Unnoteworthy Necessities in Neural Network

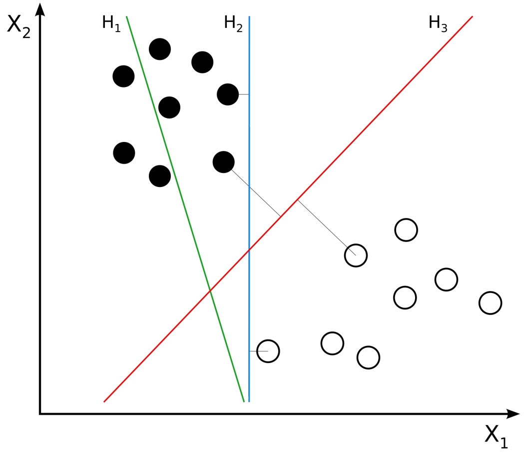
Activation Functions

Hong Kong Deep Learning

Vincent Chan, 2017/03/23

Linear Perceptron

$$
\begin{aligned}
\mathbf{x}_{i+1} &= \mathbf{W}_i\mathbf{x}_i + \mathbf{b}_i \\
&= \mathbf{W}_i(\mathbf{W}_{i-1}\mathbf{x}_{i-1} + \mathbf{b}_{i-1}) + \mathbf{b}_i \\
&= (\mathbf{W}_i\mathbf{W}_{i-1})\mathbf{x}_{i-1} + (\mathbf{W}_i\mathbf{b}_{i-1} + \mathbf{b}_i) \\
&= \mathbf{W}'\mathbf{x}_{i-1} + \mathbf{b}' \\
\Rightarrow \text{output} &= \mathbf{W}''\mathbf{x}_0 + \mathbf{b}''
\end{aligned}
$$



Multi-Layer Perceptron

$$
\begin{aligned}
\mathbf{x}_{i+1} &= \varphi(\mathbf{W}_i \mathbf{x}_i + \mathbf{b}_i) \\
&= \varphi(\mathbf{W}_i \varphi(\mathbf{W}_{i-1} \mathbf{x}_{i-1} + \mathbf{b}_{i-1}) + \mathbf{b}_i)
\end{aligned}
$$

Universal Approximability (Hornik, 1991)

Let φ be a **continuous**, **bounded**, and **non-constant** function. Then,

$$F(\mathbf{x}) = \sum_{i=1}^{N} v_i \varphi\left(\mathbf{w}_i^T \mathbf{x} + b_i\right)$$

is an approximate realization of the function f where f is independent of φ such that

$$|F(\mathbf{x}) - f(\mathbf{x})| < \varepsilon$$

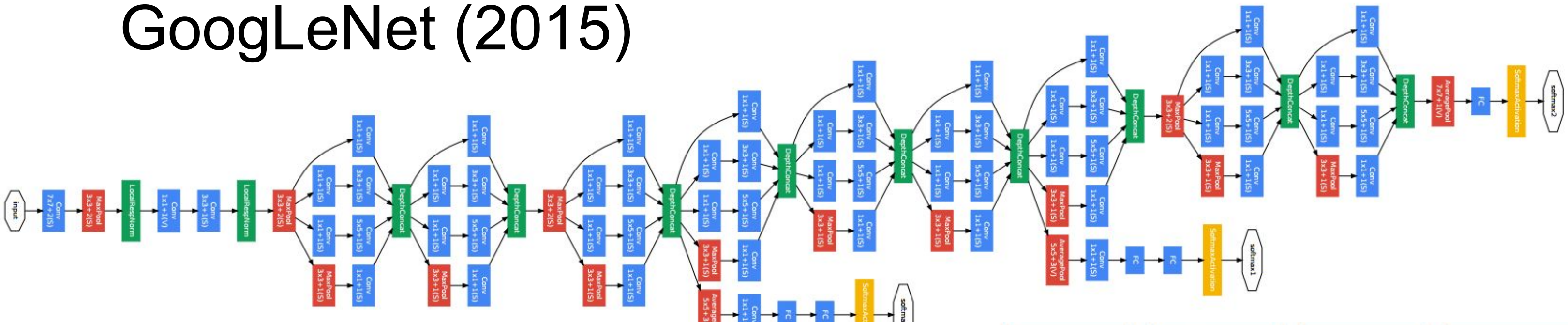
Universal Approximability (Sonoda, 2015)

