On the Vectorization for Speech-to-text

`23.09.18. Paper Study

Hyeonseo Cho

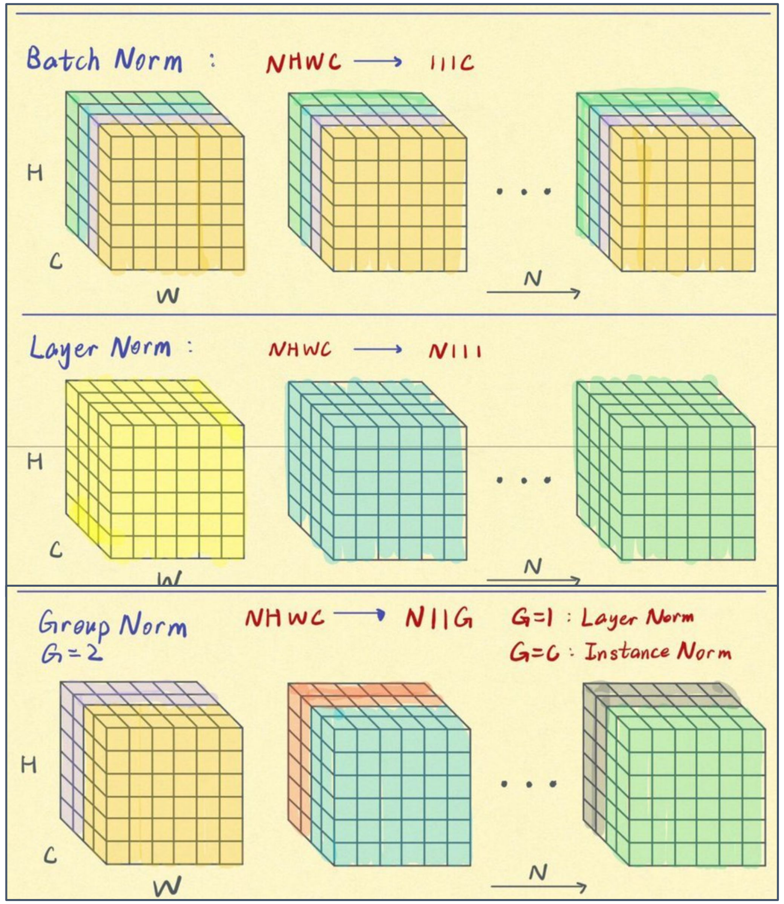
Wav2Vec

Unsupervised Learning for speech recognition

- Pre-training on unlabeled data using *contrastive learning*.
- Applying the learned representation to downstream tasks.

Group Normalization

- "We normalize *both feature and temporal dimension* for each sample which is equivalent to group normalization with a single normalization group"
- Feature dim - various feature in each time step
- Temporal dim - sequences of these features over time

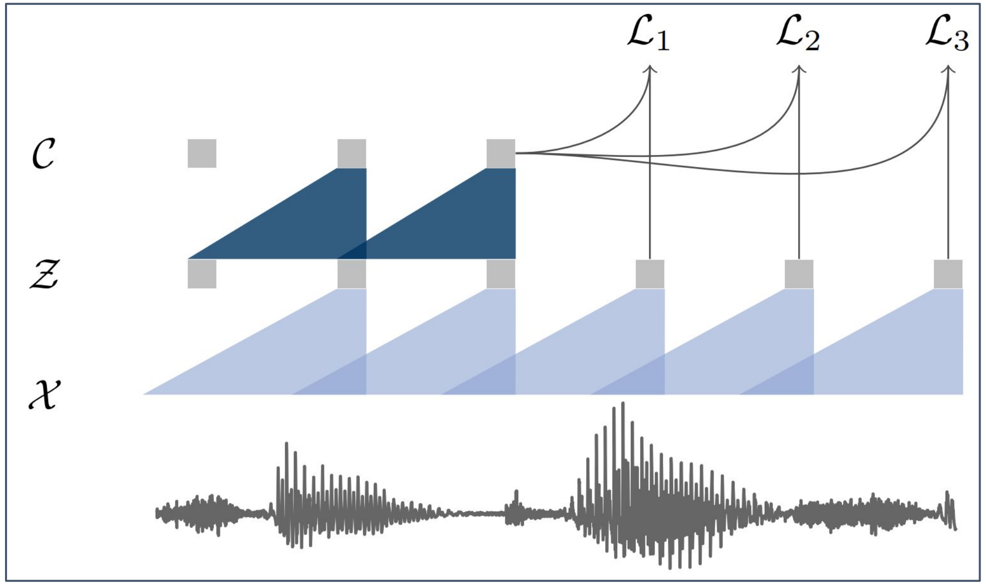


[1] wav2vec: Unsupervised Pre-Training for Speech Recognition., Schneider et al., arXiv 2019.

