Automated Image Interpretation via ConvNets & Recurrent Networks

SysEng 5212 /EE 5370 Project

Student: Xiongming Dai

Roadmap

1. Project Objectives
2. Methodology
3. Experiment
4. Result
5. Conclusion
6. References
7. Acknowledgements

1.Project Objectives

- Develop a method of automatical processing and analyzing images based on Deep Neural Network

"black and white dog jumps over bar."

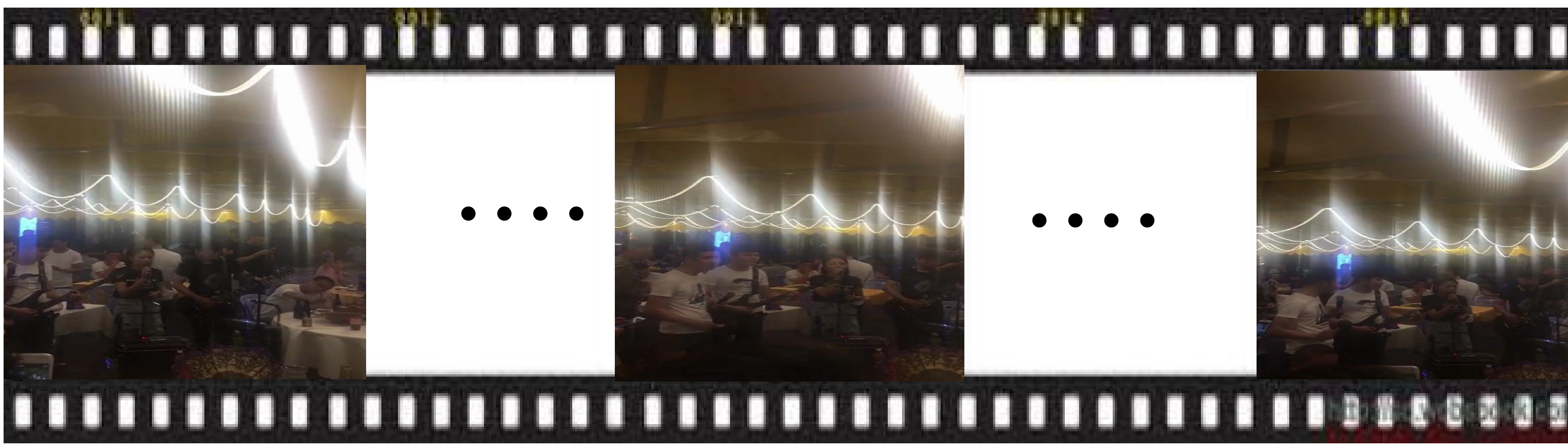
"construction worker in orange safety vest is working on road."

"girl in pink dress is jumping in air."