Let φ belongs to **Lizorkin distribution** space.
Then,

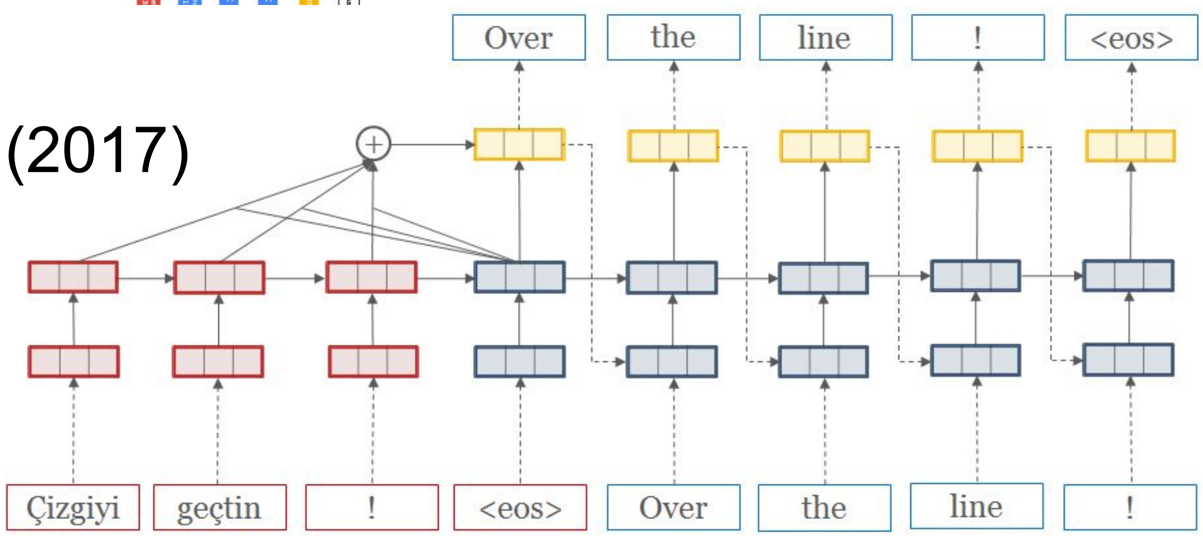
$$\begin{aligned} F(\mathbf{x}) &= \sum_{i=1}^{N} v_i \varphi\left(\mathbf{w}_i^T \mathbf{x} + b_i\right) \\ &\approx \int_{\mathbb{R}^m \times \mathbb{R}} T(\mathbf{w}, b) \varphi(\mathbf{w}^T \mathbf{x} + b) d\mu(\mathbf{w}, b) \\ &= f(\mathbf{x}) \end{aligned}$$

Towards Deep Learning

GoogLeNet (2015)



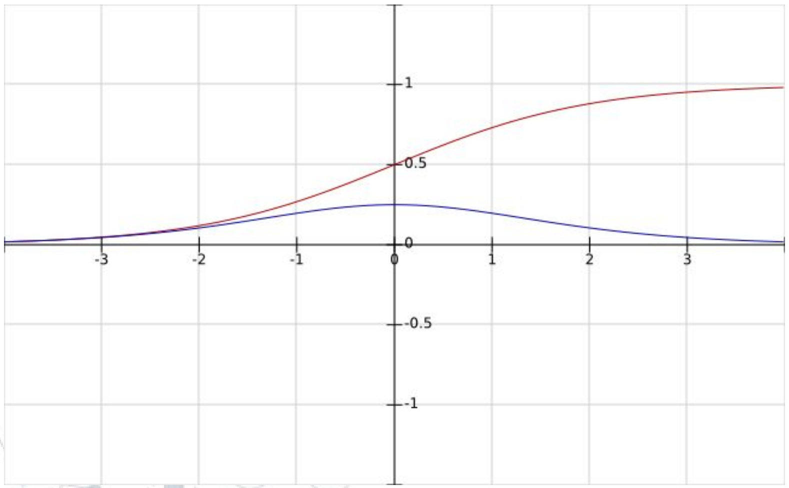
OpenNMT (2017)

