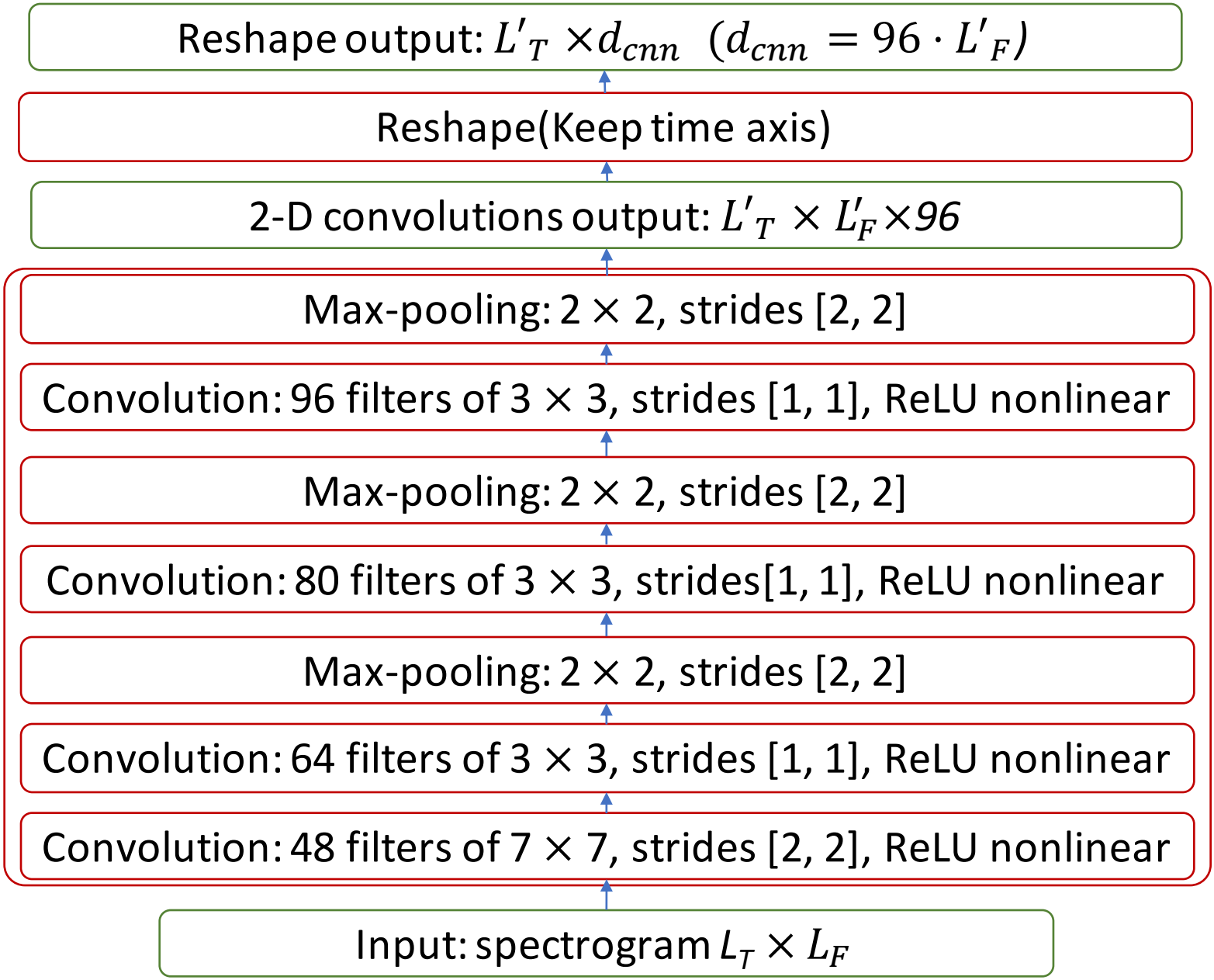
Reshape output: $L'_T \times d_{cnn}$ $(d_{cnn} = 96 \cdot L'_F)$
Reshape(Keep time axis)
2-D convolutions output: $L'_T \times L'_F \times 96$
Max-pooling: 2 × 2, strides [2, 2]
Convolution: 96 filters of 3 × 3, strides [1, 1], ReLU nonlinear
Max-pooling: 2 × 2, strides [2, 2]
Convolution: 80 filters of 3 × 3, strides[1, 1], ReLU nonlinear
Max-pooling: 2 × 2, strides [2, 2]
Convolution: 64 filters of 3 × 3, strides [1, 1], ReLU nonlinear
Convolution: 48 filters of 7 × 7, strides [2, 2], ReLU nonlinear
Input: spectrogram $L_T \times L_F$