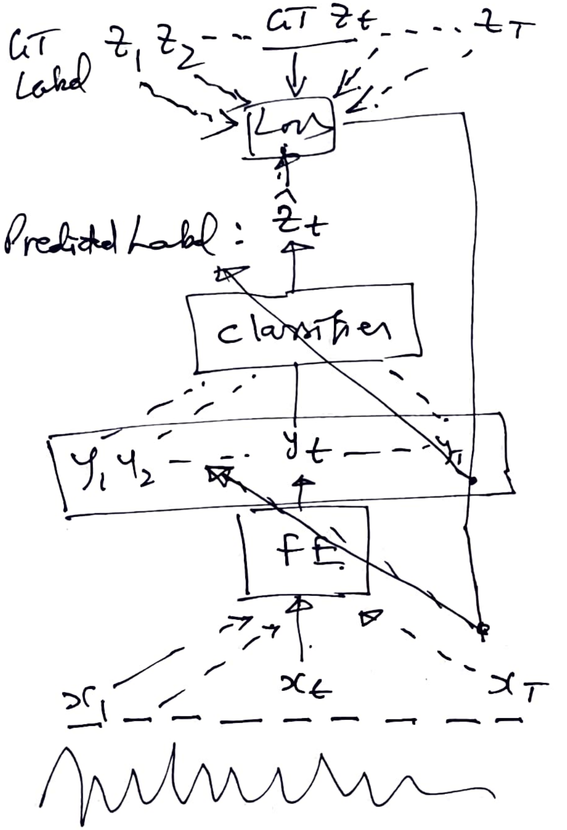
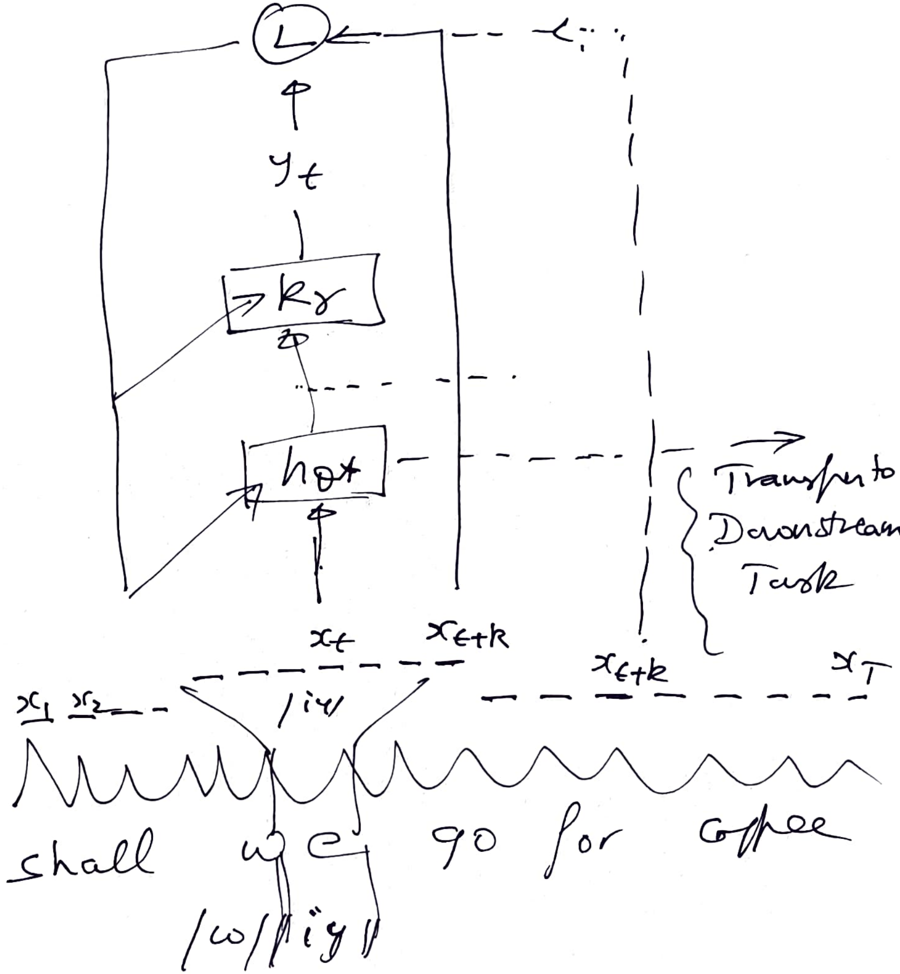
