# ADL Project: paper summary

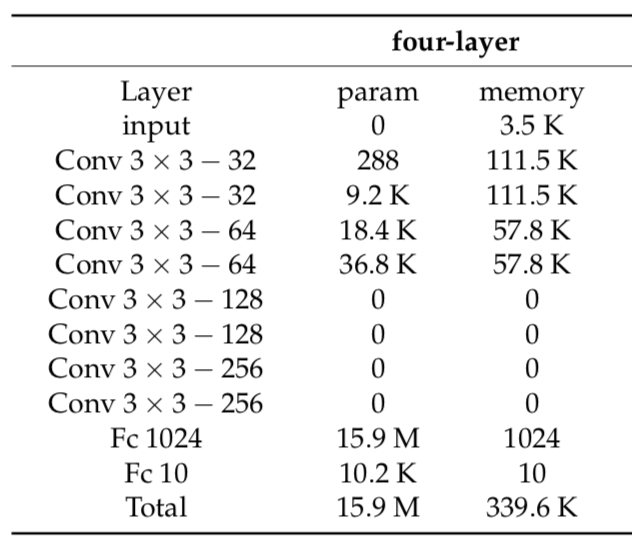
* **Title**: Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion
* Environment sound classification. Using two different aggregated features. Each train a CNN. Then combined using Dempster–Shafer evidence theory -> TSCNN-DS model.
* **Why use CNNs**: “*In recent years, with the advancement of deep learning models, the CNN becomes a primary choice in environment sounds recognition and outperform the conventional classifiers like SVM or GMM.”*
* **Why combine 2 CNNs**: “*environment sounds are mostly non-stationary signals without meaningful patterns or sub-structures, use a single feature may lead to the failure of capturing important information about environmental audio events.”*

## The CNNs (as in section 3.2):

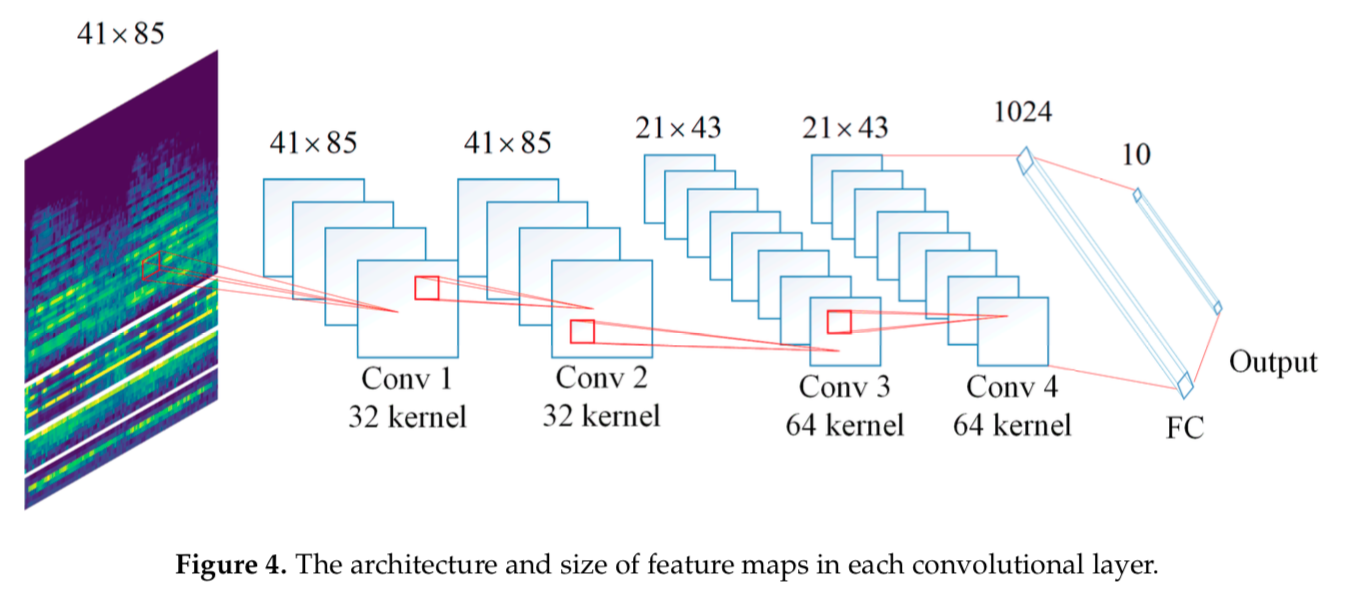
* Four convolutional layers + 1 fully connected layer.
* Conv. Layer 1: nn.conv2d(1, 32, (3,3), stride=(2,2)); batch-normalization; ReLu.
* Conv. Layer 2: nn.conv2d(32, 32, (3,3), stride=(2,2)); batch-normalization; ReLu; nn.MaxPool2d(2).
* Conv. Layer 3: nn.conv2d(32, 64, (3,3), stride=(2,2)); batch-normalization; ReLu.
* Conv. Layer 4: nn.conv2d(64, 64, (3,3), stride=(2,2)); batch-normalization; ReLu.
* Reshape to (57792,1)
* Fully connected layer: nn.Linear(57792,1024); Sigmoid.
* Output Layer: nn.Linear(1024,10); Softmax.
* **Optimizer**: Adam.
* **Dropout**: 0.5 for second, fourth and fully connected.
* **Loss**: cross-entropy.
* The **batch size** is set to 32, while all weight parameters are subjected to **L2- regularization** and **learning rate** is set to 0.001 with the **momentum** of 0.9.