1.Project Objectives

- **Input: Frames from a video or a single image**

Frame 1 **Time track** Frame N



1.Project Objectives

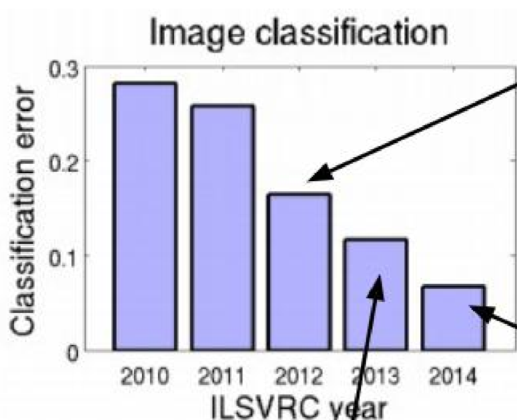
- **Output: Image Captioning**

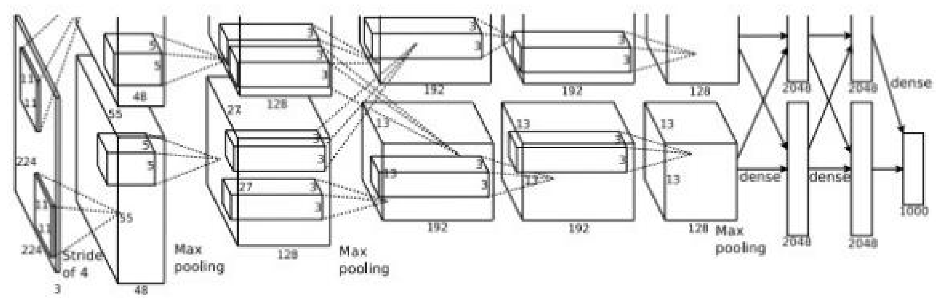
Four people are singing in a party.

2.Methodology

- image detection and classification

[Krizhevsky, Sutskever, Hinton. 2012] **16.4% error**

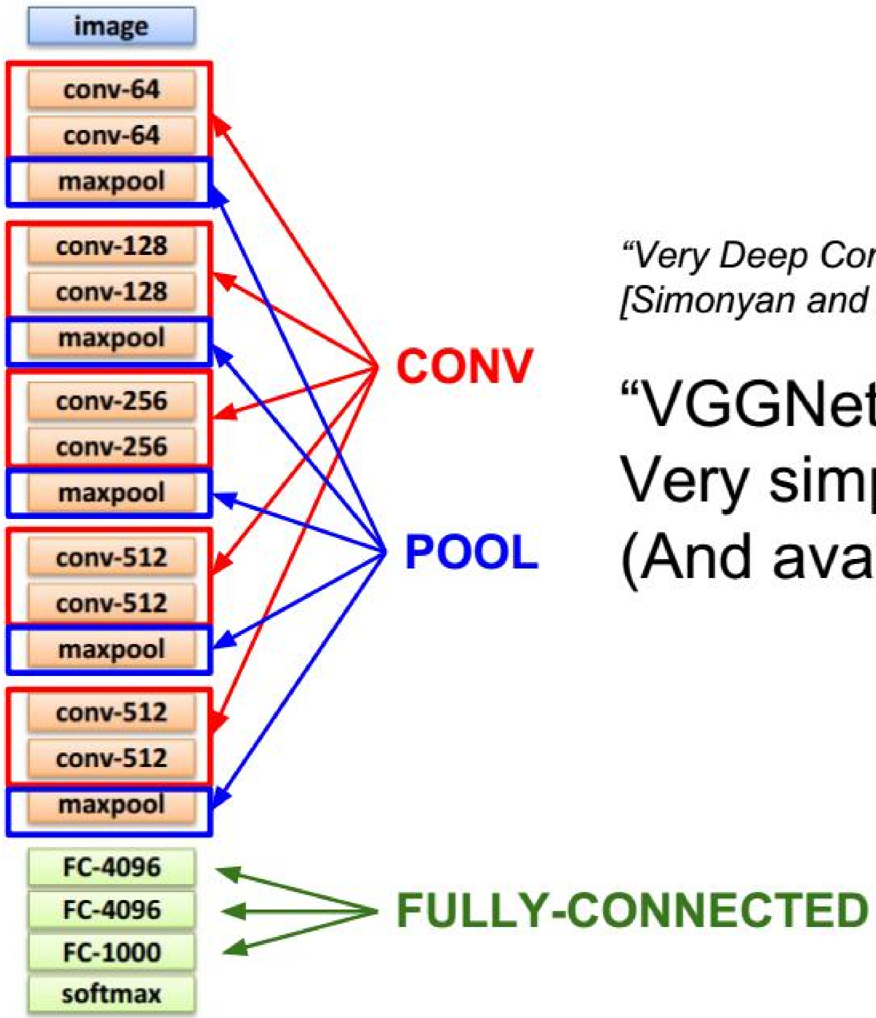




[Szegedy et al., 2014] **6.6% error**
[Simonyan and Zisserman, 2014] **7.3% error**

[Zeiler and Fergus, 2013] **11.1% error**

2.Methodology



"Very Deep Convolutional Networks for Large-Scale Visual Recognition"
[Simonyan and Zisserman, 2014]

"VGGNet" or "OxfordNet"
Very simple and homogeneous.
(And available in **Caffe**.)

