Boulder

Speech & Music; Modeling

[YouTube Playlist]

Maziar Raissi

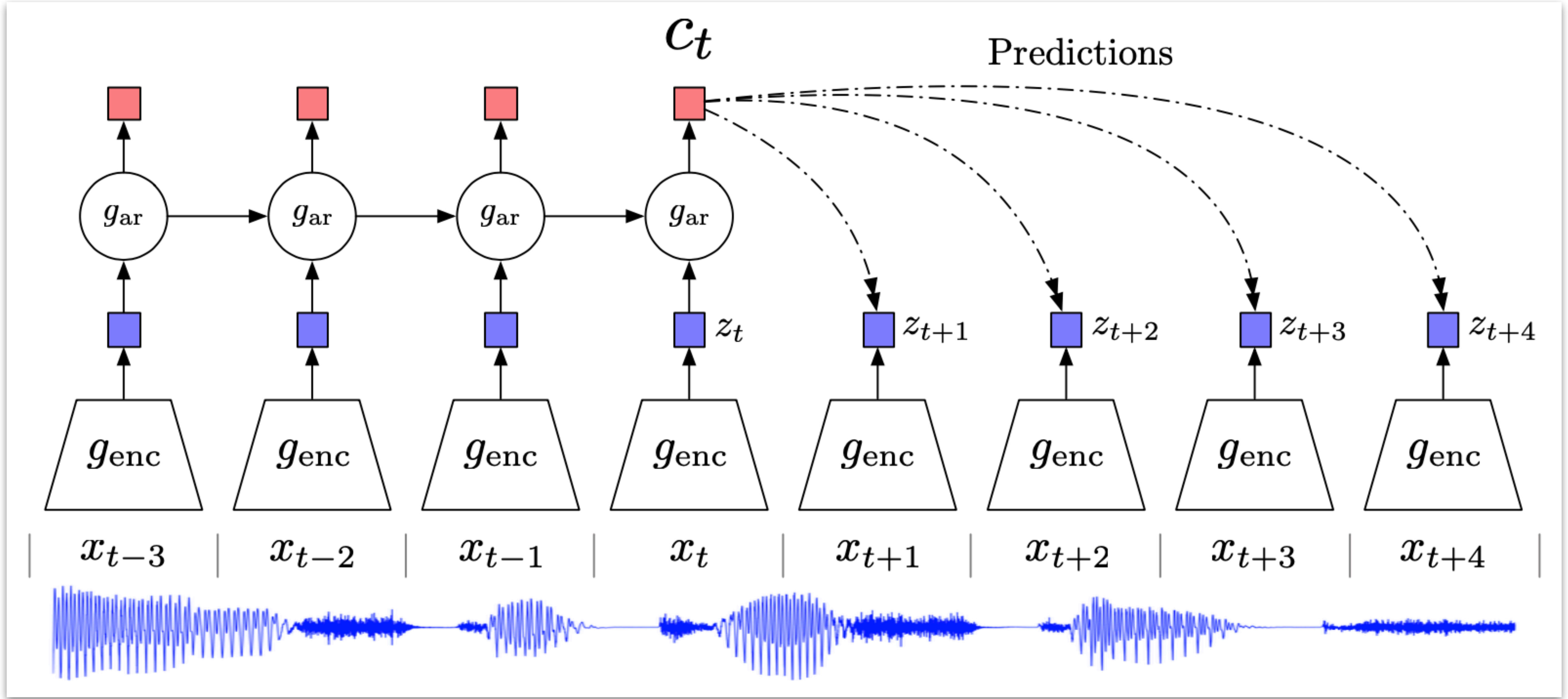
Assistant Professor

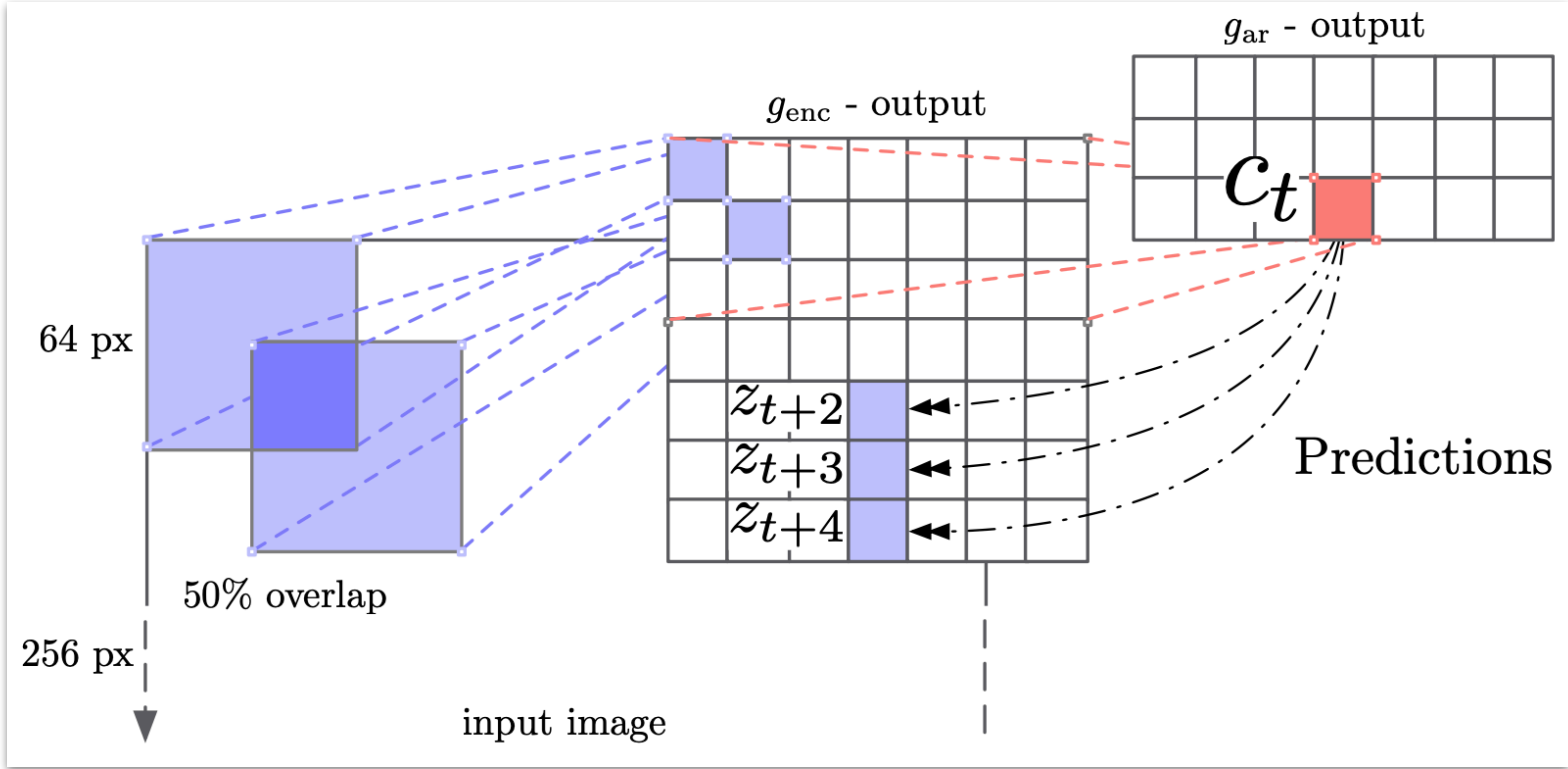
Department of Applied Mathematics

University of Colorado Boulder

maziar.raissi@colorado.edu

Representation Learning with Contrastive Predictive Coding





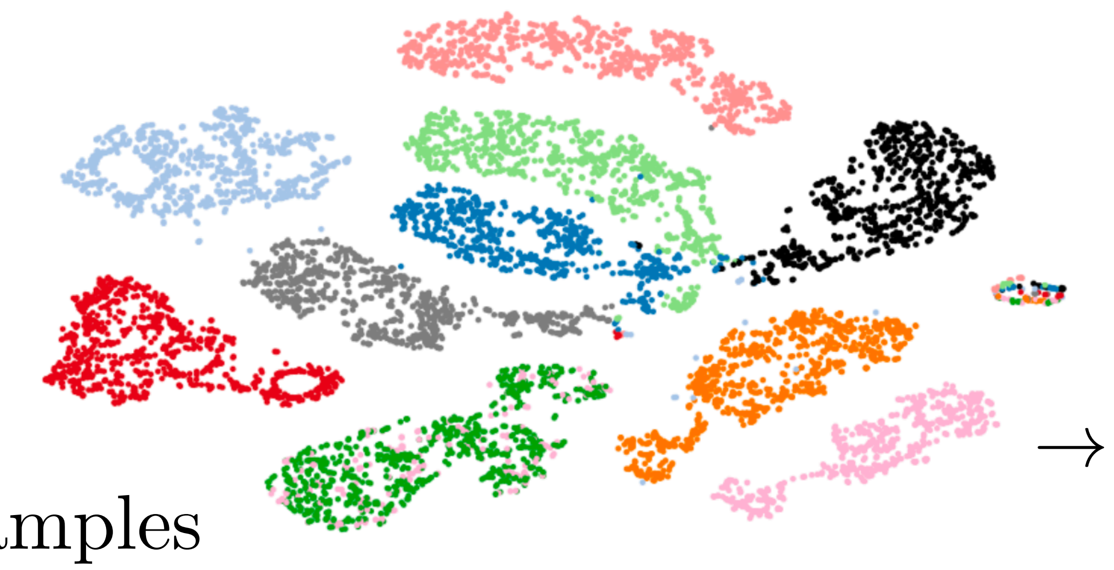
$g_{enc} \rightarrow$ non-linear encoder (e.g., strided convolutional layers with resnet blocks)
$x_t \rightarrow$ input observation
$z_t = g_{enc}(x_t) \rightarrow$ latent representation
$g_{ar} \rightarrow$ autoregressive model (e.g., GRUs)
$c_t = g_{ar}(z_{\leq t}) \rightarrow$ context latent representation (summarizing all $z_{\leq t}$ in the latent space)