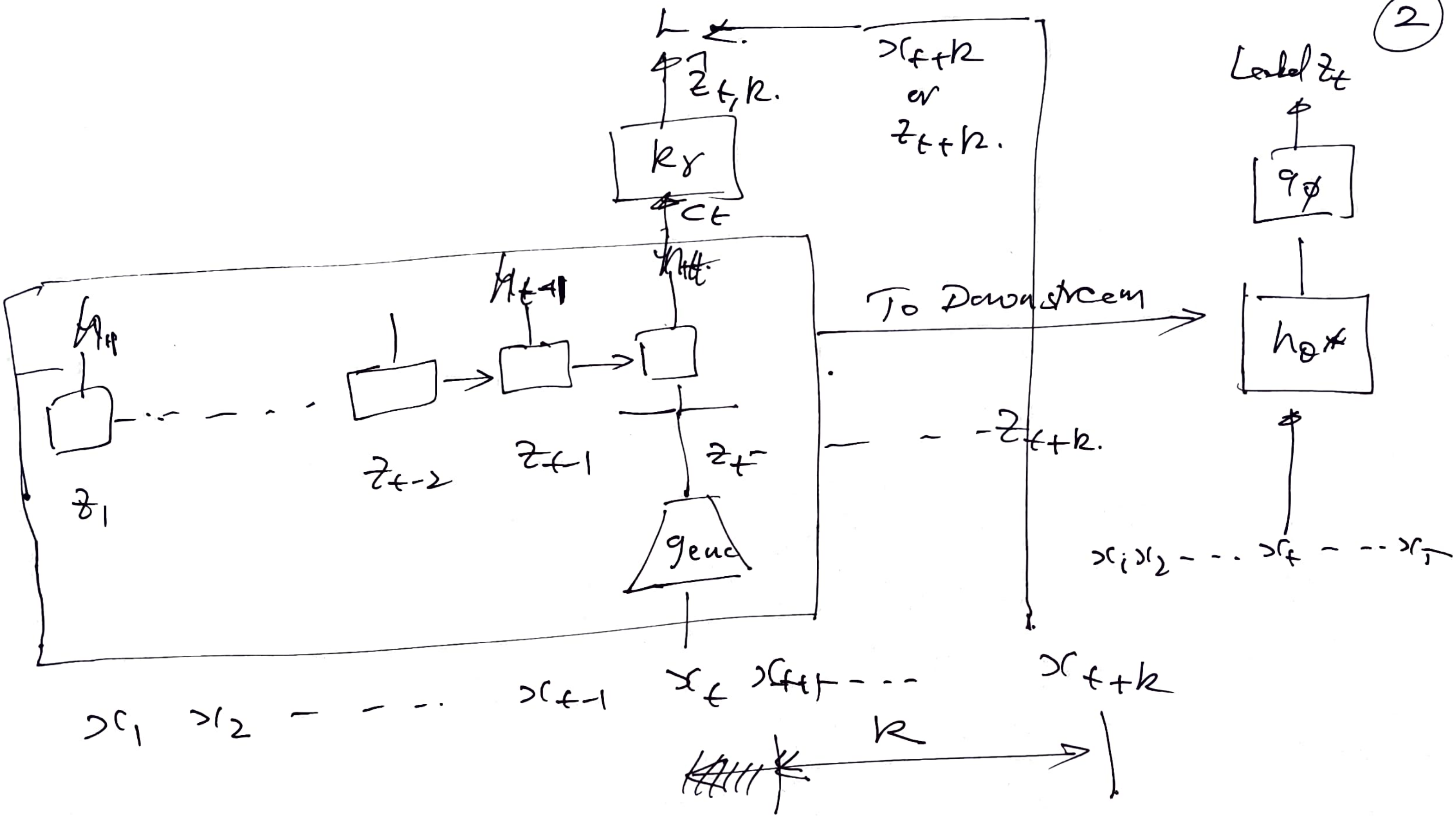
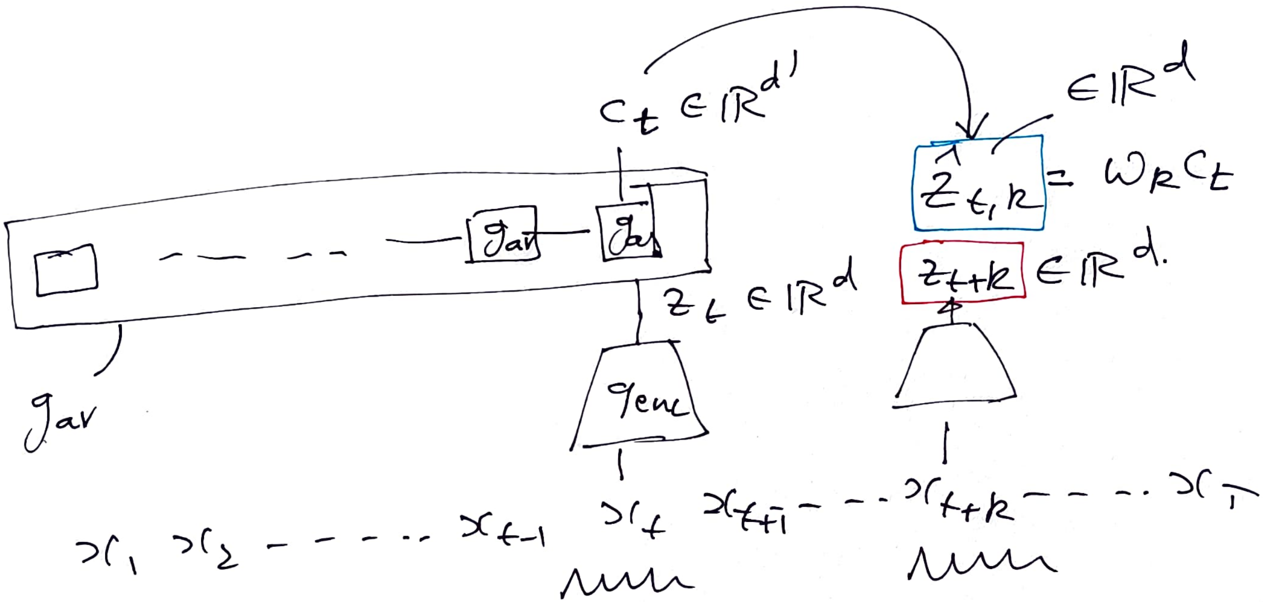


