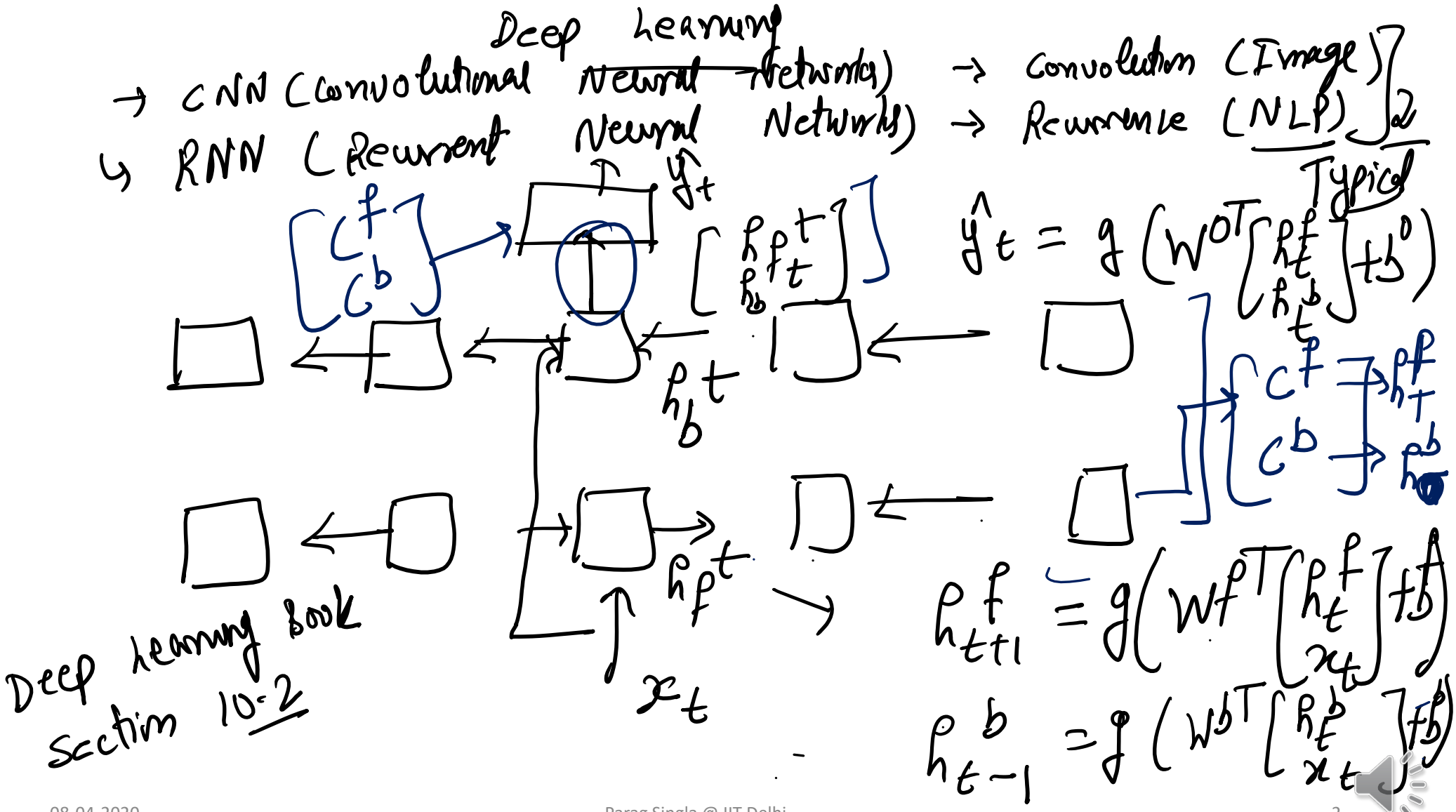
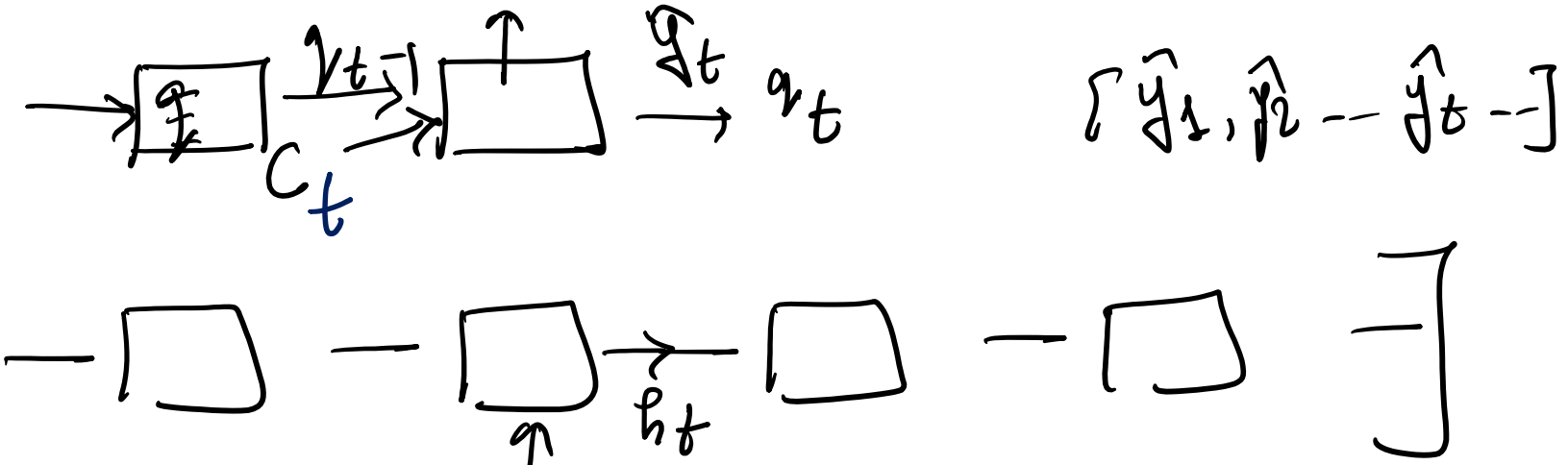
