Advanced Structures

Kuan-Yu Chen (陳冠宇)

2018/05/10 @ NTUST

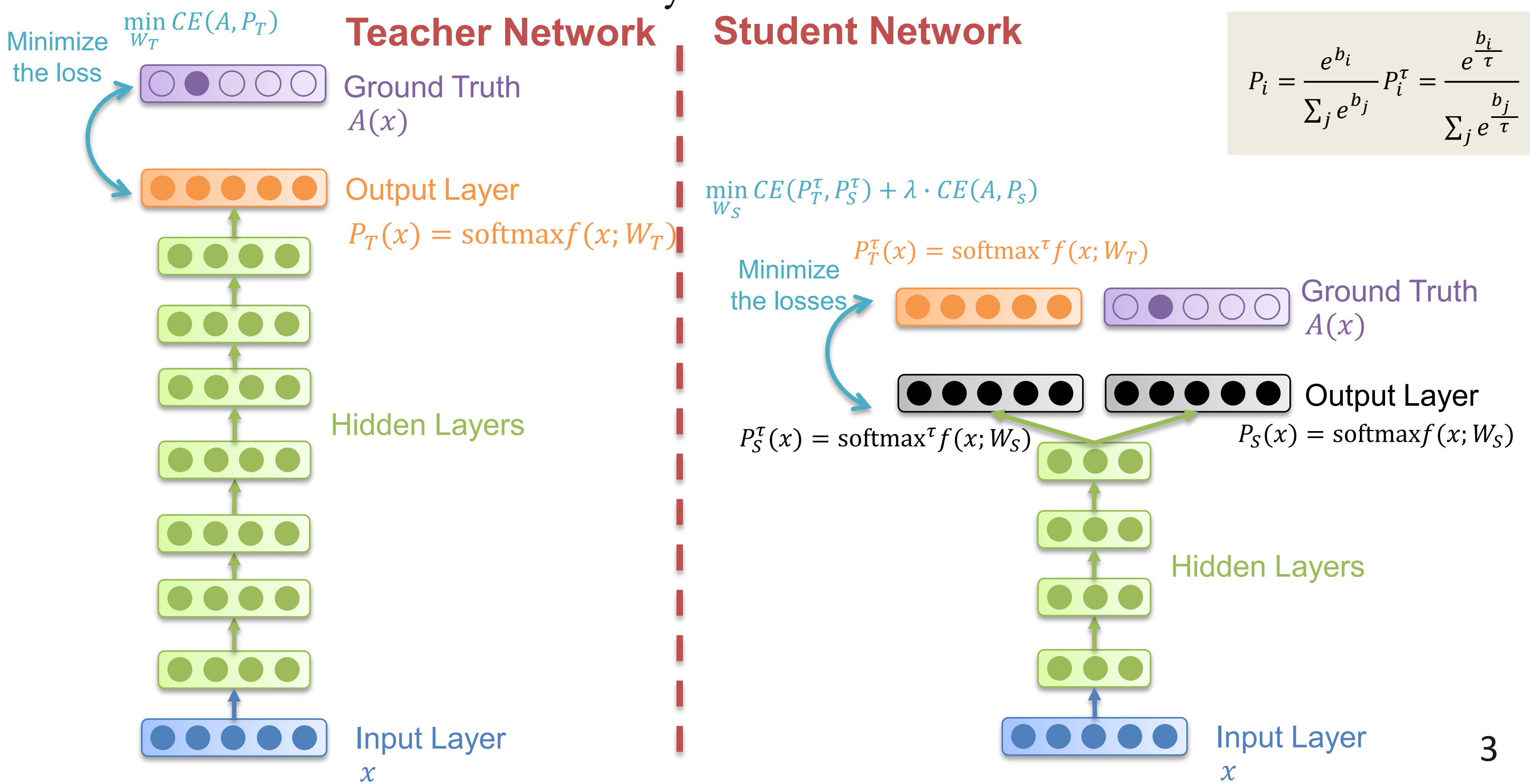
Model Compression

- Deep networks have recently exhibited state-of-the-art performance in computer vision tasks such as image classification and object detection
  - Top-performing systems usually involve very wide and deep networks, with numerous parameters
    - time consuming
    - high memory demanding
  - Knowledge Distillation (Teacher-student network) and FitNet are representatives

G. Hinton and J. Dean, “Distilling knowledge in a neural network,” in *Proceedings of NIPS*, 2014.

Knowledge Distilling

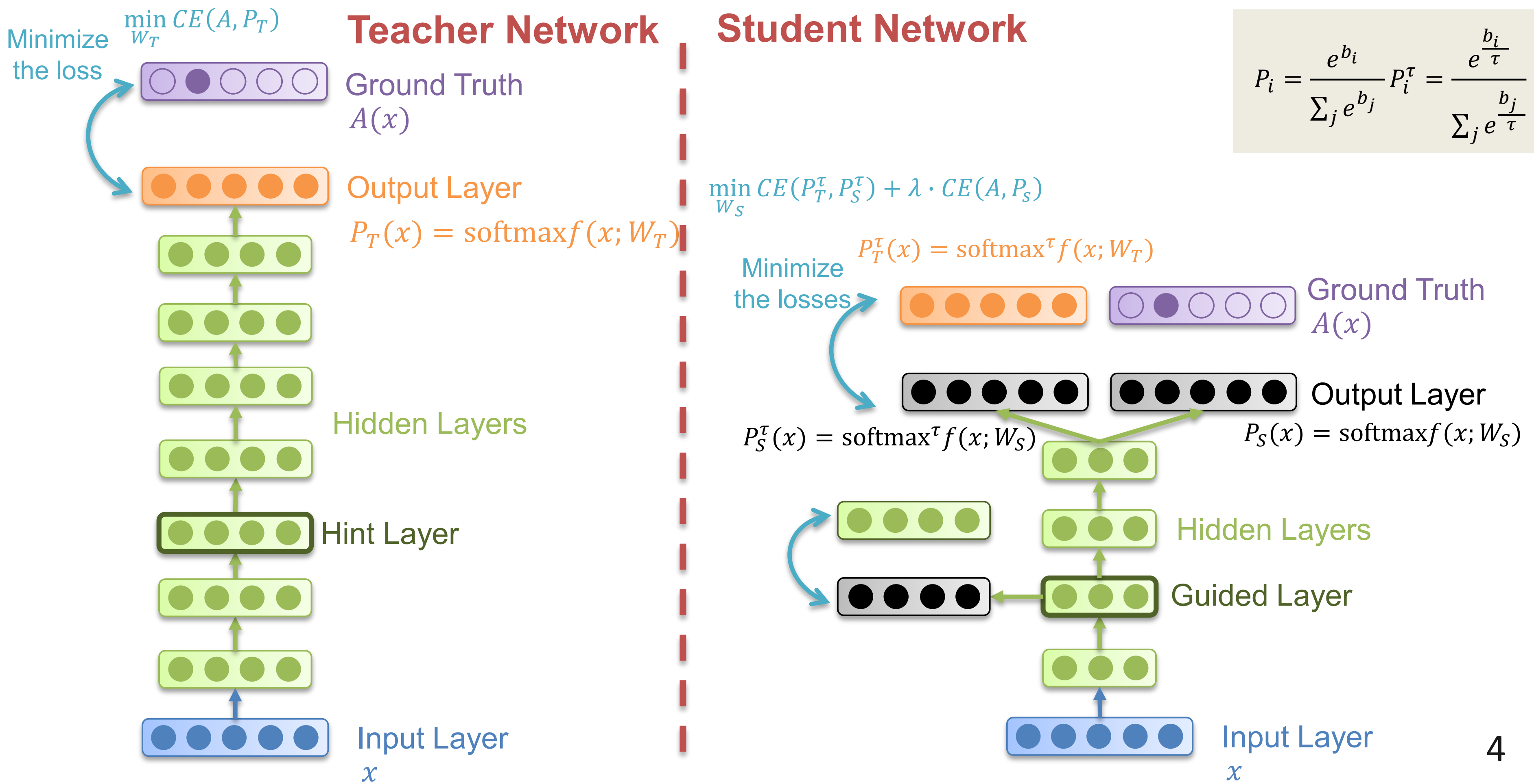
- The idea is to allow the student network to capture not only the information provided by the **true labels**, but also the finer structure learned by the **teacher network**