Wav2Vec

Model

- Encoder + Context network. 512 channels. ReLU. *Group Norm*.



[1] wav2vec: Unsupervised Pre-Training for Speech Recognition., Schneider et al., arXiv 2019.

Wav2Vec

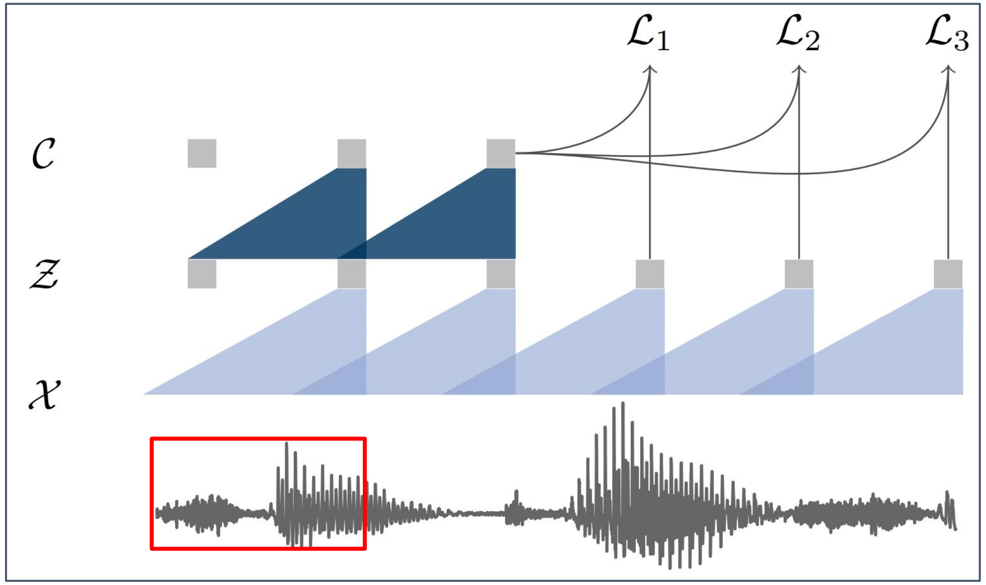
Model

- Encoder + Context network. 512 channels. ReLU. *Group Norm*.
- Encoder Net $f : \mathcal{X} \mapsto \mathcal{Z}$

  Input : raw data $\mathbf{x}_i \in \mathcal{X}$

  Output : Representation $\mathbf{z}_i \in \mathcal{Z}$

  Architecture: 5 layer conv.



Wav2Vec

Model

- Encoder + Context network. 512 channels. ReLU. *Group Norm*.
- Encoder Net $f : \mathcal{X} \mapsto \mathcal{Z}$

  Input : raw data $\mathbf{x}_i \in \mathcal{X}$

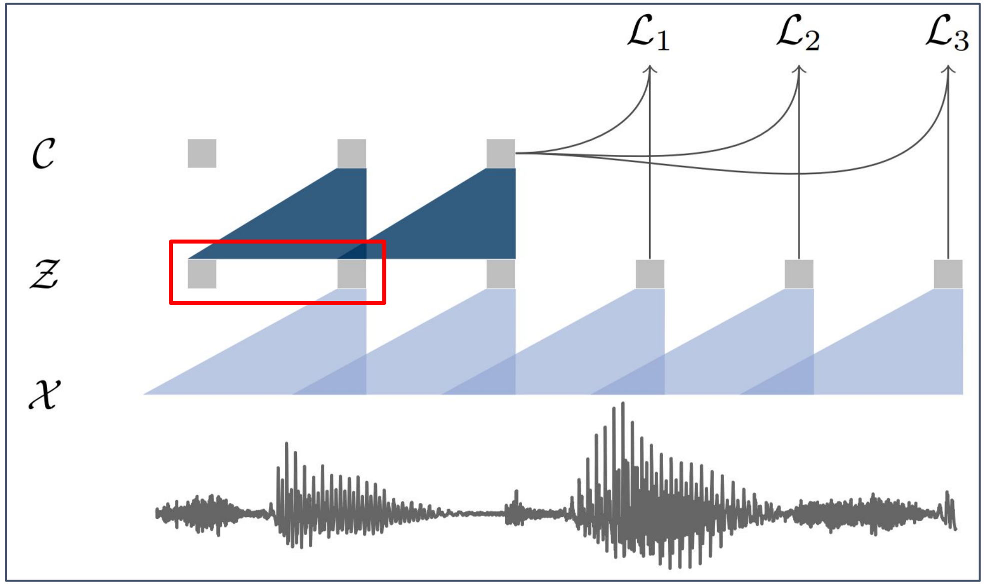
  Output : Representation $\mathbf{z}_i \in \mathcal{Z}$

  Architecture: 5 layer conv.
- Context Net

  Input : latent $\mathbf{z}_i \dots \mathbf{z}_{i-v}$

  Output : Contextualized tensor $\mathbf{c}_i = g(\mathbf{z}_i \dots \mathbf{z}_{i-v})$

  Architecture: 9 layer conv.