Conventional Supervised Pipeline for ASR

CPC




L
$z_{t,k}$
x_{t+k}
or
z_{t+k}
k_γ
c_t
h_t
h_{t-1}
h_1
To Downstream
z_{t-2}
z_{t-1}
z_t
z_1
z_{t+k}
genc
Leaked z_t
q_ϕ
h_θ
$x_1 x_2 \dots x_t \dots x_T$
x_1 x_2 x_{t-1} x_t x_{t+1} x_{t+k}
k



$$z_t = g_{enc}(x_t) \longrightarrow \text{10ms latent variable}$$

$$c_t = g_{ar}(z_{\leq t})$$

$c_t \xrightarrow{W_k} \hat{z}_{t,k} \rightarrow$ Similarity measure f_k

$z_1\ z_2\ - \quad - \cdot\ z_t \qquad z_{t+k}. \nearrow$

$x_1, x_2 - \quad - \quad -.\ x_t\ \ x_{t+1} - - \ \ x_{t+k} - - - \ x_T$

- Ideally, if we have a generative model $p(x_{t+k} \mid c_t)$ it is easy to predict future observations (as in a LM)
  - probabilistic "goodness" of $x_{t+k} \mid c_t$
- But, use a metric that preserves Mutual Information (MI) between x_{t+k} & c_t

as $$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k} \mid c_t)}{p(x_{t+k})}$$

where $$f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right)$$

i.e.