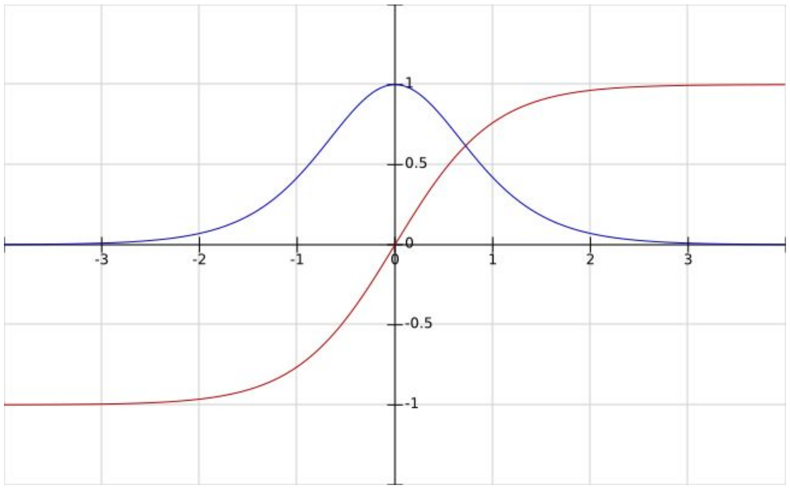


Vanishing Gradient Problem

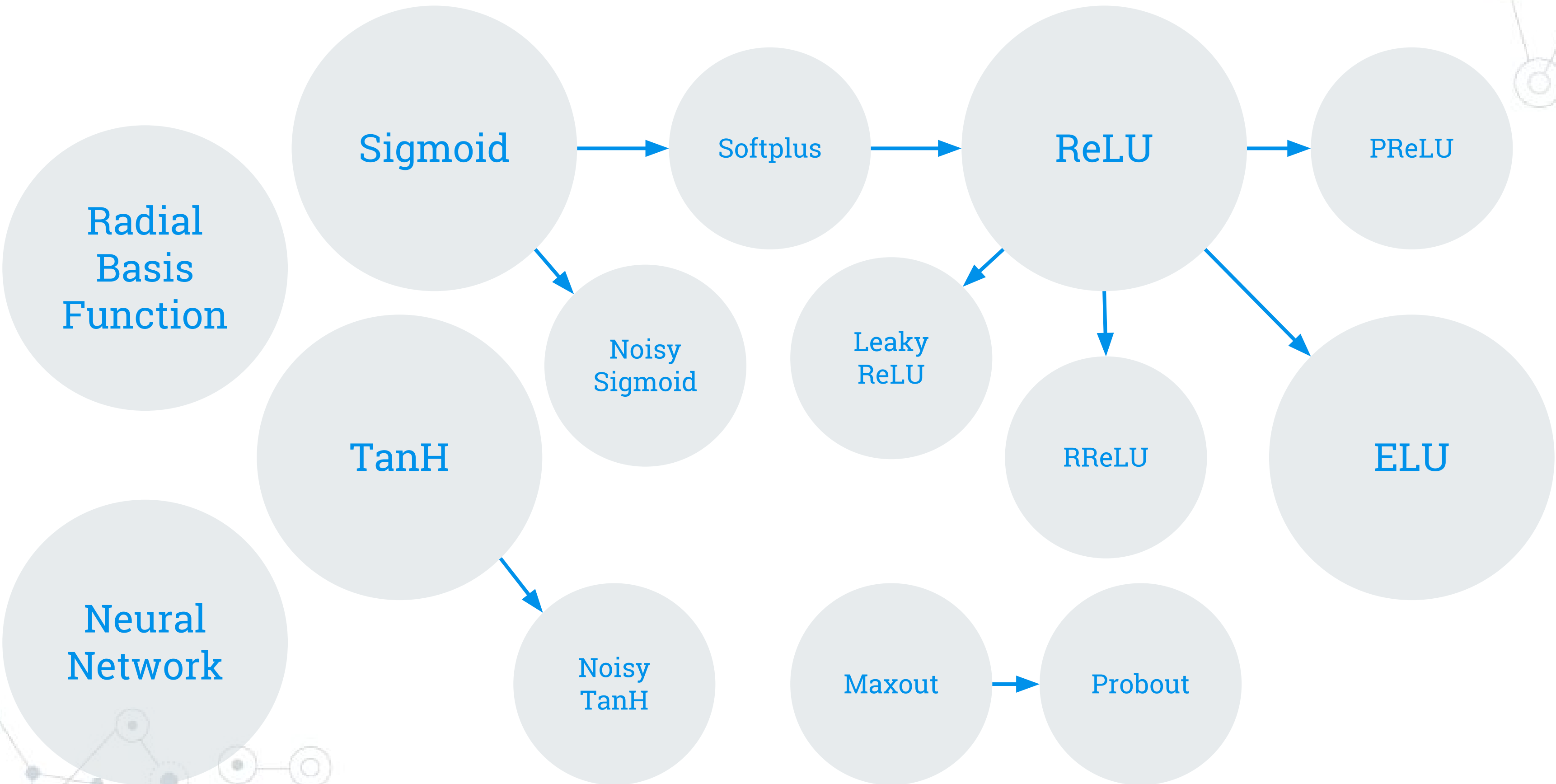
$$|\Delta \mathbf{W}_i| = \alpha \frac{\partial J}{\partial \varphi(\mathbf{net}_L)} \prod_{k=i+1}^{L} \left(\varphi'(\mathbf{net}_k)\mathbf{W}_k\right) \varphi'(\mathbf{net}_i)\mathbf{x}_i$$