## Inconsistencies:

Table



* If every convolutional layer had stride = 2 as described in section 3.2, the size of the channels would half with each layer. Thus, ignoring the max-pooling, the size of each channel after the fourth layer would be 3x6, which would mean the fully connected layer would have 3x6x64x1024=1.2M parameters (contradicting Table 1).
* Assuming Table 1 is right, each channel needs to be of size 11x22, as 11x22x64x1024=15.9M. This means that either only two layers had a stride step of 2, as described, and max pooling wasn’t done, or max pooling was applied, and only one conv. layer had stride (layer 4).



* *“It can be seen from Figure 4 that the feature maps derived from first and second convolutional layers have the same size as the input feature. After 2 × 2 max pooling processing, the size of input feature maps for third convolutional layer is 21 × 43. Since the max pooling is not performed after convolutional layer 3, so that the size of input features for 4th convolutional layer is 21 × 43 as well. Then, features with size of 11 × 22 are derived from the last hidden layer and feed to the fully-connected layer which has 1024 hidden units. The output is a 1 × 10 tensor according to the number of classed of UrbanSound8K dataset is 10.“* **Section 4 paragraph 3**. It doesn’t say how features with size 11x22 are derived from the last hidden layer.
* **Note**: Max Pool with kernel size 2 has the same effect on the dimensions of the channel as the convolutional layer having stride (2,2).
* In figure 4 we can see the input size for each convolutional layer, this shows clearly that max pooling with kernel size 2 is applied after the second convolutional layer, as described in section 3.2. For this to be consistent with Table 1, Conv 4 must have stride step 2, as suggested in section 3.2, or max pooling be performed after this layer.
* **Conclusion**: the architecture is as follows (described in pyTorch):
  + Conv1 = nn.conv2d(1,32,3) *(41x85)*
  + Conv2 = nn.conv2d(32,32,3) *(41x85)*
  + nn.MaxPool2d(2) *(41x85) -> (21x43)*
  + conv3 = nn.conv2d(64,64,3) *(21x43)*
  + conv4 = nn.conv2d(64,64,3,stride=2) *(21x43) -> (11x22)*
  + fc1 = nn.Linear(15488,1024)
  + fc2 = nn.Linear(1024,10)

## The Features:

* Five auditory features extracted: log-mel spectrogram (LM), MFCC, chroma, spectral contrast and tonnetz (last 3 = CST).
* LM + CST are combined to form feature LMC (41x85).
* MFCC + CST are combined to form feature MC (41x85).
* MFCC + LM + CST are combined to form feature MLMC (41x145).