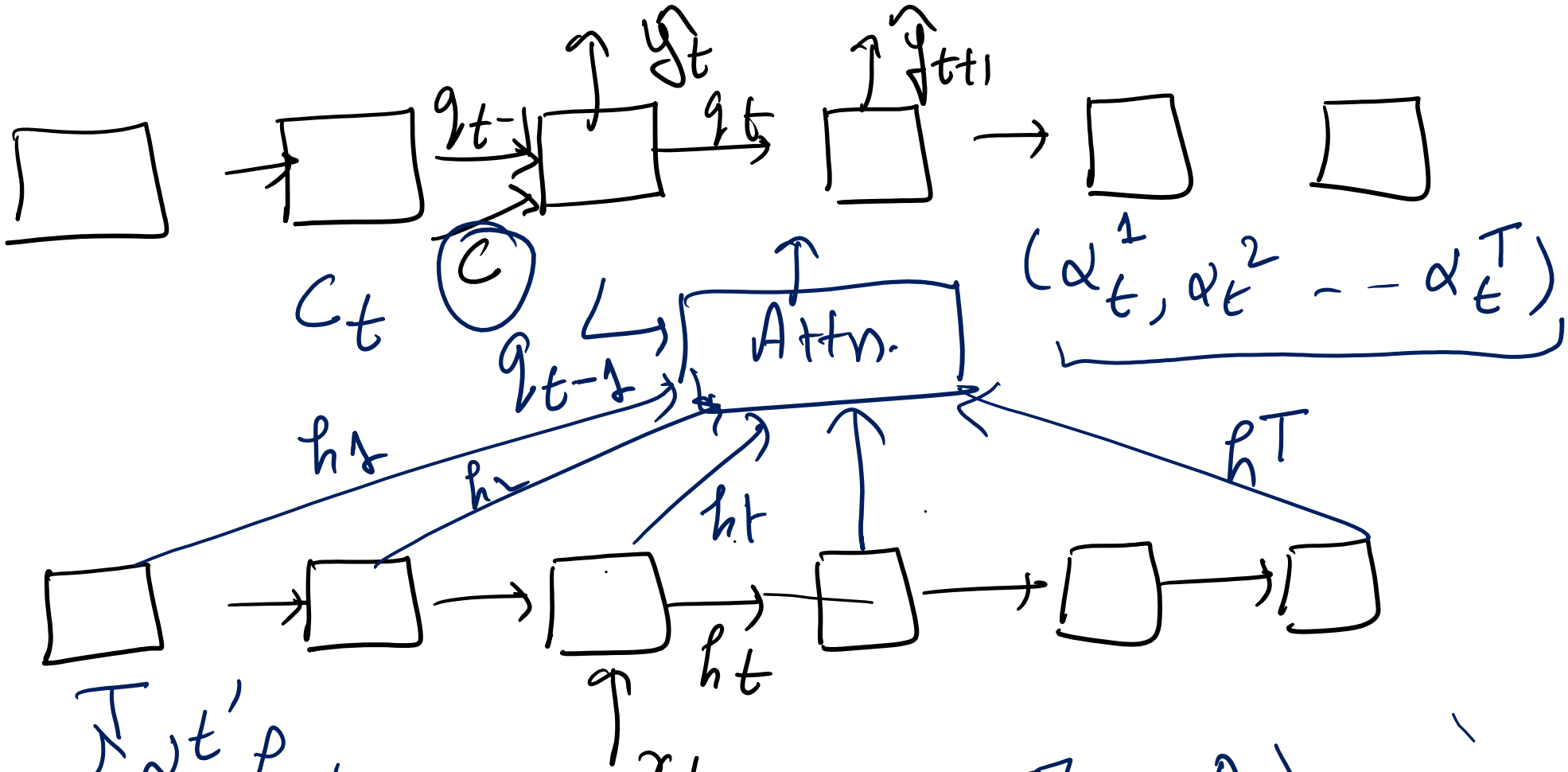

