x}, \phi]] \right] d\mathbf{x}.
 \end{aligned}$$

- Equation 16.2 (last line). For some reason, this didn't print properly, although it looks fine in my original pdf. Should be:

$$\begin{aligned}\hat{\phi} &= \operatorname{argmax}_{\phi} \left[ \prod_{i=1}^I Pr(x_i|\phi) \right] \\ &= \operatorname{argmin}_{\phi} \left[ \sum_{i=1}^I -\log [Pr(x_i|\phi)] \right] \\ &= \operatorname{argmin}_{\phi} \left[ \sum_{i=1}^I \log \left[ \left| \frac{\partial f[z_i, \phi]}{\partial z_i} \right| \right] - \log [Pr(z_i)] \right],\end{aligned}$$

- Equation 16.25.  $\phi$  should change to  $\hat{\phi}$  on left hand side. Correct version is:

$$\hat{\phi} = \operatorname{argmin}_{\phi} \left[ \text{KL} \left[ \sum_{i=1}^I \delta[\mathbf{x} - \mathbf{f}[\mathbf{z}_i, \phi]] \middle| \middle| q(\mathbf{x}) \right] \right].$$

- Equation 16.26.  $\phi$  should change to  $\hat{\phi}$  on left hand side. Correct version is:

$$\hat{\phi} = \operatorname{argmin}_{\phi} \left[ \text{KL} \left[ \frac{1}{I} \sum_{i=1}^I \delta[\mathbf{x} - \mathbf{x}_i] \middle| \middle| Pr(\mathbf{x}_i, \phi) \right] \right].$$

- Equation 18.24 has a minor formatting mistake. Better written as:

$$\begin{aligned}\log \left[ \frac{Pr(\mathbf{x}, \mathbf{z}_{1...T} | \phi_{1...T})}{q(\mathbf{z}_{1...T} | \mathbf{x})} \right] \\ &= \log \left[ \frac{Pr(\mathbf{x} | \mathbf{z}_1, \phi_1)}{q(\mathbf{z}_1 | \mathbf{x})} \right] + \log \left[ \frac{\prod_{t=2}^T Pr(\mathbf{z}_{t-1} | \mathbf{z}_t, \phi_t) \cdot q(\mathbf{z}_{t-1} | \mathbf{x})}{\prod_{t=2}^T q(\mathbf{z}_{t-1} | \mathbf{z}_t, \mathbf{x}) \cdot q(\mathbf{z}_t | \mathbf{x})} \right] + \log [Pr(\mathbf{z}_T)] \\ &= \log [Pr(\mathbf{x} | \mathbf{z}_1, \phi_1)] + \log \left[ \frac{\prod_{t=2}^T Pr(\mathbf{z}_{t-1} | \mathbf{z}_t, \phi_t)}{\prod_{t=2}^T q(\mathbf{z}_{t-1} | \mathbf{z}_t, \mathbf{x})} \right] + \log \left[ \frac{Pr(\mathbf{z}_T)}{q(\mathbf{z}_T | \mathbf{x})} \right] \\ &\approx \log [Pr(\mathbf{x} | \mathbf{z}_1, \phi_1)] + \sum_{t=2}^T \log \left[ \frac{Pr(\mathbf{z}_{t-1} | \mathbf{z}_t, \phi_t)}{q(\mathbf{z}_{t-1} | \mathbf{z}_t, \mathbf{x})} \right],\end{aligned}$$

- Equation 18.34 missing indices on noise term:

