CONFORMER BASED AUTOMATIC SPEECH RECOGNITION WITH CTC

About Dataset

- Dataset: LibriSpeech “train-clean-100” subset
- Sampling rate: 16 kHz
- Average duration per utterance: ~10 seconds
- 100 hours of clean English speech
- Used for both training and testing

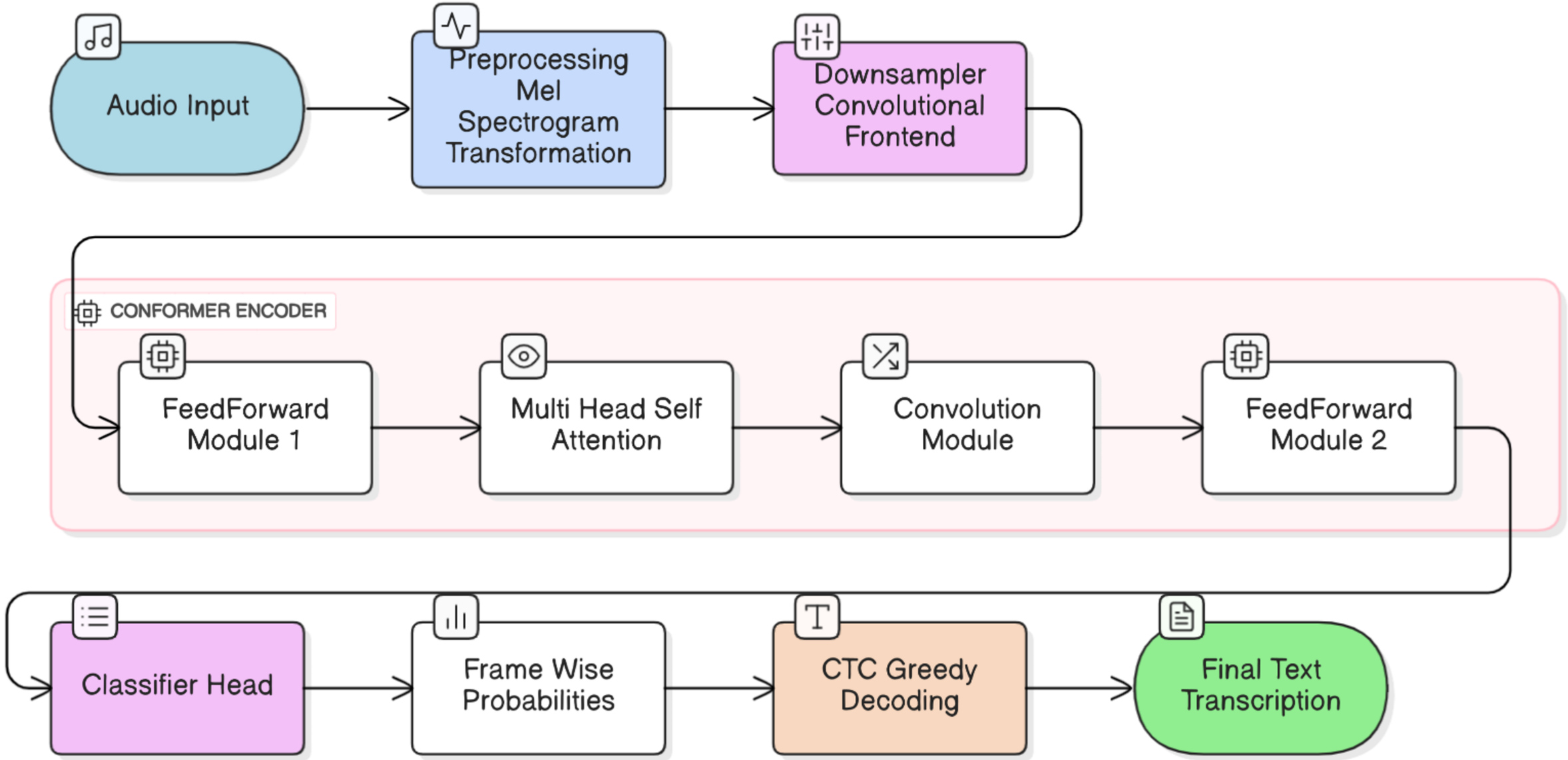
Baseline Model Setup

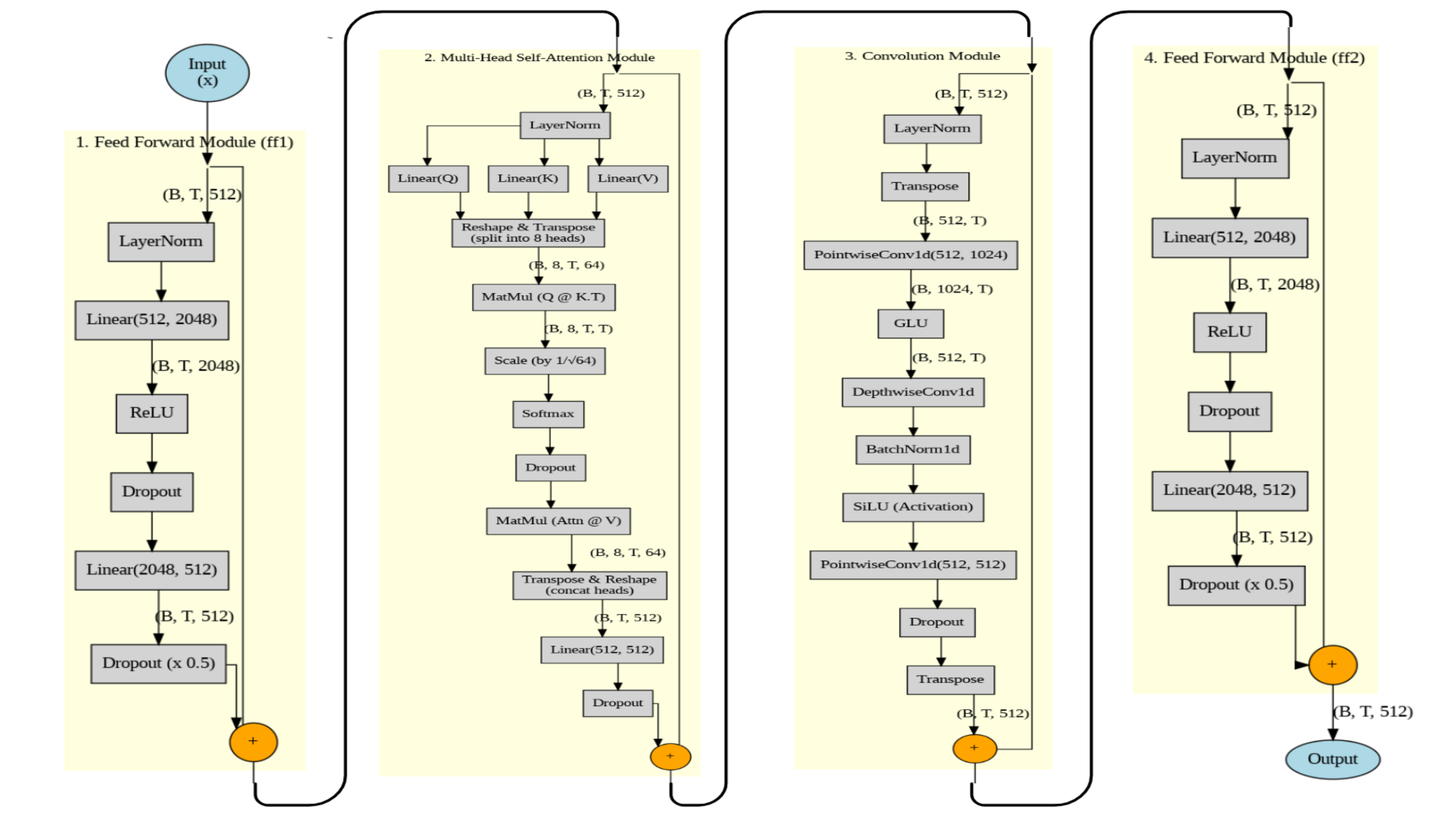
- Trained, full-precision (FP32) Conformer-CTC model as the reference.
- Evaluate on a subset of LibriSpeech ‘test-clean’ to record baseline metrics:
- Word Error Rate (WER) – measures recognition accuracy
- Model size – memory footprint
- Inference latency – time per input

Baseline ASR

- **Input:** Audio Input from the LibriSpeech dataset (Train/Test splits, 16 kHz).
- **Pre-processing:** Initial audio processing such as resampling and cropping.
- **Feature Extraction:** Preprocessing Mel Spectrogram Transformation.
- **Temporal Modeling:** The Downsampler Convolutional Frontend followed by the CONFORMER ENCODER
- **Alignment:** The Classifier Head produces Frame Wise Probabilities which are aligned to the final transcription via CTC Greedy Decoding.
- **Training:** Adam optimizer, CTC loss.
