**Deep learning material**

Neural network questions:

1. Loss function的选择:

Logistic loss (我们需要算probability不只是0或1)

cross entropy for soft max（soft max是占总体百分比，我们各个类别总和不为1）

1. activation function的选择:

sigmoid (vanishing gradient problem是否不同layer需要用不同activation fuction)

ReLU (parameter learned by gradient descent是否必须和gradient descent合用)

1. 基本参数的选择：

number hidden layers（flatter is better）

number of neurons (Ng说通常每层neuron数一致)

batch size （如何确定batch size？）

number of epochs (early stopping 感觉可以从小试起)

optimization method (gradient descent, LG, GS, contract descent, etc.)

Adaptive learning rate (Adagrad 如果不用gradient descent效果如何, momentum如何设定常数值)

Avoid overfitting:

1. early stopping
2. weight decay（是否可以和dropout合用？）
3. dropout（被dropout部分neuron testing时theta如何决定？）
4. cross validation method (k-fold 具体split比例如何决定？)

Deep learning network:

1. CNN (subsampling: 动物，汽车等显著特征)
2. RNN (p222: video caption; p289: acoustic features 是否通过识别语音生成字幕快速甄别视频类别？)
3. Ultra deep network (FractalNet, residual network, highway network)

**Youtube-8M-paper**

Classes:

4800 entities from YouTube video annotation system, 24 top-level categories

Models:

a). frame-level

1. Frame-level model and average pooling

Train 4800 independent 1 vs all classifiers to all entities, and then aggregate the score for annotations to video-level score using simple average

1. Deep Bag of Frame (DBoF) pooling (regardless of sequence of frames)

Layers: input layer (N\*k) 🡪 fully connected layer (M, ReLU) 🡪 max pooling layer 🡪 … fully connected layer … 🡪 output layer (Logistic or Softmax)

Optimization: Stocastic Gradient Descent (SGD)

Loss function: logistic loss for logistic layer and cross-entropy loss for softmax layer

1. Long Short-Term Memory (LSTM) (include the effect of frame sequence)

2 layers with 1024 units, linearly increasing per-frame weights from 1/N to 1 for the last frame

b). video-level

1. Logistic Regression
2. Hinge Loss
3. Mixture of Experts (MoE)

Perform very well for multi-modal classification

Conclusion:

In terms of frame-level models, LSTM performs best;

In terms of video-level models, MoE performs best;

MoE and LSTM perform almost the same because the underlying frame-level features are already very strong

**目前模型思路总结**

1. Frame-level 和 video-level得分开来，video-level其实是通过frame-level的信息aggregate得到的数据，所以不管视频长短，video-level的feature length都是相同的

2. 先说video-level，基本网上的paper都是讲关于frame-level的，只有Youtube8M paper这个讲到了video-level modeling，主要结论是因为youtube这个frame-level的feature足够强，所以video-level model的效果其实就很好了，因为video-level数据是frame-level aggregate出来的。然后 video-level应该不需要用到神经网络，因为数据比较简单，在 (logistic, SVN, etc.) 这些模型里面，MoE模型的效果应该是最好的，这个感谢裴龙他已经在跑这个了，应该我们很快就能看到结果。

3. frame-level的话，也就是细分到每一个画面的数据，这个目前比较效果比较好的方法主要有三种：

1). Deep-Bag-of-Frame (DBoF)

这种就是把不同时间的single-frame的数据屯到一块儿， 再用不同的pooling方法把temporal的信息融入进去，这种算法主要区别在于选择不同的pooling方法，然后它不包含frame的顺序信息。

2). LSTM (Long-Short-Term-Memory) 或LSTM & CNN

这种就是主要用RNN，记忆门的原理。缺点是LSTM原来不是为video classification设计的，视频相邻时间的数据会很接近，这可能会影响LSTM的效果。将所有single-frame数据先CNN再用LSTM会得到很高的正确率，但这样运算量十分巨大

3). RCNN

这个就是完全将RNN和CNN结合在一起的一种模型，效率会比上面提到的先CNN再RNN的那种高，但预测结果来说稍微差一点。在sampling的时候我们需要使用random sampling而不是sequential sampling，原因就像上面说的因为视频前后关联性很强，用sequential sampling会有redundancy的情况。

4. 其他问题：

在源数据方面，有几个paper在用frame-level之余还用到了optical flow，把optical flow和frame-level一起train会得到更高的正确率，因为考虑到了更多motion的信息，不知道我们能否获得这个的数据。基本所有paper都用了stochastic gradient descent (SGD)来取代gradient descent因为learning rate不robust。