2.Methodology

1) Detection (by convolutional neural network)

- 19 locations for detecting the objects based on Region- CNN
- The embedding vectors based on the pixels I_b

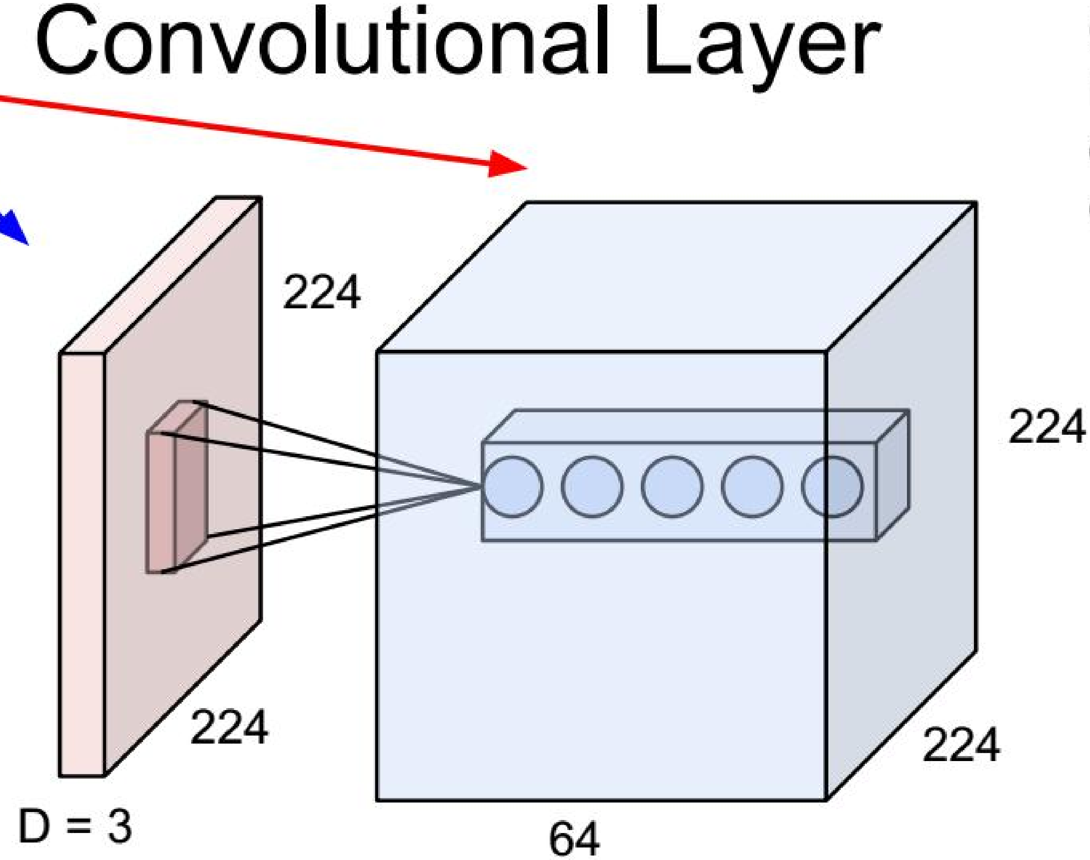
Inside the bounding box :

$$v = W_m[CNN_{\theta_c}(I_b)] + b_m.$$

Where $CNN_{\theta_c}(I_b)$ takes the image inside a given bounding box and returns the 4096-dimensional activations of the fully connected layer before the classifier, nearly 60 million parameters.