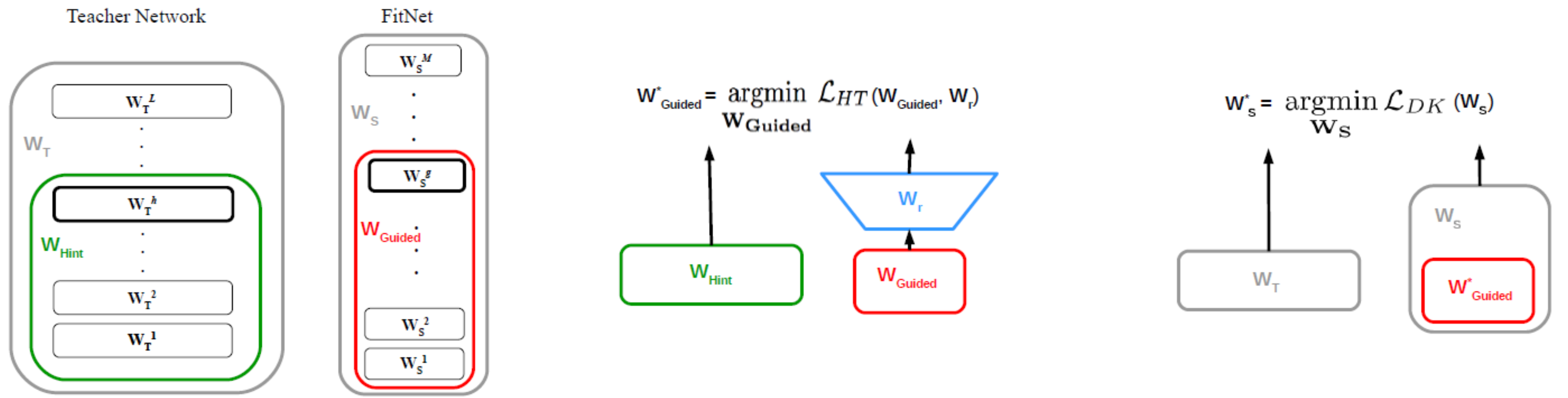
A. Romero, et al., “Fitnets: Hints for thin deep nets,” in *Proceedings of ICLR*, 2015.

FitNets.

- The FitNet is trained in a stage-wise fashion
  - The core idea is that layer-wise information should also be obtained



FitNets..



(a) Teacher and Student Networks (b) Hints Training (c) Knowledge Distillation

1. Starting from a trained teacher network and a randomly initialized FitNet
2. Adding a regressor W_r on top of the FitNet guided layer and train the FitNet parameters W_{Guided}
3. Based on the pre-trained parameters W_{Guided}, we train the parameters of whole FitNet, W_S

Highway Networks.

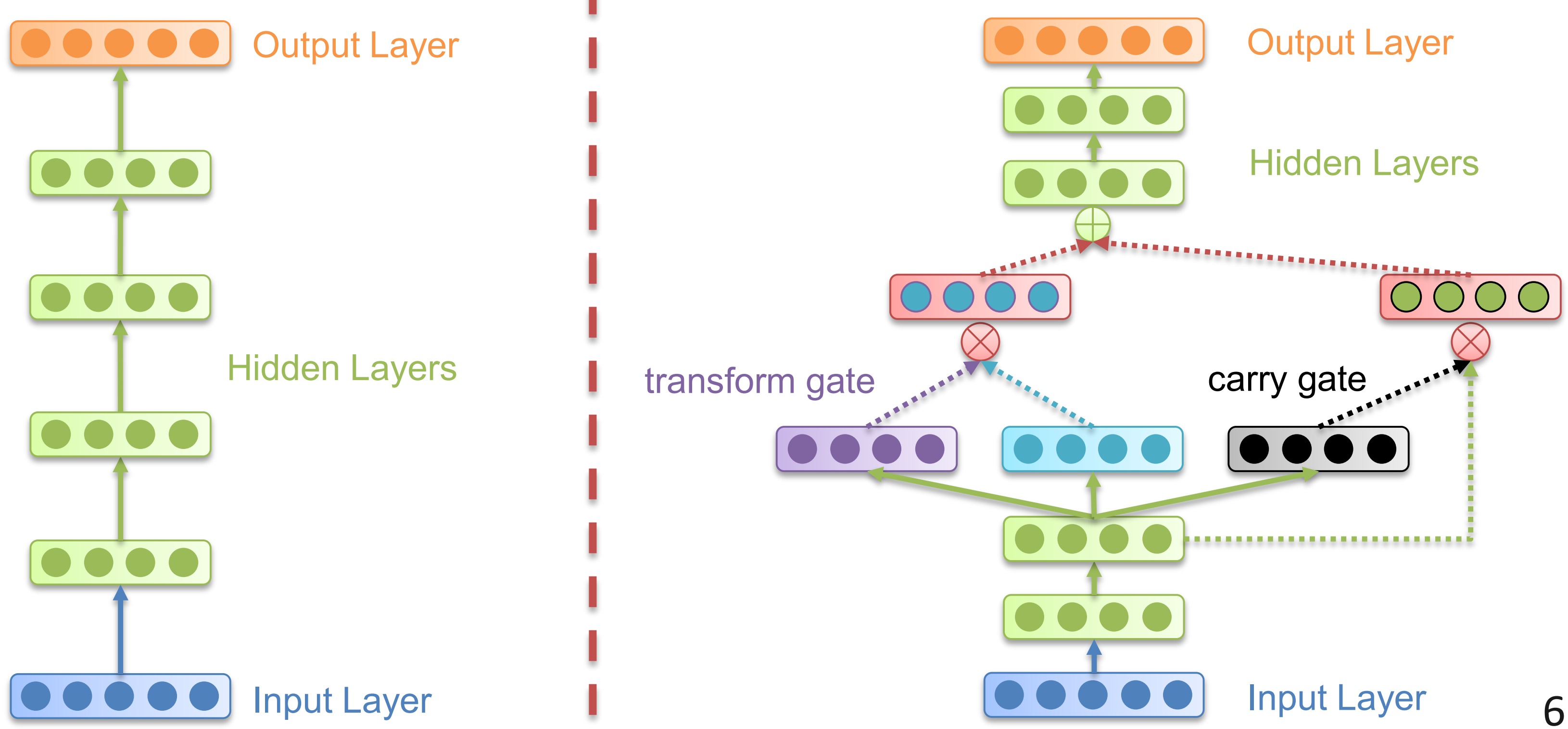
- Highway networks allow unimpeded information flow across several layers on information highways