$$f_k(x_{t+k}, c_t) := \exp(z_{t+k}^T \underbrace{W_k c_t}_{\hat{z}_{t+k}})$$

InfoNCE Loss

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

$X = \{x_1, x_2, \ldots, x_N\} \rightarrow$ set of N random samples



$\rightarrow$ t-SNE visualization of speech (each color represents a different speaker)

Mutual Information Estimation

The optimal value for $f(x_{t+k}, c_t)$ is proportional to $\frac{p(x_{t+k}|c_t)}{p(x_{t+k})}$!

No need to predict future observations x_{t+k} directly with a generative model $p(x_{t+k}|c_t)$. InfoNCE (Noise Contrastive Estimation) relieves the model from modeling the high dimensional distributions x_{t+k}.

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

$$I(x; c) = \sum_{x,c} p(x, c) \log \frac{p(x|c)}{p(x)} \rightarrow \text{mutual information}$$

Minimizing $\mathcal{L}_N$ implies maximizing $I(x_{t+k}; c_t)$.

Speech, images, text and reinforcement learning!

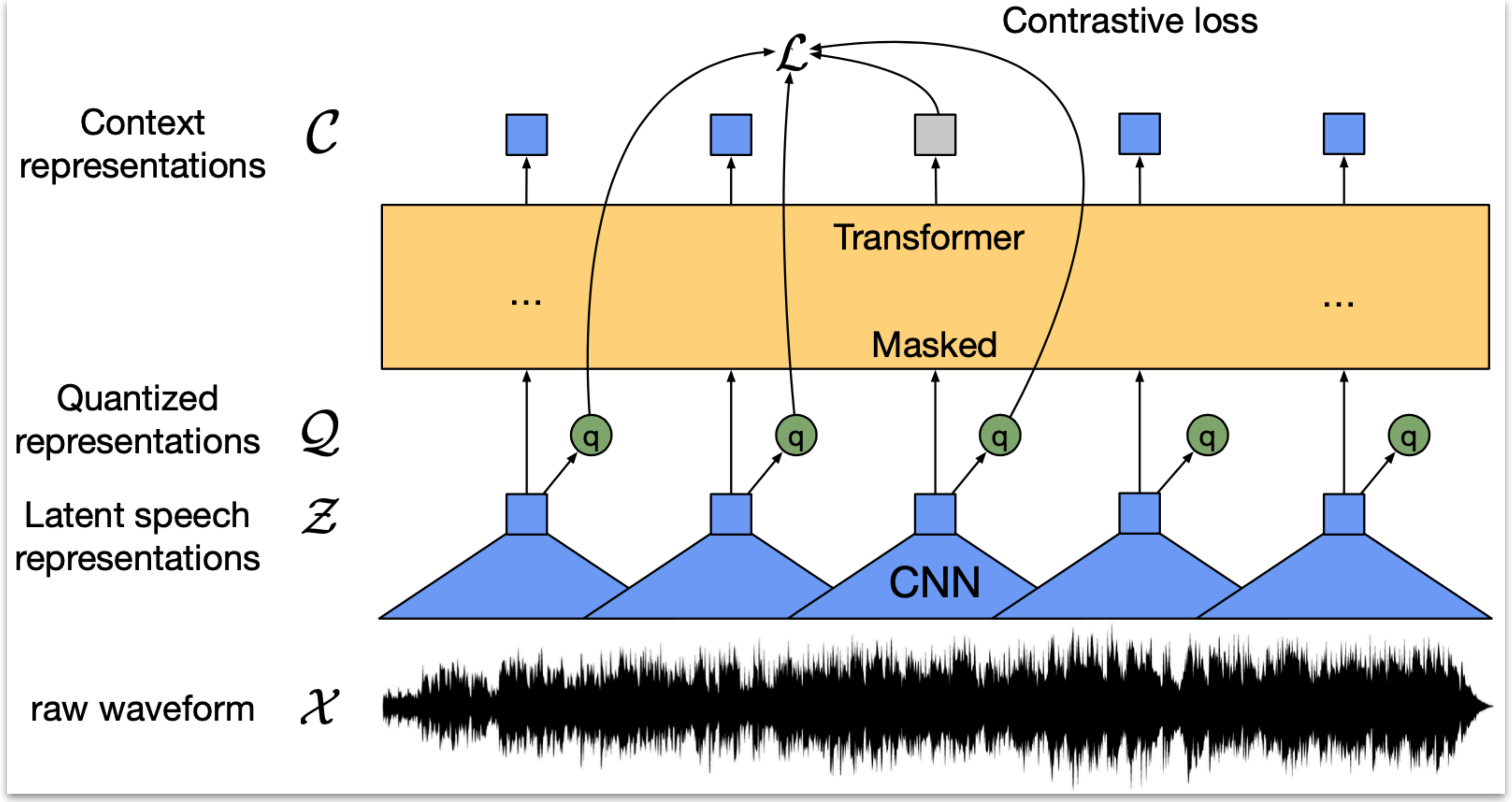
Containing one positive sample from $p(x_{t+k}|c_t)$ and $N-1$ negative samples from the proposal distribution $p(x_{t+k})$