2.Methodology

Convolutional Layer

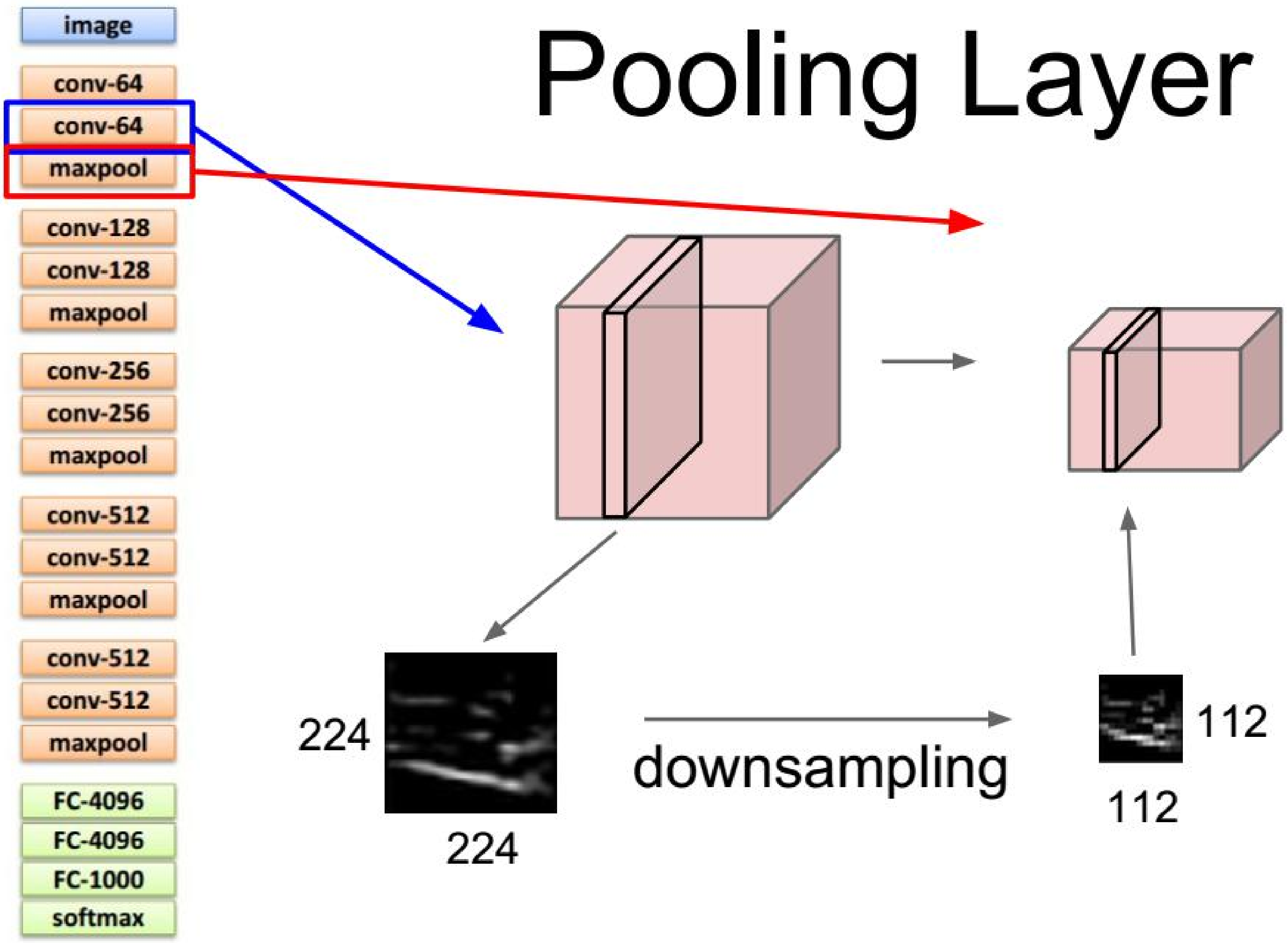


Can be implemented efficiently with convolutions