Vanilla Networks

$$x_{t+1} = H(x_t, W_H)$$

Highway Networks

$$x_{t+1} = H(x_t, W_H) \cdot T(x_t, W_T) + x_t \cdot C(x_t, W_C)$$



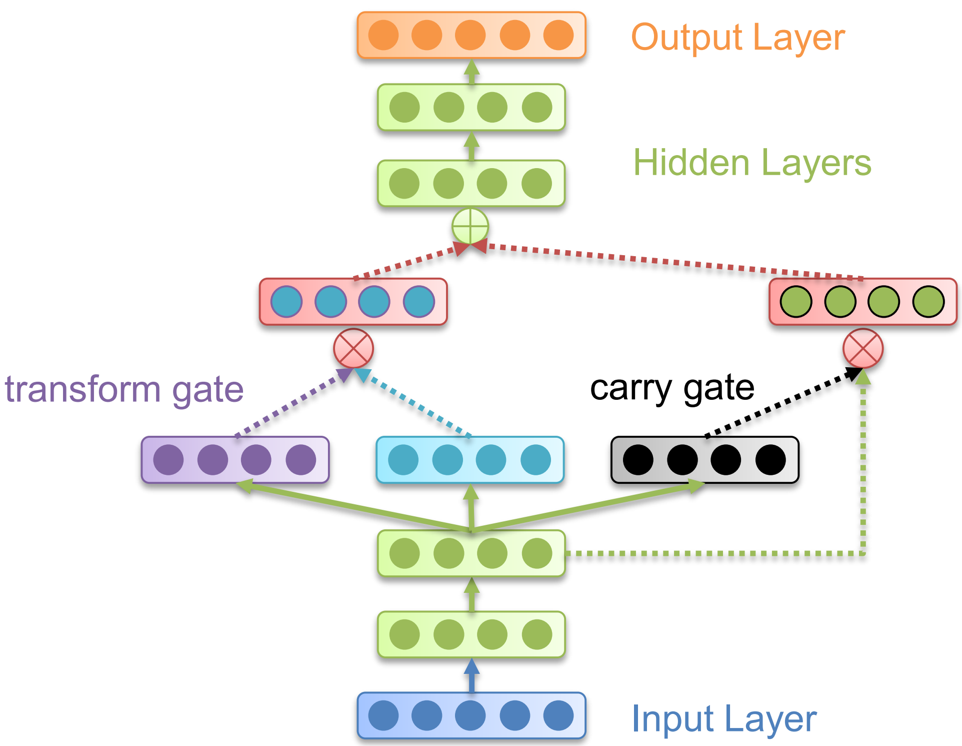
R. Srivastava, et al., "Training very deep networks," in *Proceedings of NIPS*, 2015.

Highway Networks..

- A simplified variant is to set carry gate equal to one minus transform gate

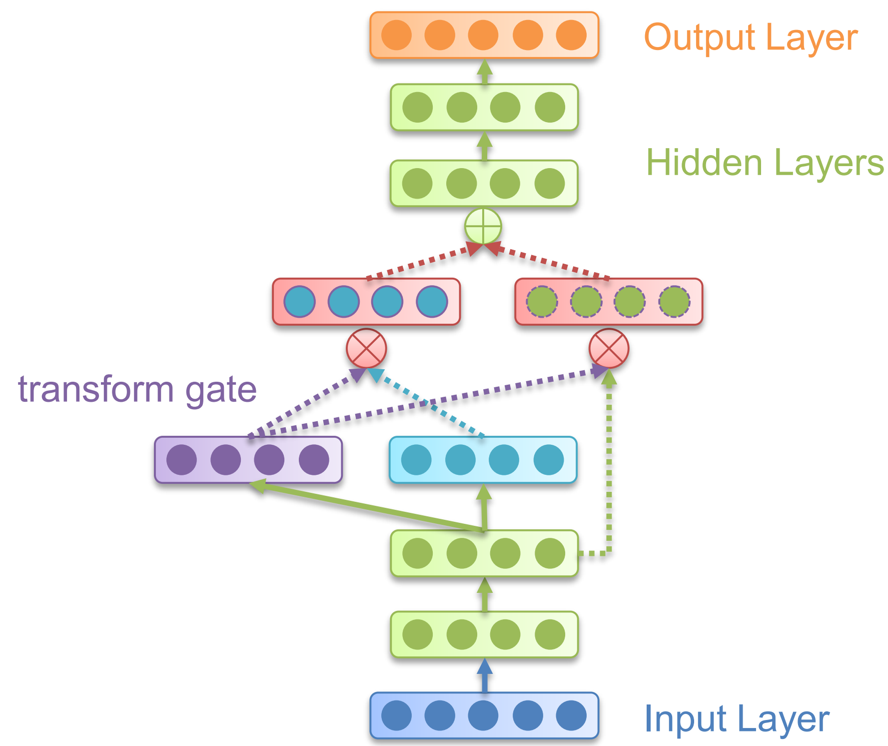
Highway Networks

$$x_{t+1} = H(x_t, W_H) \cdot T(x_t, W_T) + x_t \cdot C(x_t, W_C)$$



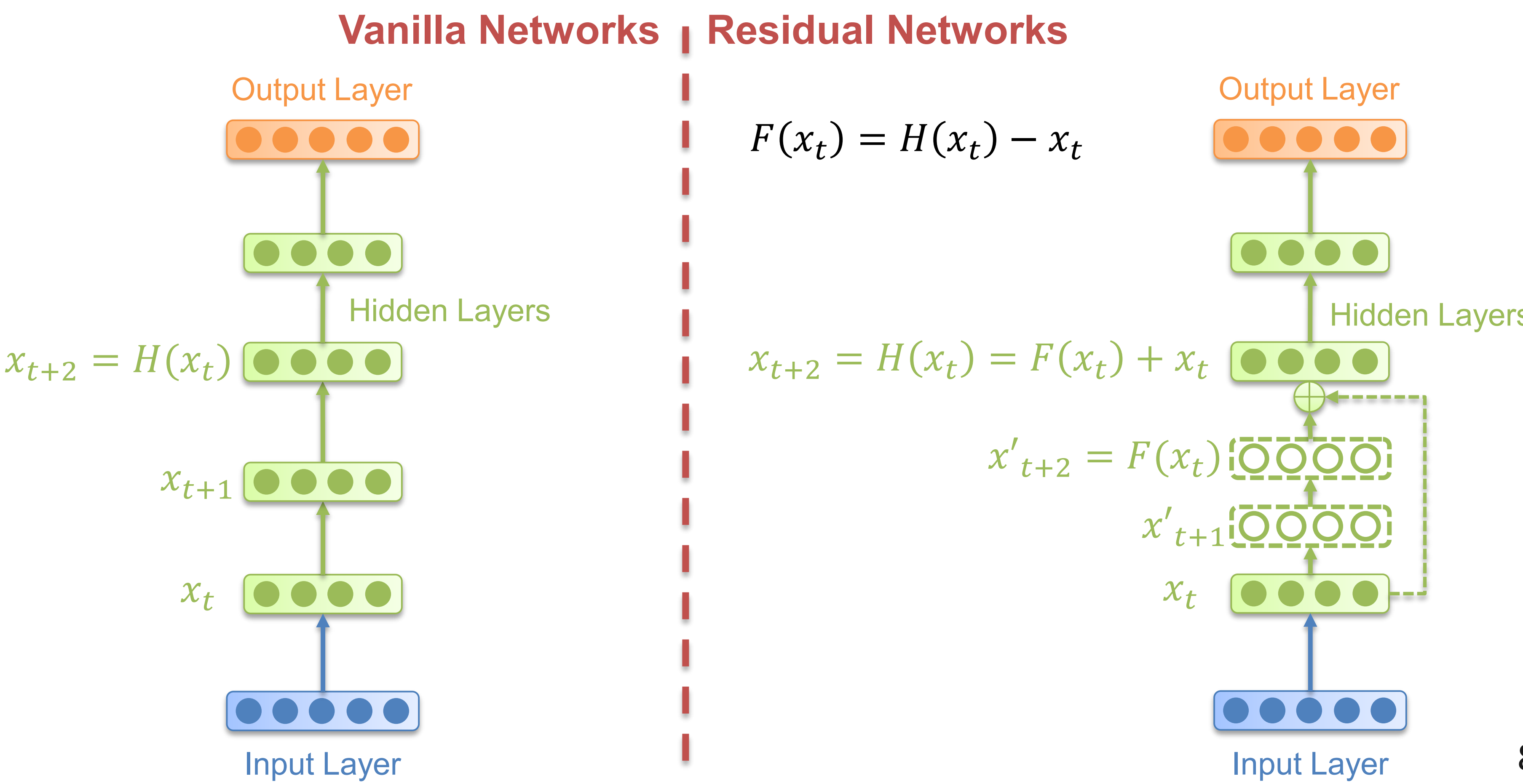
Simplified Highway Networks

$$x_{t+1} = H(x_t, W_H) \cdot T(x_t, W_T) + x_t \cdot [1 - T(x_t, W_T)]$$



Residual Networks

- ResNet hypothesizes that it is easier to optimize the **residual mapping** than to optimize the original, unreferenced mapping

