Top Equations in Deep Learning

March 2, 2020

References

[1] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

[3] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[4] D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2013.

[5] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. , 323(6088):533–536, Oct. 1986.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.

	Equation		Use	Citation
1	Stochastic Gradient Descent	$\theta_{i+1} \leftarrow \theta_i - \alpha \sum_{x \in B} \frac{\partial L(x)}{\partial \theta}$	Optimization	
2	Backpropagation	$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial \theta}$	Computing Gradients	[5]
3	Linear Layer	$y = Wx + b$	Base Layer in Deep Learning	
4	Convolutions	$(f * g)[n] = \sum_{m=0}^{K} f[n-m] \cdot g[m]$	Image Networks	
5	ReLU	$\mathrm{ReLU}\,(x) = \max(0, x)$	Activation Function	[1]
6	Recurrent Cells	$h_{t+1} = \tanh\left(W_{ih} x_t + b_{ih} + W_{hh} h_t + b_{hh}\right)$	Sequence Modeling	[5]
7	Self Attention	$\mathrm{Attention}\,(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$	Transformers	[6]
8	Generative Adversarial Netowrks	$\min_G \max_D \underset{x \sim p_{data}(x)}{\mathbb{E}} [\log D(x)] + \underset{z \sim p_z(z)}{\mathbb{E}} [\log\left(1 - D(G(z))\right)]$	Generative Modeling	[2]
9	REINFORCE	$\triangle\theta = \alpha r \frac{\partial \log p\left(a \mid \pi^\theta(s)\right)}{\partial \theta}$	Reinforcement Learning	
10	(Variational) Auto-Encoder		Liklihood Modeling	[4]
11	Categorical Cross Entropy	$L\,(x, y) = -x_y + \log\left(\sum_j \exp\left(x_j\right)\right)$	Multiclass Classification	
12	Bellman Equation	$V(s_t) = \max_{a_t}\left(R(s_t, a_t) + \gamma \sum_{s_{t+1}} p(s_t, a_t, s_{t+1}) V(s_{t+1})\right)$ $Q(s_t, a_t) = Q(s_t, a_t) + \alpha \cdot \left(R(s_t, a_t) + \gamma \cdot \max_a Q\left(s_{t+1}, a\right) - Q\left(s_t, a_t\right)\right)$	Reinforcement Learning	
13	Reparametrization Trick	$\underset{z \sim p(z\|\theta)}{\mathbb{E}} [f(z)] = \underset{\epsilon \sim p(\epsilon)}{\mathbb{E}} [f(g(\epsilon, \theta))]$	Differential Sampling	
14	ℓ_p Norm	$\\|x\\|_p = \left(\sum_{i=1}^{n} \|x_i\|^p\right)^{1/p} \qquad \\|A\\|_F \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \|a_{ij}\|^2}$	Distance, Regression Loss	
15	Batch Normalization	$y = \frac{x - \mathbb{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta$	Regularization	[3]
16	Mutual Information			

Table 1