Deep learning
→ CNN (Convolutional Neural Networks) → Convolution (Image)
↳ RNN (Recurrent Neural Networks) → Recurrence (NLP)
Typical
Deep learning book
Section 10.2



Attention

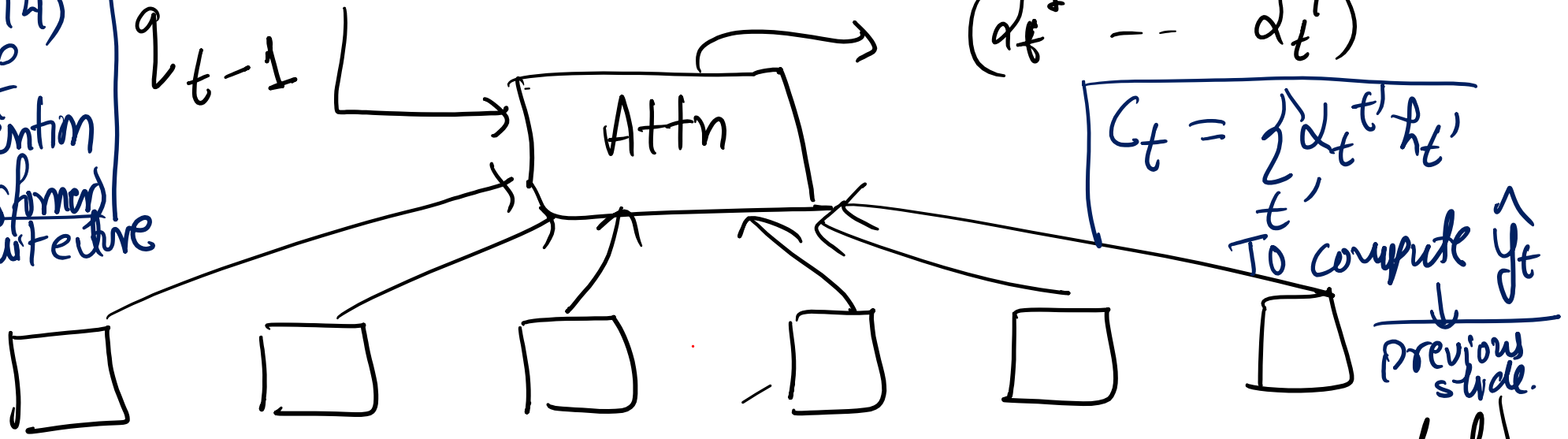
$$\hat{y}_t = g\left(W^o \begin{bmatrix} q_{t-1} \\ c \end{bmatrix} + b^o\right)$$