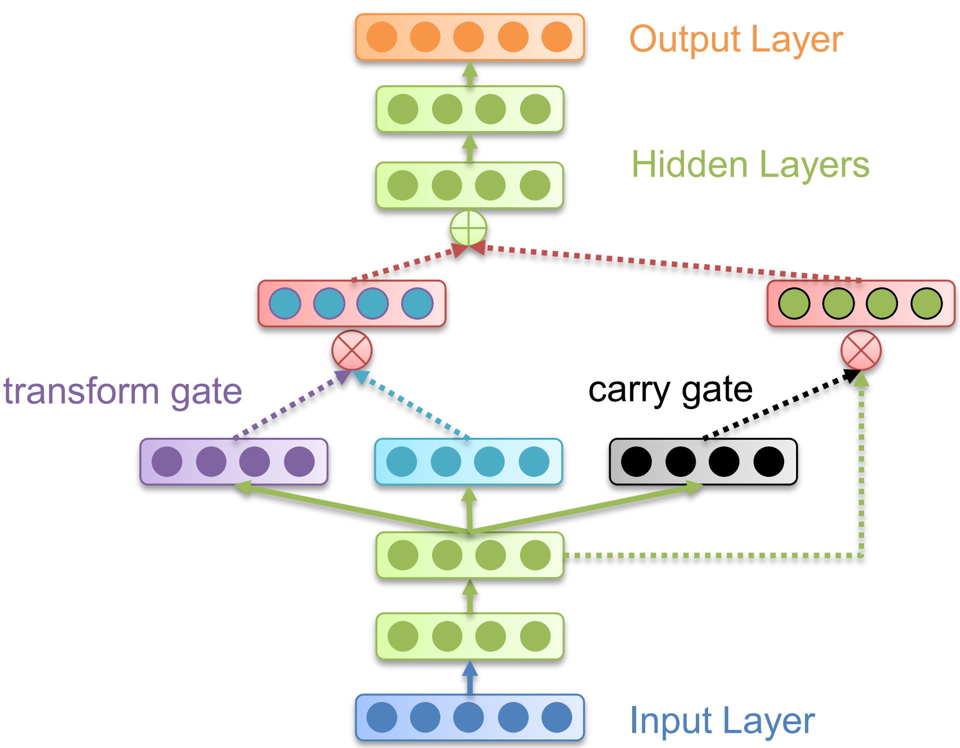


Highway vs. ResNet

- ResNet is a short name for Residual Network
  - ResNet usually refers to the classic CNN-based residual learning
- When $T(\cdot) = C(\cdot) = 1$, ResNet ~ Highway