$$\varphi'(x) < C \Rightarrow \frac{|\Delta \mathbf{W}_L|}{|\Delta \mathbf{W}_i|} < C^{L-i}$$





Family of Activation Functions



(Hypothesized) Factors on Training

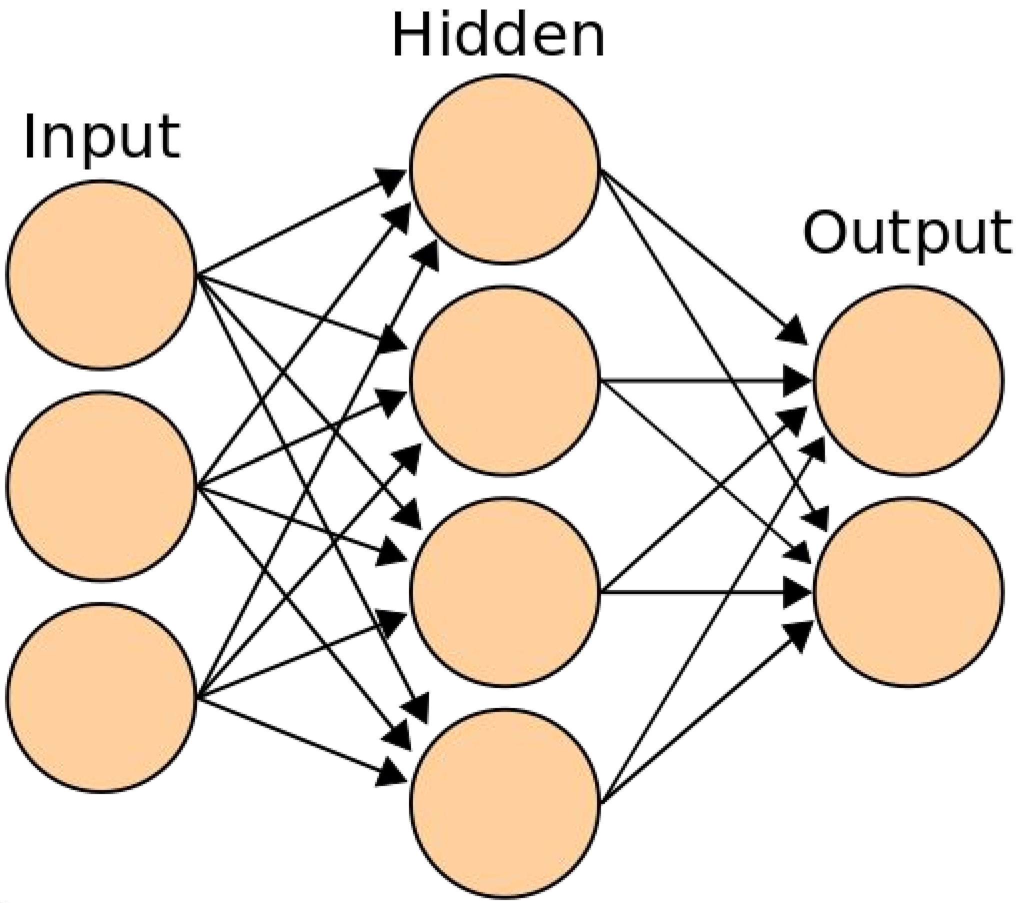
- Representation of information
  - Disentangling
  - Effective variable size
  - Dying neurons
- Internal covariate shift
- Stochasticity
  - Escaping local minima
  - Reducing overfitting
- Computational complexity