Bahdanau et al (2014)


y_t
y_{t+1}
q_{t-1}
q_t
c_t
c
Attn.
q_{t-1}
$(\alpha_t^1, \alpha_t^2 \cdots \alpha_t^T)$
h_1
h_2
h_t
h_T
h_t
x_t
$c_t = \sum_{t'=1}^{T} \alpha_t^{t'} h_{t'}$
$\hat{y}_t = g\left(W^0 \begin{bmatrix} q_{t-1} \\ c_t \end{bmatrix} + b^0\right)$

① Bahdanau et al. (2014)

② self Attention (Transformer) Architecture



$$a_t^{t'} = f\left(W^{att}\begin{bmatrix} q_{t-1} \\ h_{t'} \end{bmatrix} + b^{att}\right)$$ (unnormalized score)

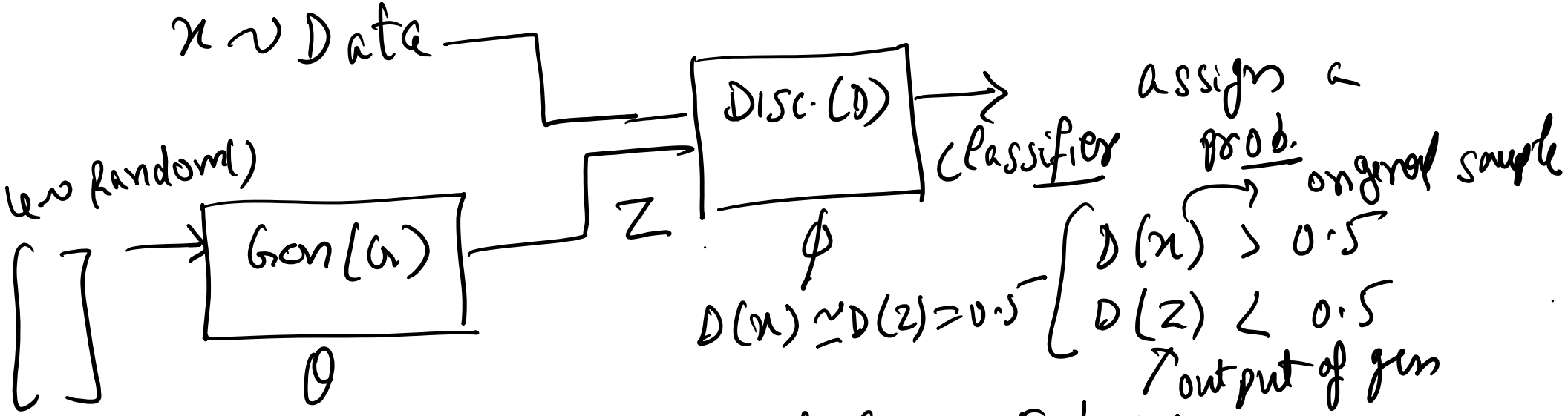
(softmax) $$\alpha_t^{t'} = \left[\frac{e^{a_t^{t'}}}{\sum_{t'} e^{a_t^{t'}}}\right]$$ (generalization of logistic)

H.W.

Relevant Paper links will be posted on the course website:

1. Attention (for Machine Translation)
   (Bahdanau et al. 2014)
2. Transformer Architecture/Self Attention
   (Vaswani et al. 2017)
3. Any others..