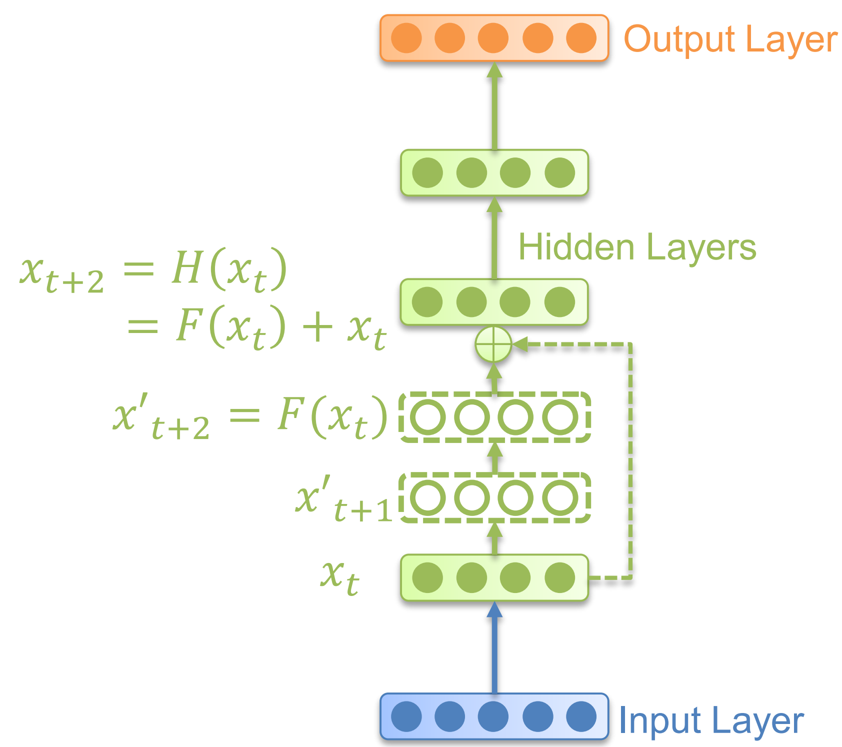
Highway Networks

$$x_{t+1} = H(x_t, W_H) \cdot T(x_t, W_T) + x_t \cdot C(x_t, W_C)$$

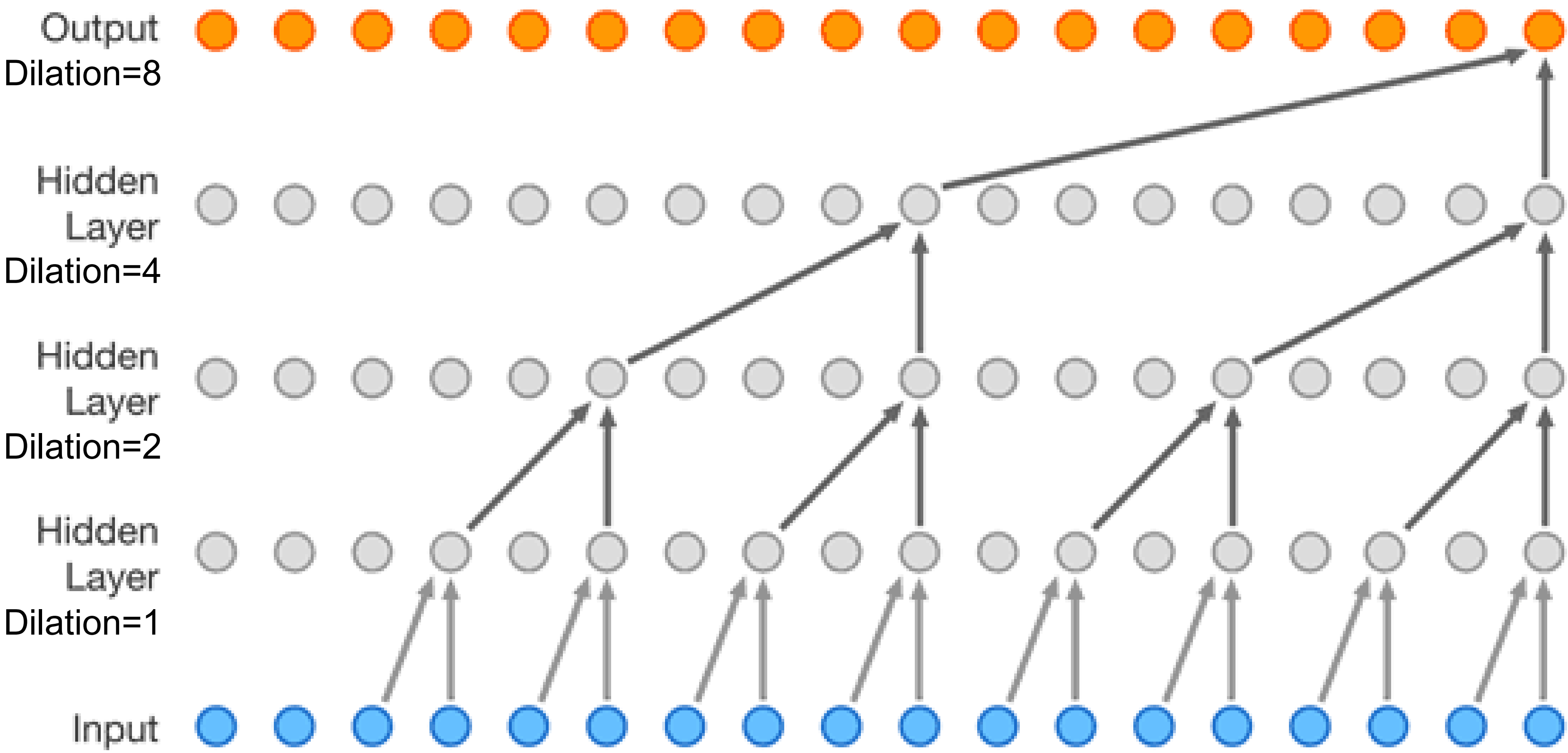


Residual Networks

$$x_{t+2} = F(x_t) + x_t \qquad F(x_t) = H(x_t) - x_t$$

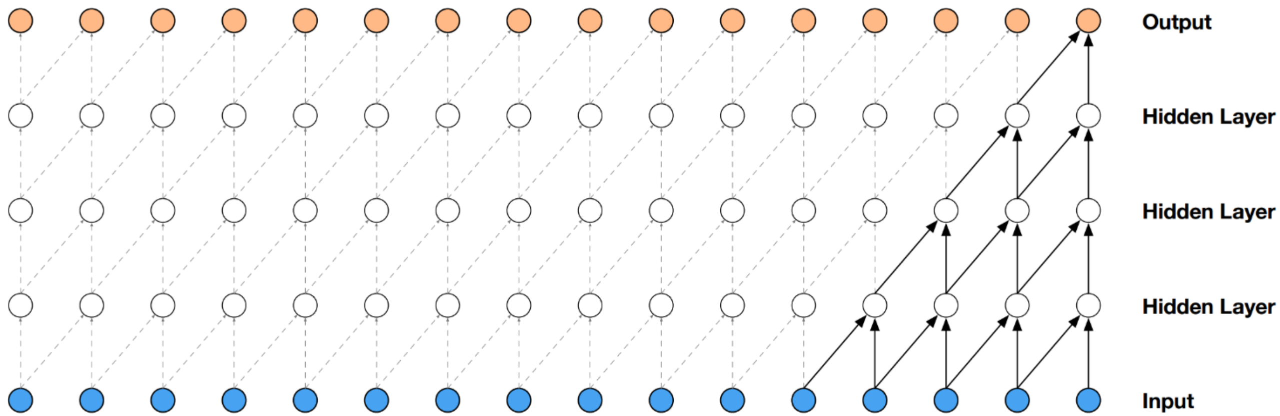


Dilated Convolutions.

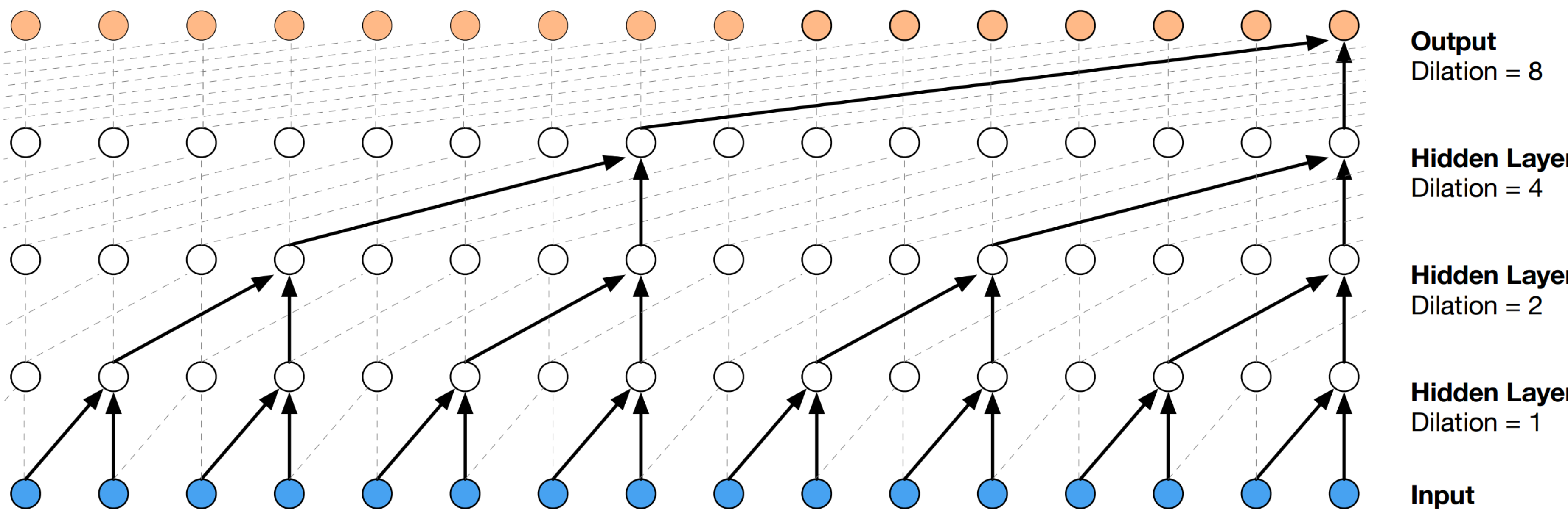


Dilated Convolutions..