Wav2Vec

Contrastive Loss

- Draw Representations of genuine future audio samples closer, while distancing the representations of distractor samples.

$$\mathcal{L}_k = -\sum_{i=1}^{T-k} \left(\log \sigma(\mathbf{z}_{i+k}^\top h_k(\mathbf{c}_i)) + \lambda \mathop{\mathbb{E}}_{\tilde{\mathbf{z}} \sim p_n} \left[\log \sigma(-\tilde{\mathbf{z}}^\top h_k(\mathbf{c}_i))\right] \right)$$

Wav2Vec

Contrastive Loss

- Draw Representations of genuine future audio samples closer, while distancing the representations of distractor samples.

$$\mathcal{L}_k = -\boxed{\sum_{i=1}^{T-k}} \left(\log \sigma(\mathbf{z}_{i+k}^\top h_k(\mathbf{c}_i)) + \lambda \mathop{\mathbb{E}}_{\tilde{\mathbf{z}} \sim p_n} \left[\log \sigma(-\tilde{\mathbf{z}}^\top h_k(\mathbf{c}_i)) \right] \right)$$

Wav2Vec

Contrastive Loss

- Draw Representations of genuine future audio samples closer, while distancing the representations of distractor samples.

$$\mathcal{L}_k = -\sum_{i=1}^{T-k} \left(\log \boxed{\sigma(\mathbf{z}_{i+k}^\top h_k(\mathbf{c}_i))} + \lambda \mathop{\mathbb{E}}_{\tilde{\mathbf{z}} \sim p_n} [\log \sigma(-\tilde{\mathbf{z}}^\top h_k(\mathbf{c}_i))] \right)$$

Wav2Vec

Contrastive Loss

- Draw Representations of genuine future audio samples closer, while distancing the representations of distractor samples.

$$\mathcal{L}_k = -\sum_{i=1}^{T-k} \left(\log \sigma(\mathbf{z}_{i+k}^\top h_k(\mathbf{c}_i)) + \lambda \mathop{\mathbb{E}}_{\tilde{\mathbf{z}} \sim p_n} \left[\boxed{\log \sigma(-\tilde{\mathbf{z}}^\top h_k(\mathbf{c}_i))}\right] \right)$$

Wav2Vec

Contrastive Loss

- Draw Representations of genuine future audio samples closer, while distancing the representations of distractor samples.

$$\mathcal{L}_k = -\sum_{i=1}^{T-k} \left(\log \sigma(\mathbf{z}_{i+k}^\top h_k(\mathbf{c}_i)) + \lambda \mathop{\mathbb{E}}_{\tilde{\mathbf{z}} \sim p_n} \left[\log \sigma(-\tilde{\mathbf{z}}^\top h_k(\mathbf{c}_i))\right] \right) \qquad \mathcal{L} = \sum_{k=1}^{K} \mathcal{L}_k$$

Wav2Vec

Contrastive Loss

- Draw Representations of genuine future audio samples closer, while distancing the representations of distractor samples.

$$\mathcal{L}_k = -\sum_{i=1}^{T-k}\left(\log\sigma(\mathbf{z}_{i+k}^{\top}h_k(\mathbf{c}_i)) + \lambda \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p_n}\left[\log\sigma(-\tilde{\mathbf{z}}^{\top}h_k(\mathbf{c}_i))\right]\right) \qquad \mathcal{L} = \sum_{k=1}^{K}\mathcal{L}_k$$

Decoding

- Transform the acoustic model's output into a sequence of words or characters.

$$\max_{\mathbf{y}} f_{\mathrm{AM}}(\mathbf{y}|\mathbf{c}) + \alpha\log p_{\mathrm{LM}}(\mathbf{y}) + \beta|\mathbf{y}| - \gamma\sum_{i=1}^{T}[\pi_i = \text{`}|\text{`}]$$

References

[1] wav2vec: Unsupervised Pre-Training for Speech Recognition., Schneider et al., arXiv 2019.