$$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k} \mid c_t)}{p(x_{t+k})}$$

realized as a scalar "similarity" score between x_{t+k} [i.e. z_{t+k}] & c_t [i.e. $W_k c_t = \hat{z}_{t,k}$]

$$f_k(x_{t+k}, c_t) = \exp\Big(\underbrace{z_{t+k}^{T}\, \overbrace{W_k\, c_t}^{\hat{z}_{t,k}}}_{\text{dot product}}\Big)$$

$$\begin{tikzcd}
c_t \arrow[r, bend left, "W_k"] & \hat{z}_{t,k} \arrow[r] & \boxed{f_k(x_{t+k}, c_t)} \\
z_t \arrow[u] & z_{t+k} \arrow[u] \arrow[ur] & \\
x_t \arrow[u] & x_{t+k} \arrow[u] &
\end{tikzcd}$$

Mutual Information [Information Gain]

- Between 2 random variables X & Y
- Measure of the Mutual Dependence between the 2 variables
- Quantifies "amount of information" obtained about one r.v by observing the other r.v.
- Uncertainty about X (or Y) is reduced once Y (or X) is observed
- → Information gain.

MI

$$I(X;Y) = D_{KL}\left(P_{(X,Y)} \,\|\, P_X \otimes P_Y\right)$$

$\Rightarrow$ for PMFs

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} P_{X,Y}(x,y) \log\left[\frac{P_{X,Y}(x,y)}{P_X(x) \cdot P_Y(y)}\right]$$

- $I(X,Y) = 0$ if and only if X & Y are independent variables

ie $\underline{P_{X,Y}(x,y) = P_X(x) \cdot P_Y(y)}$

In our setting

$$X \Rightarrow x \;;\quad Y \Rightarrow c$$

x → signal x_t, x_{t+k} etc

c → Context c_t vector

$$I(x;c) = \sum_{x,c} p(x,c) \log \frac{p(x,c)}{p(x)\cdot p(c)}$$

$$= \sum_{x,c} p(x,c) \log \frac{p(x/c)}{p(x)}$$

$p(x,c) = p(x/c)\cdot p(c)$

Maximise MI between encoded representations $\underline{c_t \;\&\; z_{t+k}}$ → in the place of x_{t+k}

~~in place of~~

of interest

$$I(x_{t+k}, c_t) = \sum_{x, c} p(x_{t+k}, c_t) \log \frac{p(x_{t+k} | c_t)}{p(x_{t+k})}$$

If Loss between c_t & x_{t+k} (or z_{t+k}) is of the form

$$\mathcal{L}_N = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

$$\boxed{X : N \text{ samples} \begin{cases} 1 \text{ +ve } x_{t+k} \\ N-1 \text{ -ve } x_j \end{cases}}$$

Then it can be shown

$$I(x_{t+k}, c_t) \geq \log(N) - \mathcal{L}_N$$

Ideally, we need to optimise g_{enc}, g_{ar} & W_k to maximize <u>mutual information</u> between c_t & x_{t+k}

- Instead, work with MI between c_t & z_{t+k}
- As a proxy to <u>max MI</u> $\Rightarrow$ <u>min InfoNCE loss $\mathcal{L}_N$</u>

c_t → <u>MI or $\mathcal{L}_N$</u>

<u>MI</u>

$$
\begin{array}{ccccccccc}
 & & & c_t & & & & & \\
 & & & \uparrow & & & & & \\
 & & & z_t & & & z_{t+k} & & \\
 & & \uparrow & \uparrow & & & & & \\
x_1, x_2 & \cdots & x_{t-1} & x_t & x_{t+1} & \cdots & x_{t+k} & \cdots & x_T
\end{array}
$$

- $I(x_{t+k}, c_t) \geq \log N - \mathcal{L}_N \longrightarrow$ <u>InfoNCE loss</u>