- Classic CNNs

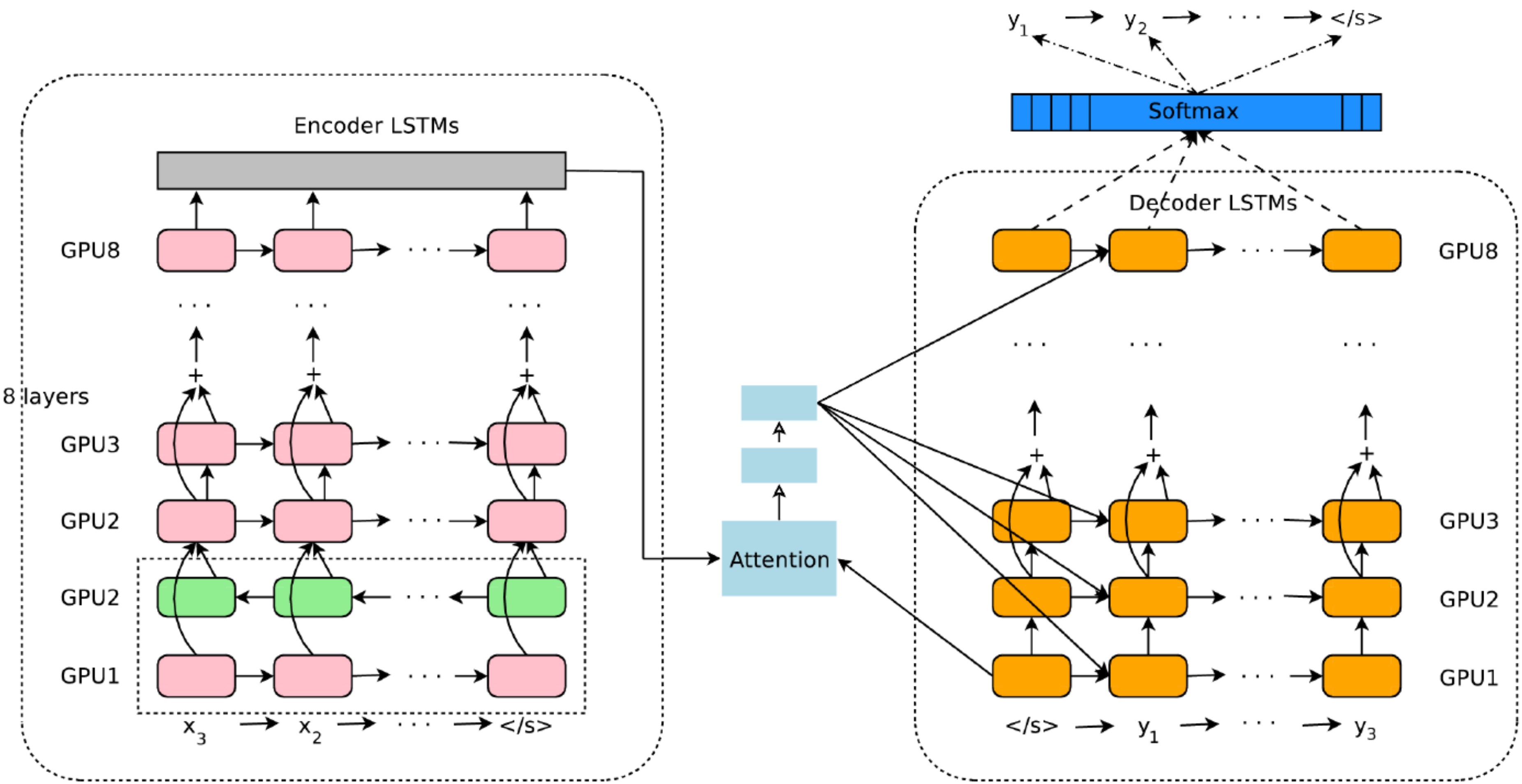


- Dilated Convolutions



Google’s Neural Machine Translation.

- A conventional encoder-decoder architecture with attention



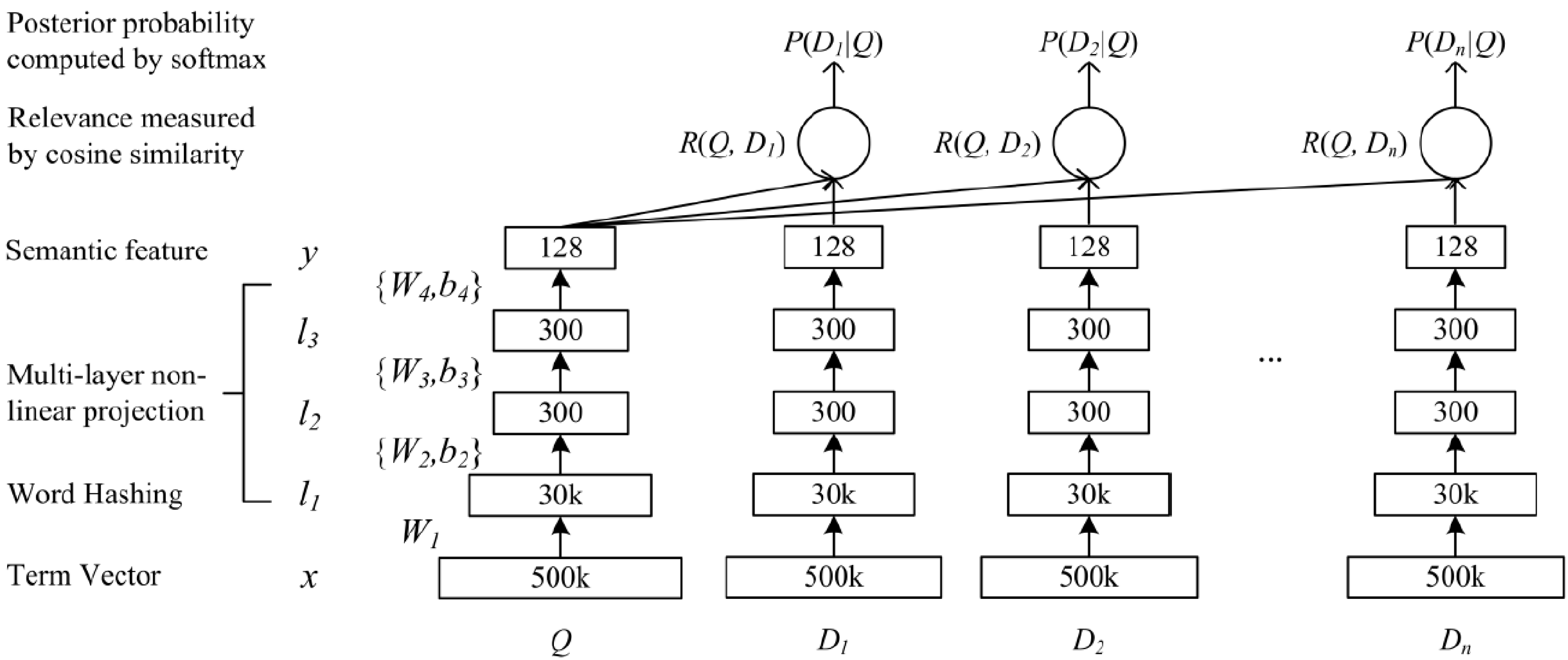
Google's Neural Machine Translation..

- To be able to make use of multilingual data within a single system, GNMT proposes one simple modification to the input data
  - An artificial token is introduced at the beginning of the input sentence to indicate the target language the model should translate to

```
How are you? -> ¿Cómo estás?
```

```
<2es> How are you? -> ¿Cómo estás?
```

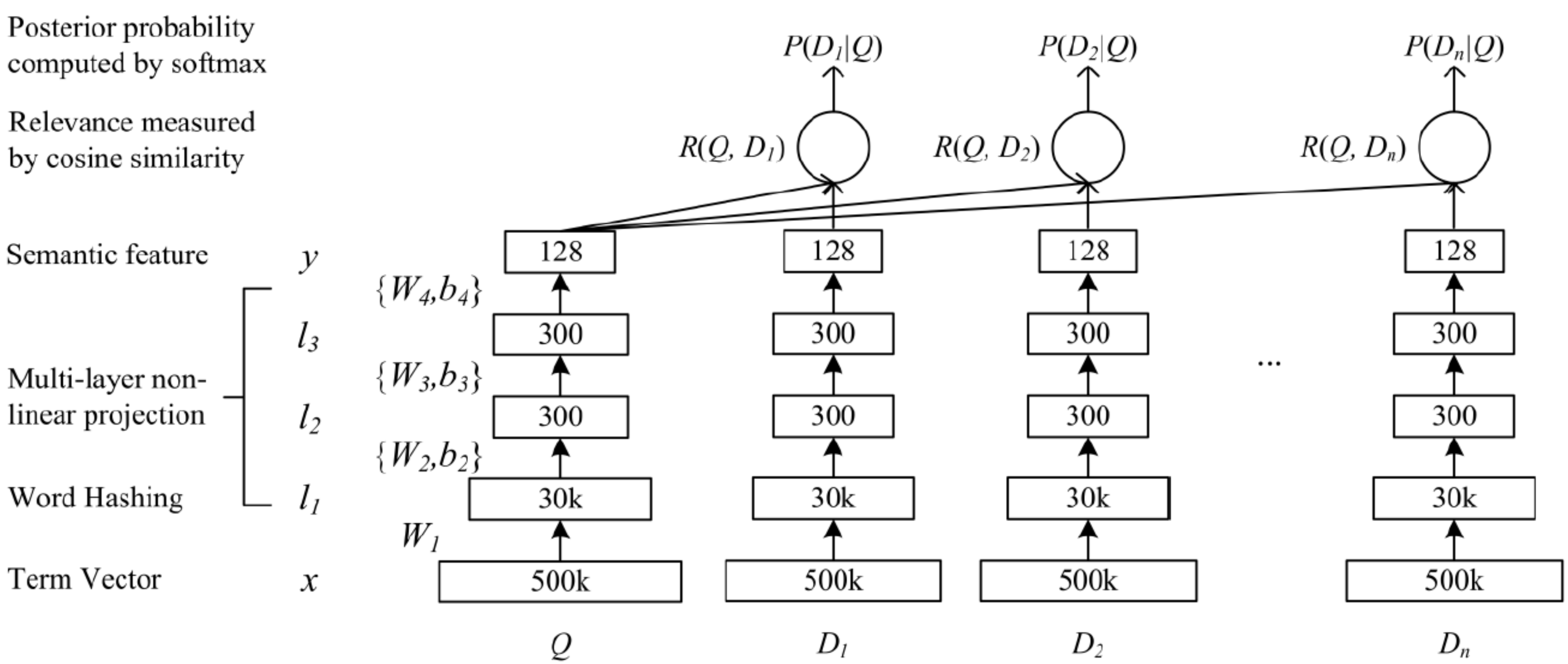
Deep Structured Semantic Model (DSSM).