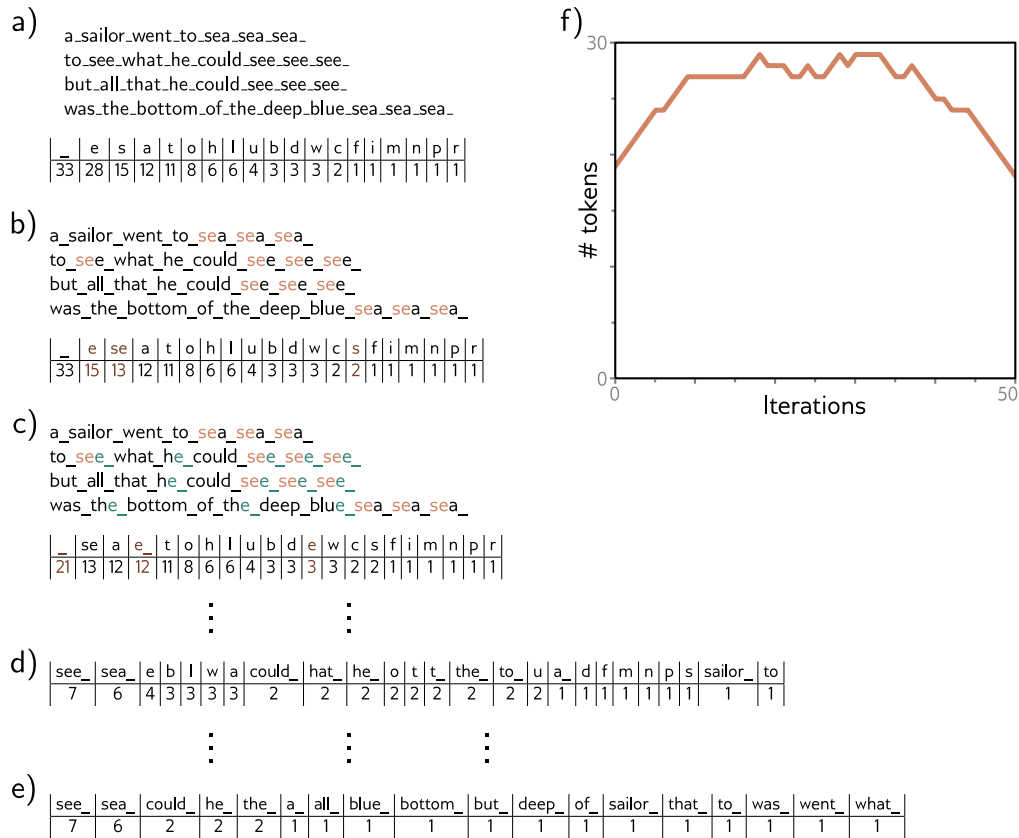
$$\begin{aligned}L[\phi_{1...T}] &= \sum_{i=1}^I -\log [\text{Norm}_{\mathbf{x}_i} [\mathbf{f}_1[\mathbf{z}_{i1}, \phi_1], \sigma_1^2 \mathbf{I}]] \\ &\quad + \sum_{t=2}^T \frac{1}{2\sigma_t^2} \left\| \left( \frac{1}{\sqrt{1-\beta_t}} \mathbf{z}_{it} - \frac{\beta_t}{\sqrt{1-\alpha_t}\sqrt{1-\beta_t}} \boldsymbol{\epsilon}_{it} \right) - \mathbf{f}_t[\mathbf{z}_{it}, \phi_t] \right\|^2.\end{aligned}\tag{1.10}$$

- Section 18.6.1 (Second printing) The obvious architectural choice for this image-to-image mapping is the U-Net (figure 11.10).
- Section 20.2.2 Another possible explanation for the ease with which models are trained is that some regularization methods like L2 regularization (weight decay) make the loss surface flatter and more convex.
- Section 20.2.4 For example, Du et al. (2019a) show that residual networks converge to zero training loss when the width of the network  $D$  (i.e., the number of hidden units) is  $\Omega[I^4 K^2]$  where  $I$  is the amount of training data, and  $K$  is the depth of the network.
- Section 20.5.1 (Second printing) In general, the smaller the model, the larger the proportion of weights that can... Note that this statement is only true for pure pruning methods, and not for lottery tickets where the pruned network is retrained from scratch, and so it has been removed.
- Section 21.7 the Conference on AI, Ethics, and Society
- Appendix A. The notation  $\{0, 1, 2, \dots\}$  denotes the set of non-negative integers.
- Appendix A ...big- $O$  notation, which represents an upper bound...
- Appendix A.  $f[n] < c \cdot g[n]$  for all  $n > n_0$
- Equation B. 18

$$\begin{aligned}
 y_1 &= \phi_{10} + \phi_{11}z_1 + \phi_{12}z_2 + \phi_{13}z_3 \\
 y_2 &= \phi_{20} + \phi_{21}z_1 + \phi_{22}z_2 + \phi_{23}z_3 \\
 y_3 &= \phi_{30} + \phi_{31}z_1 + \phi_{32}z_2 + \phi_{33}z_3.
 \end{aligned} \tag{1.11}$$

- Appendix C.5.4 Accent in wrong place: The Fréchet and Wasserstein distances...
- Equation C.32.

$$\begin{aligned}
 D_{KL} \left[ \text{Norm}[\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1] \middle| \middle| \text{Norm}[\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2] \right] &= \\
 \frac{1}{2} \left( \log \left[ \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} \right] - D + \text{tr} \left[ \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1 \right] + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \right).
 \end{aligned}$$



Corrected version of figure 12.8