$\mathcal{L}_N \downarrow \longrightarrow$ Lower Bound of $I(x_{t+k}, c_t) \uparrow$

$\longrightarrow$ Good for MI(x_{t+k}, c_t)

$N \uparrow \longrightarrow$ Good for "

Implication of Bound-Result

$$\text{Max } I(x_{t+k}, c_t)$$

$$\Longrightarrow \min_{g_{enc}, g_{ar}, W_k} \mathcal{L}_N$$

$$= \min_{g_{enc}, g_{ar}, W_k} \left[-\mathbb{E}_X \left(\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right) \right]$$

$$= \min_{g_{enc}, g_{ar}, W_k} -\mathbb{E}_X \left[\log \frac{\exp(z_{t+k}^T W_k c_t)}{\sum_{x_j \in X} \exp(z_j^T W_k c_t)} \right]$$

Self-Supervision by CPC : Find optimal Model Parameters

$$\theta^* = g_{enc}^*,\ g_{ar}^*,\ W_k^*.$$

$$= \underset{g_{enc},\, g_{ar},\, W_k}{\operatorname{argmin}} - E_x \left[\log \frac{\exp(z_{t+k}^T W_k c_t)}{\sum_{x_j \in X} \exp(z_j^T W_k c_t)} \right]$$

$$\Rightarrow \quad \begin{aligned} z_{t+k} &= g_{enc}(x_{t+k}) \\ c_t &= g_{ar}(z_{\leq t}) \\ z_j &= g_{enc}(x_j) \end{aligned}$$

$\mathcal{L}_N$ is a NCE loss — towards maximizing MI.

$\mathcal{L}_N$: Given a set of $X = \{x_1, x_2, \ldots x_N\}$
of N random samples
— has 1 +ve sample from $p(x_{t+k}/c_t)$
— N−1 −ve samples from $p(x_{t+k})$

Data Distn
$p(x_{t+k}/c_t)$: ~~Each~~ Conditional pdf of x_{t+k} given c_t

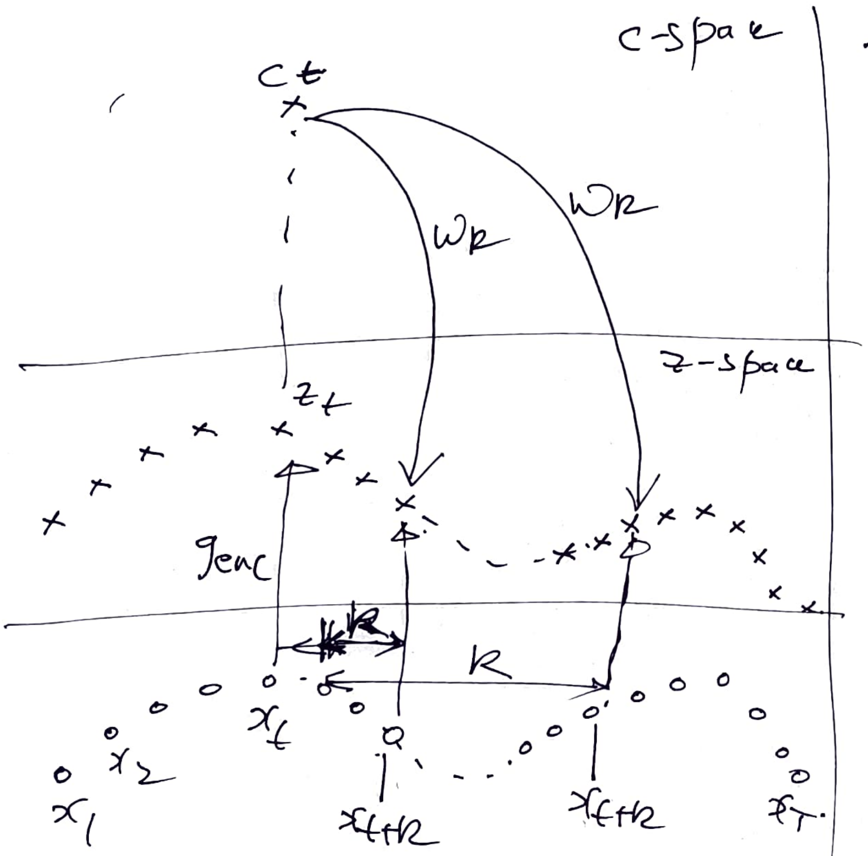
MI Concept $\rightarrow$ $\left\{ \begin{array}{l} \text{Once } c_t \text{ is observed available,} \\ \text{uncertainty about } x_{t+k} \text{ reduces} \end{array} \right.$

