Applied Data science

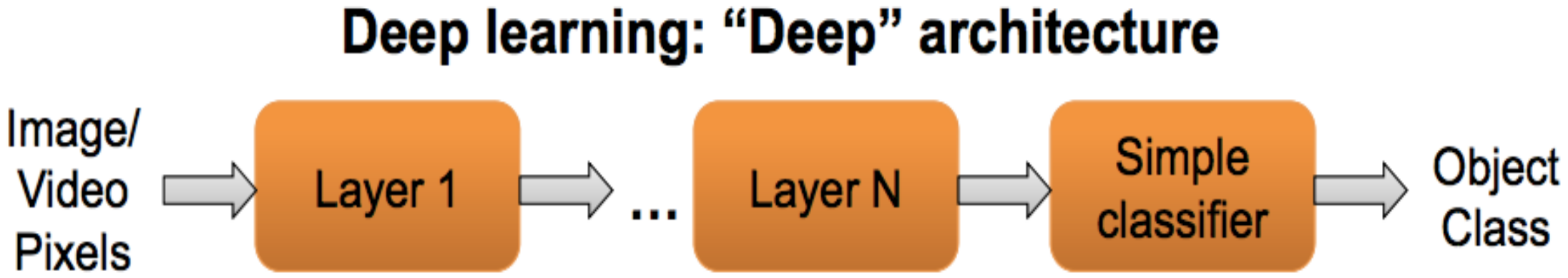
Spring 2017_Project 3_Group 14

- Han, Ke (kh2793)
- Li, Mengchen (ml3890)
- Mison, Virgile (vcm2114)
- Pan, Yijia (yp2424)
- Xiang, Yi (yx2365)

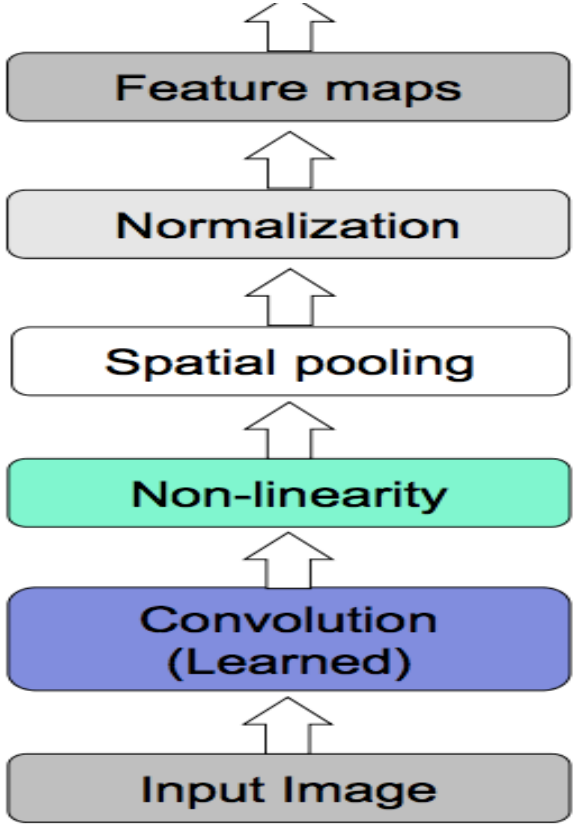
Models Comparison

feature	model	parameters	accuracy
sift	**GBM**	**ntree = 100, depth =1, shrinkage = 0.1 (CV)**	**0.73**
sift (LASSO)	GBM	ntree = 100, depth =1, shrinkage = 0.1 (CV)	0.64
sift (LASSO)	KNN	k = 3	0.65
sift (LASSO)	XG boost	objective = 'logistic', max_depth = 7, eta = 0.11, gamma = 0.01	0.68
sift (LASSO)	random forest	ntree = 600 (CV)	0.67
sift (LASSO) + texture	majority vote (XG boost)	...	0.69
CNN	**CNN**	**num.round = 750, learning.rate = 0.3, dropout = 0.7**	**0.81**

Deep Learning (CNN)



Additional Feature Extraction (CNN)



Problems + Strategies

- 2000 images -> 1600 train + 400 test
- **Deep Learning (mxnet)**
- **Results:** train acc=95% vs. test acc=75%
- **Problem:** Overfitting !!!
- **Reason:** Deep networks need to be trained on a huge number of training images to achieve satisfactory performance
- **How to solve?** Ans: **data augmentation** (add to the general training dataset images that have been flipped horizontally and also with one rotation of small angles)
- **Strategies:** 2000 images -> 4000 images, improved!