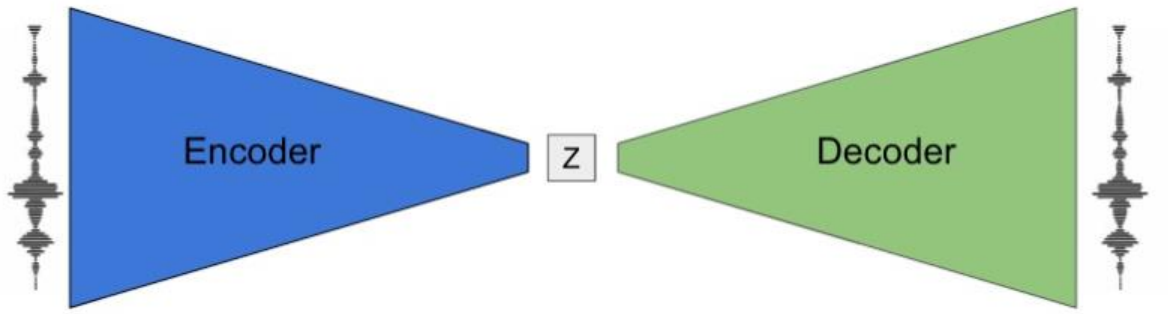
More Music Representations

- Train decoder separately: Multi-Task framework
  - Hierarchical decoder
- Pretrain & finetune: BERT

Hierarchical Encoder

- <A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music>
- high dimensional data -> low dimensional vector
-
- Variational Autoencoder (VAE) has proven to be an effective model for producing semantically meaningful latent representations for natural data.
-
- Bidirectional Encoder
- Hirarchical Decoder

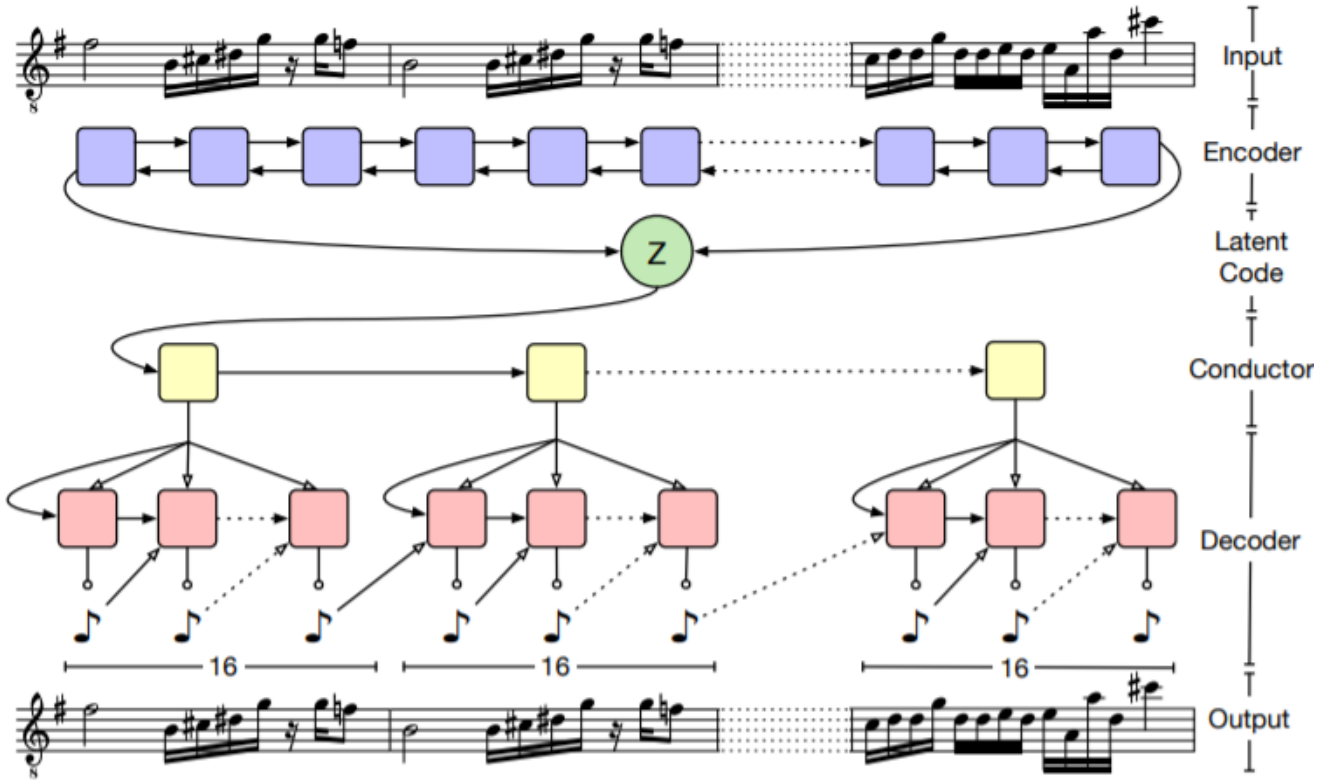


BERT

- VideoBERT: Video -> features vectors -> hierarchical clustering to tokenize visual features
- Pretraning & finetune -> MusicBERT?
-
- Train representation model by performing:
  - Masked prediction -> bar prediction of music
  - Next sentence prediction -> next sentence prediction of music

Hierarchical Encoder

- High abstraction ability of latent vector along with multi-task framework?



More Temporal Representations

- Trained encoder separately
- <Describing Video With Attention-Based Bidirectional LSTM>
- <Weakly Supervised Dense Video Captioning>

Dense Captioning

- describe a video shot with multiple informative and diverse sentences
- Temporal action location & captioning