	Letter-Bigram		Letter-Trigram	
Word Size	**Token Size**	**Collision**	**Token Size**	**Collision**
40k	1107	18	10306	2
500k	1607	1192	30621	22

#good# => [#go, goo, ood, od#]

DSSM..

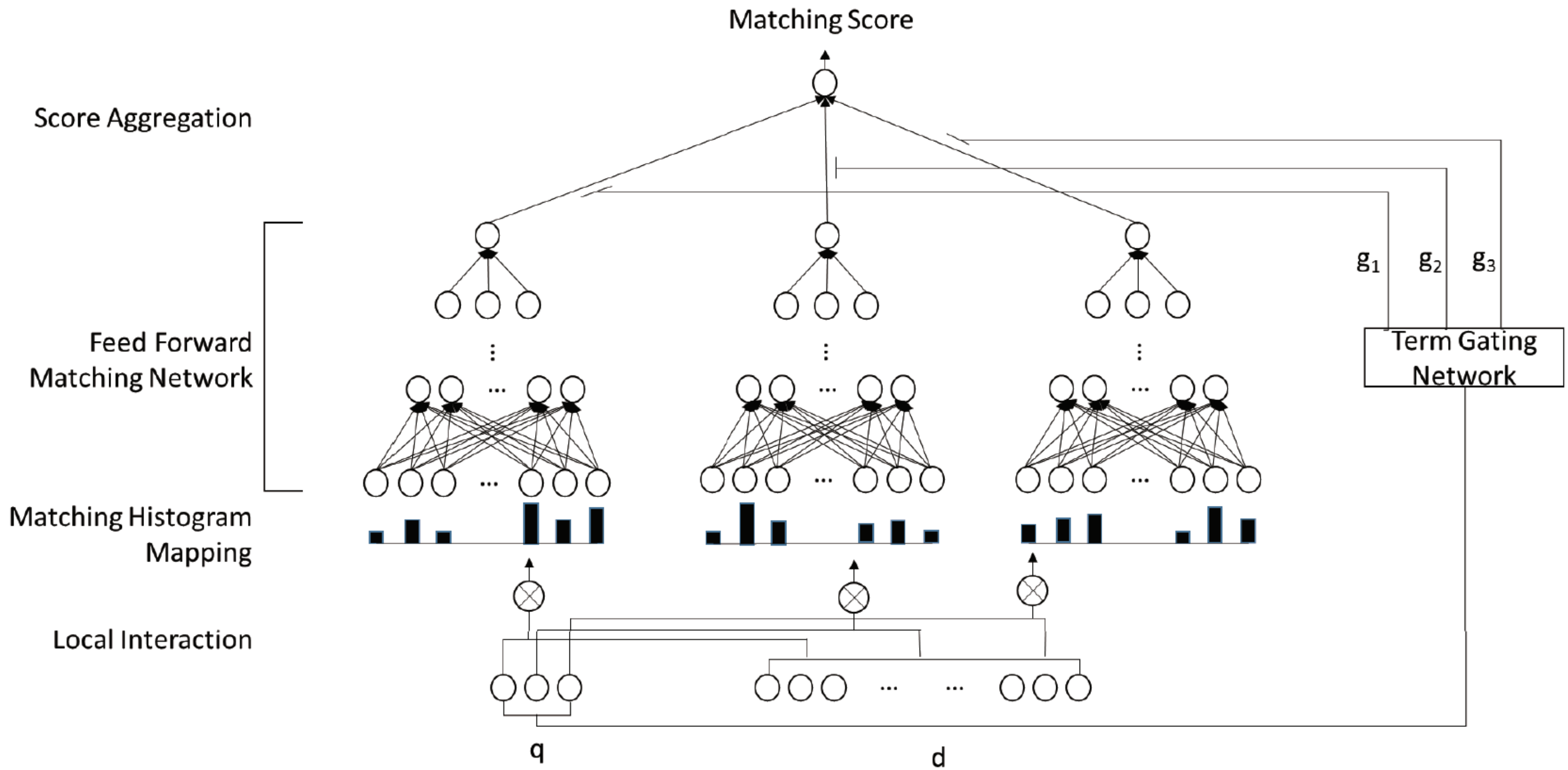


$$R(q,d) = \cos(\vec{q},\vec{d})$$

$$P(d|q) = \frac{\exp(R(q,d))}{\sum_{d'} \exp(R(q,d'))}$$

$$L = \prod_{d \in R_q} P(d|q)$$

Deep Relevance Matching Model

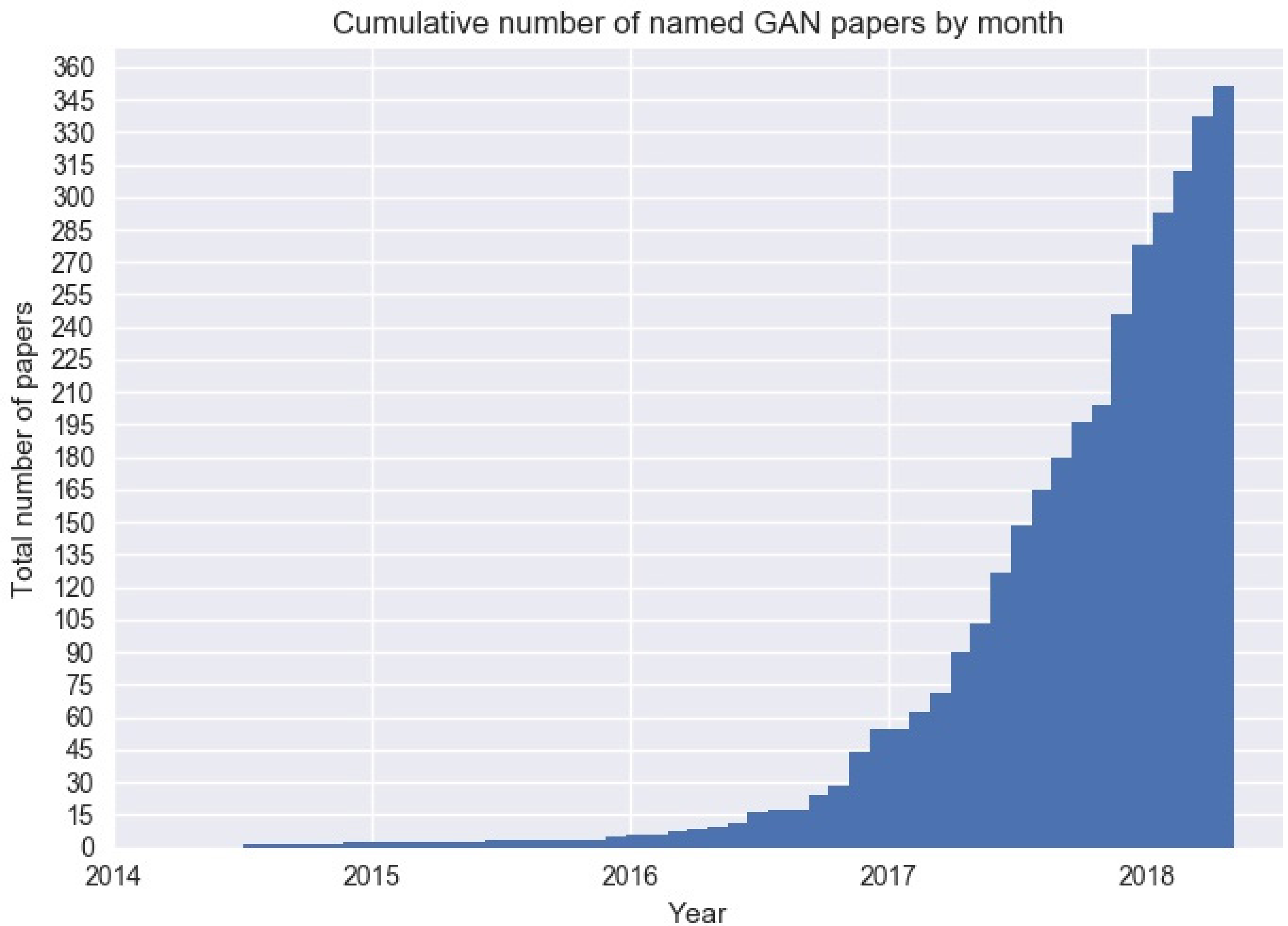


Query: "car to go"
Document: "car, rent, truck, bump, injunction, runway"
Five Bins: {[-1,-0.5), [-0.5,0), [0,0.5), [0.5,1), [1,1]}
Local Interaction for "car": (1, 0.2, 0.7, 0.3, -0.1, 0.1)
Matching Histogram for "car": [0, 1, 3, 1, 1]

Generative Adversarial Networks.

Generative Adversarial Networks..