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." *arXiv preprint arXiv:1807.03748* (2018).

wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

Boulder

[YouTube Video](#)



$f : \mathcal{X} \mapsto \mathcal{Z}$

multi-layer convolutional feature encoder

$\mathcal{X} \rightarrow$ input raw audio

$\mathcal{Z} = (z_1, z_2, \ldots, z_T) \rightarrow$ latent speech representations

$g : \mathcal{Z} \mapsto \mathcal{C}$

transformer

$\mathcal{C} = (c_1, c_2, \ldots, c_T) \rightarrow$ representations capturing information from the entire sequence

```
Instead of fixed positional embeddings which encode absolute positional
information, use a convolutional layer which acts as relative positional embedding.
```

$\mathcal{Z} \mapsto \mathcal{Q}$

quantization module

$\mathcal{Q} = (q_1, q_2, \ldots, q_T)$

diversity loss: encourage the model to use the codebook entries equally often $\leftarrow$

Quantization Module

$G \rightarrow$ number of codebooks/groups

$V \rightarrow$ number of entries

$e \in \mathbb{R}^{V \times d/G}$

	avg. WER	std.
Continuous inputs, quantized targets (Baseline)	7.97	0.02
Quantized inputs, quantized targets	12.18	0.41
Quantized inputs, continuous targets	11.18	0.16
Continuous inputs, continuous targets	8.58	0.08

Choose one entry/row from each codebook e and concatenate the resulting vectors $e_1, \ldots, e_G$ and apply a linear transformation $\mathbb{R}^d \rightarrow \mathbb{R}^f$ to obtain $q \in \mathbb{R}^f$. The Gumbel softmax enables choosing discrete codebook entries in a fully differentiable way!

TIMIT phoneme recognition accuracy in terms of phoneme error rate (PER).

	dev PER	test PER
CNN + TD-filterbanks [59]	15.6	18.0
PASE+ [47]	-	17.2
Li-GRU + fMLLR [46]	–	14.9
wav2vec [49]	12.9	14.7
vq-wav2vec [5]	9.6	11.6
This work (no LM) Connectionist Temporal Classification (CTC) loss		
LARGE (LS-960)	7.4	8.3

$z \mapsto l$

$z \rightarrow$ feature encoder output

$l \in \mathbb{R}^{G \times V} \rightarrow$ logits

$$p_{g,v} = \frac{\exp(l_{g,v} + n_v)/\tau}{\sum_{k=1}^{V} \exp(l_{g,k} + n_k)/\tau}$$

probability of choosing the v-th codebook entry for group g

$\tau \rightarrow$ non-negative temperature

$n = -\log(-\log(u)) \rightarrow$ Gumbel noise $\quad u \sim U(0,1)$

Forward pass: $i = \arg\max_j p_{g,j} \rightarrow$ codeword i

Backward pass: true gradient of the Gumbel softmax outputs

Pre-training

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)} \qquad sim(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{b}/\|\mathbf{a}\|\|\mathbf{b}\|$$

contrastive loss: identify the true quantized latent speech representation for a masked time step within a set of distractors.

$$\mathcal{L}_d = \frac{1}{GV}\sum_{g=1}^{G} -H(\bar{p}_g) = \frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V} \bar{p}_{g,v}\log\bar{p}_{g,v} \qquad \mathcal{L} = \mathcal{L}_m + \alpha\mathcal{L}_d$$

Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *arXiv preprint arXiv:2006.11477* (2020).

Boulder

Questions?

YouTube Playlist