Noise Distn
$p(x_{t+k}) \rightsquigarrow$ pdf when no knowledge of c_t is available.

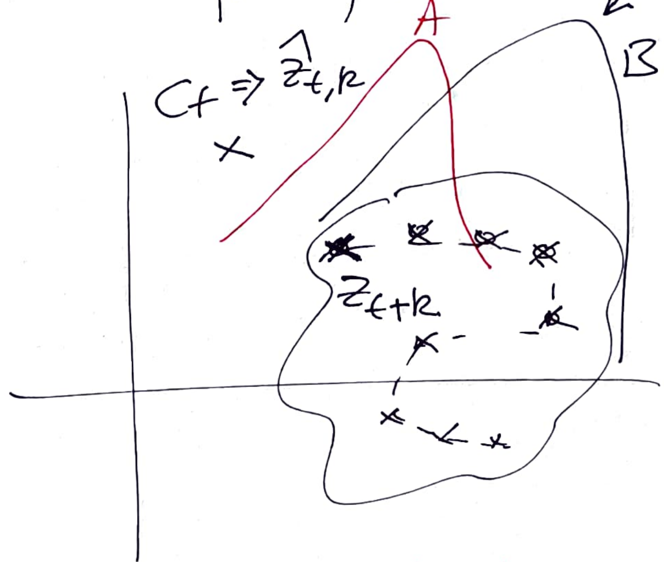
$c_t \to z_{t+k}$, $c_t \to x_{t+k}$

z_1 - - - z_{t-1} z_t z_{t+k}

x_1, x_2 - - - x_{t-1} x_t x_{t+1} - - - x_{t+k} - - - x_T



Typically x_{t+k} is drawn from $p(x_{t+k})$



A: $p(x_{t+k} \mid c_t)$

B: $p(x_{t+k})$

$x_t \longrightarrow z_t \longrightarrow c_t$

On knowing c_t, pick x_{t+k} for some 'k'

then given | on observing c_t

$p(x_{t+k}) \longrightarrow p(x_{t+k} / c_t)$

Noise distribution | Data Distribution

[−ve samples are drawn from this distr] | [+ve sample x_{t+k} is supposed to come from this distribution]

Verify this $\longrightarrow$ NCE

InfoNCE loss $\mathcal{L}_N = - \underset{X}{E} \left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right]$

Is a categorical cross-entropy loss

- of classifying the +ve sample 'correctly'

i.e. as coming from Data Distri. & Not the Noise Distn

P [x_{t+k} was drawn from the conditional distn $p(x_{t+k}/c_t)$ rather than the proposal or 'noise' distn $p(x_{t+k})$]

$= p(x_{t+k}$ is a +ve sample $| X, c_t)$

$$P\left(x_{t+k} \text{ is } \overset{\text{a +ve sample}}{\text{from Data Distn}} \,\middle|\, X, c_t\right) = \frac{p(x_{t+k}, c_t)}{p(c_t)}$$

$$= \frac{\dfrac{p(x_{t+k} \mid c_t)}{p(x_{t+k})}}{\displaystyle\sum_{j=1}^{N} \frac{p(x_j \mid c_t)}{p(x_j)}}$$

jump

Eqn (5)

$$P(x_{t+k}^{+ve} \mid c_t) \cdot P(c_t) = P(x_{t+k}, c_t) \mid X$$

This is the Prob associated with correct class @ optimal (minimization of InfoNCE Loss $\mathcal{L}_N$)