- **Test-Output:** Final Text Transcription.

TRAINING


Audio Input
Preprocessing
Mel
Spectrogram
Transformation
Downsampler
Convolutional
Frontend
CONFORMER ENCODER
FeedForward
Module 1
Multi Head Self
Attention
Convolution
Module
FeedForward
Module 2
Classifier Head
Frame Wise
Probabilities
CTC Greedy
Decoding
Final Text
Transcription


Input
(x)
1. Feed Forward Module (ff1)
(B, T, 512)
LayerNorm
Linear(512, 2048)
(B, T, 2048)
ReLU
Dropout
Linear(2048, 512)
(B, T, 512)
Dropout (x 0.5)
+
2. Multi-Head Self-Attention Module
(B, T, 512)
LayerNorm
Linear(Q)
Linear(K)
Linear(V)
Reshape & Transpose
(split into 8 heads)
(B, 8, T, 64)
MatMul (Q @ K.T)
(B, 8, T, T)
Scale (by 1/√64)
Softmax
Dropout
MatMul (Attn @ V)
(B, 8, T, 64)
Transpose & Reshape
(concat heads)
(B, T, 512)
Linear(512, 512)
Dropout
+
3. Convolution Module
(B, T, 512)
LayerNorm
Transpose
(B, 512, T)
PointwiseConv1d(512, 1024)
(B, 1024, T)
GLU
(B, 512, T)
DepthwiseConv1d
BatchNorm1d
SiLU (Activation)
PointwiseConv1d(512, 512)
Dropout
Transpose
(B, T, 512)
+
4. Feed Forward Module (ff2)
(B, T, 512)
LayerNorm
Linear(512, 2048)
(B, T, 2048)
ReLU
Dropout
Linear(2048, 512)
(B, T, 512)
Dropout (x 0.5)
+
(B, T, 512)
Output