Improving Activation Functions

- Randomization
- Self adjustability
- Approximation (rectification / smoothing)
- Violating all seen so far!

Q & A

Neural Network Structure



Backpropagation

$$J = f\left(\varphi\left(\mathbf{W}_i \mathbf{x}_i + \mathbf{b}_i\right)\right)$$

$$\frac{\partial J}{\partial \mathbf{W}_i} = \frac{\partial J}{\partial \varphi(\mathbf{net}_L)} \prod_{k=i+1}^{L} \left(\frac{\partial \varphi(\mathbf{net}_k)}{\partial \mathbf{net}_k} \frac{\partial \mathbf{net}_k}{\partial \varphi(\mathbf{net}_{k-1})} \right) \frac{\partial \varphi(\mathbf{net}_i)}{\partial \mathbf{net}_i} \frac{\partial \mathbf{net}_i}{\partial \mathbf{W}_i}$$

$$\Delta \mathbf{W}_i = -\alpha \frac{\partial J}{\partial \mathbf{W}_i} = -\alpha \frac{\partial J}{\partial \varphi(\mathbf{net}_L)} \prod_{k=i+1}^{L} \left(\varphi'(\mathbf{net}_k) \mathbf{W}_k \right) \varphi'(\mathbf{net}_i) \mathbf{x}_i$$

Universal Approximability (Cybenko, 1989)

Let φ be a **continuous**, and **sigmoidal** function, and ε>0. We may define

$$F(\mathbf{x}) = \sum_{i=1}^{N} v_i \varphi\left(\mathbf{w}_i^T \mathbf{x} + b_i\right)$$

as an approximate realization of the function f where f is independent of φ; that is,

$$|F(\mathbf{x}) - f(\mathbf{x})| < \varepsilon$$

References

- Universal approximability
  - Hornik, Kurt. "Approximation capabilities of multilayer feedforward networks." *Neural networks* 4.2 (1991): 251-257.
  - Sonoda, Sho, and Noboru Murata. "Neural network with unbounded activation functions is universal approximator." *Applied and Computational Harmonic Analysis* (2015).
- Vanishing gradient problem
  - Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks." *Proceedings of The 30th International Conference on Machine Learning*. 2013.

References

ReLUs / ELU

- Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Deep Sparse Rectifier Neural Networks." *International Conference on Artificial Intelligence and Statistics*. 2011.
- Xu, Bing, et al. "Empirical evaluation of rectified activations in convolutional network." *arXiv preprint arXiv:1505.00853* (2015).
- Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter. "Fast and accurate deep network learning by exponential linear units (elus)." *arXiv preprint arXiv:1511.07289* (2015).

Maxout / Probout

- Goodfellow, Ian, et al. "Maxout Networks." *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 2013.
- Springenberg, Jost Tobias, and Martin Riedmiller. "Improving deep neural networks with probabilistic maxout units." *arXiv preprint arXiv:1312.6116* (2013).

References

- Network in network
  - Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." *Proceedings of the 2nd International Conference on Learning Representations*. (2014).
- Noisy activation function
  - Gulcehre, Caglar, et al. "Noisy Activation Functions." *Proceedings of The 33rd International Conference on Machine Learning*. 2016.