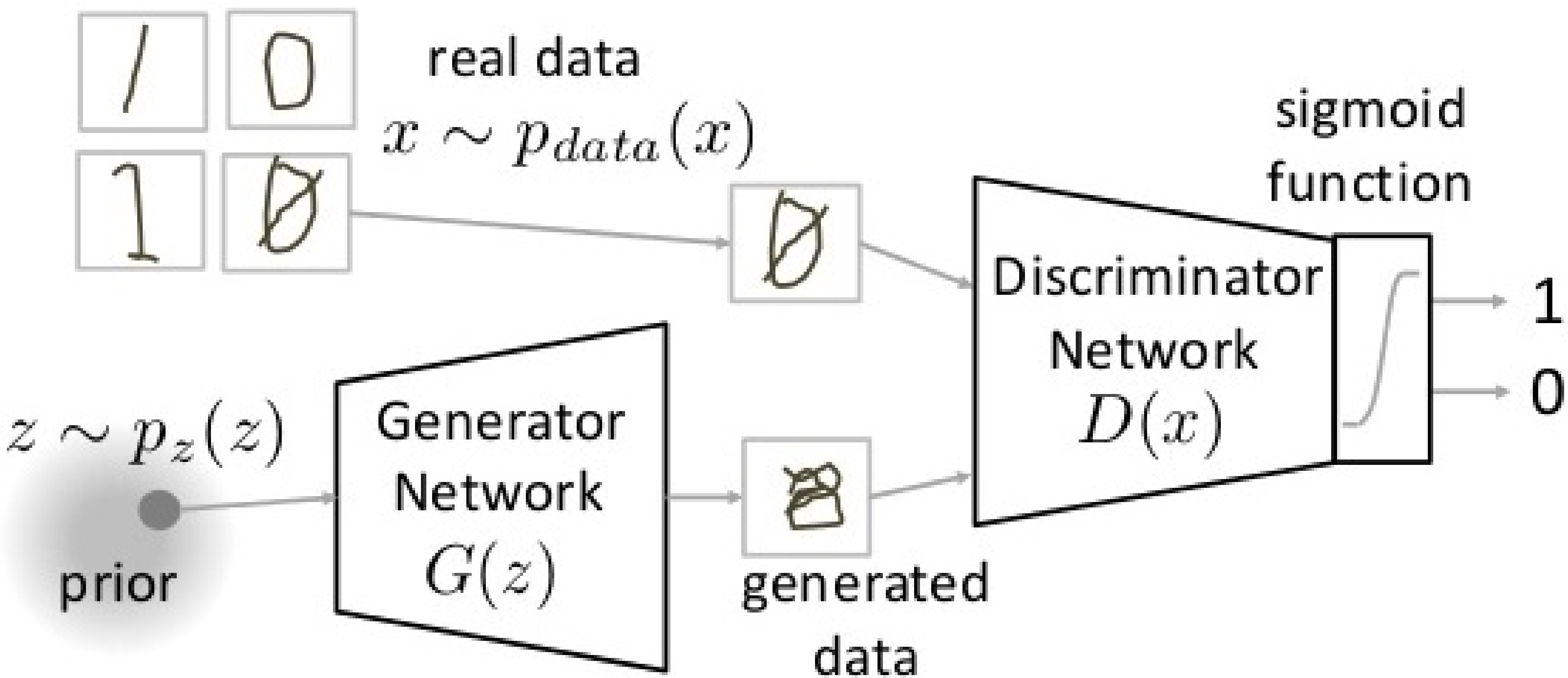
- https://github.com/hindupuravinash/the-gan-zoo



Generative Adversarial Networks…

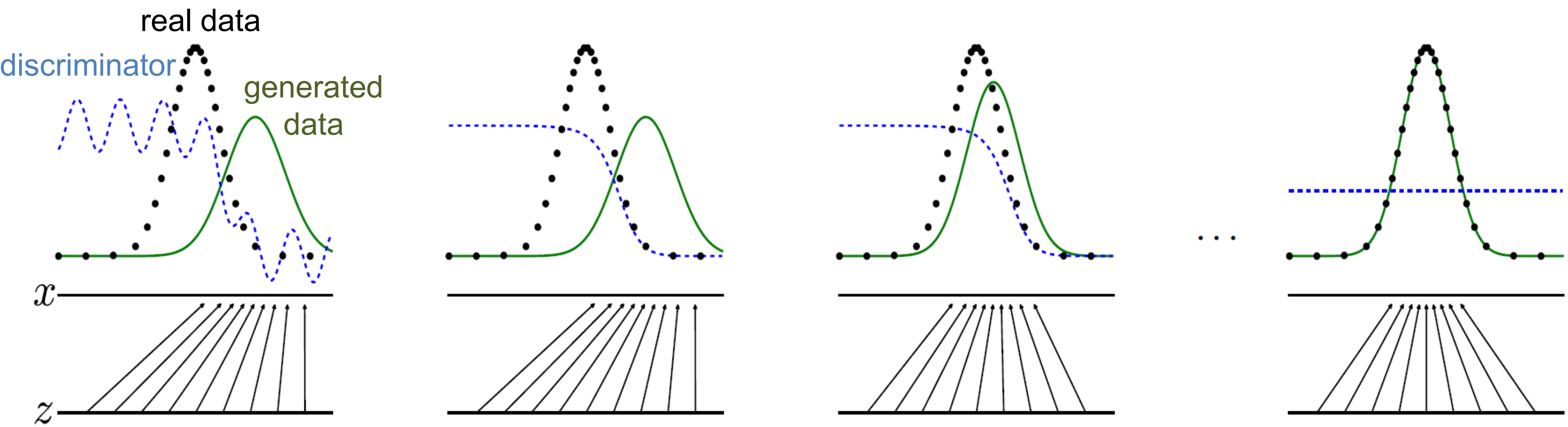
- Discriminator is used to criticize the results produced by generator
- The ultimate goal of generator is to cheat the discriminator, thus the generator can create potential objects

$$\min_{G}\max_{D} V(D,G) = \mathbf{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbf{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$



Generative Adversarial Networks….

$$\min_G \max_D V(D,G) = \mathbf{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbf{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$



Questions?

kychen@mail.ntust.edu.tw