Generative Adversarial Networks (GANs)



$L(\theta, \phi)$

$L_{cross\text{-}entropy} = \sum_{y \in \{1 \ldots r\}} -1[y=c] \log p_c$

Cross Entropy

X_1, X_2 → same sample space

Cross Entropy (X_1, X_2) → Entropy

$= \sum_x - p(X_1 = x) \log p(X_2 = x)$

$$L(\theta, \phi) = -\left[E_{x \sim Data}[\log D_\phi(x)] + E_{z \sim Gen(u)}[\log(1 - D_\phi(z))]\right]$$

θ → parameters of Generator

ϕ → parameters of Discriminator

Disc:- $\min_{\phi} L(\theta, \phi)$

Gen:- $\max_{\theta} L(\theta, \phi)$

$\max_{\theta}\left[\min_{\phi} L(\theta, \phi)\right]$

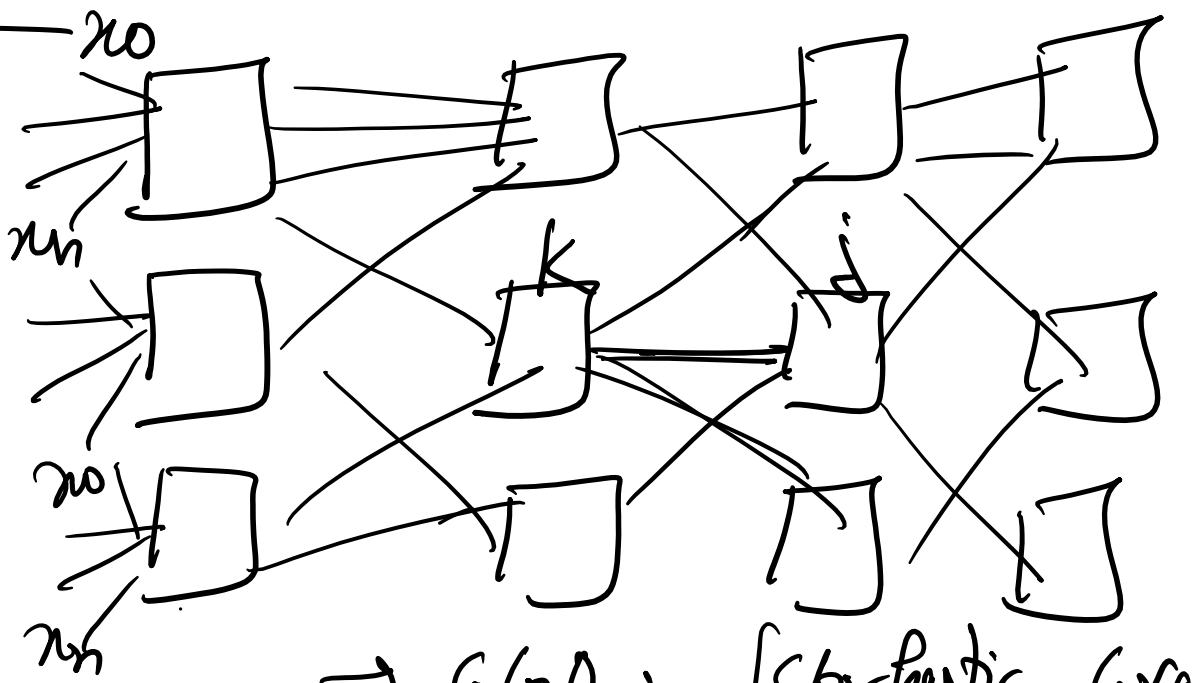
Alternate min/max problem

Goodfellow (2014) et al.

→ ① Mode collapse

→ ② Stability of training

Dropouts :-



Srivastava et al. (2014)

Batch Normalization

↓ Relevant

→ SGD :- [Stochastic Gradient Descent]

With p_{drop} :- $\theta_{jk} = 0$] $\forall jk$]

↳ (Robust)

At prediction :- (Ensemble)