Every blue neuron is connected to a **3x3x3** array of inputs

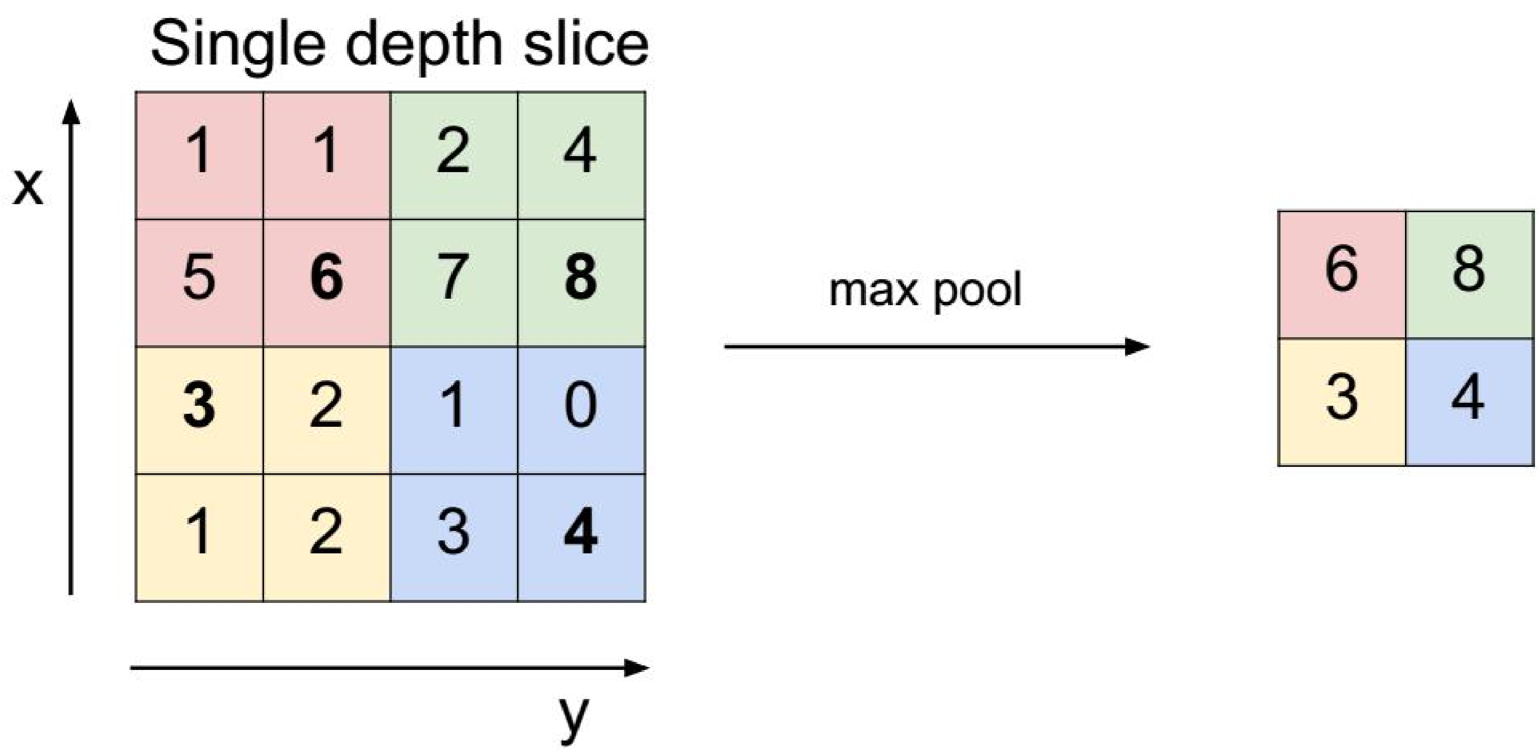
2.Methodology

Pooling Layer



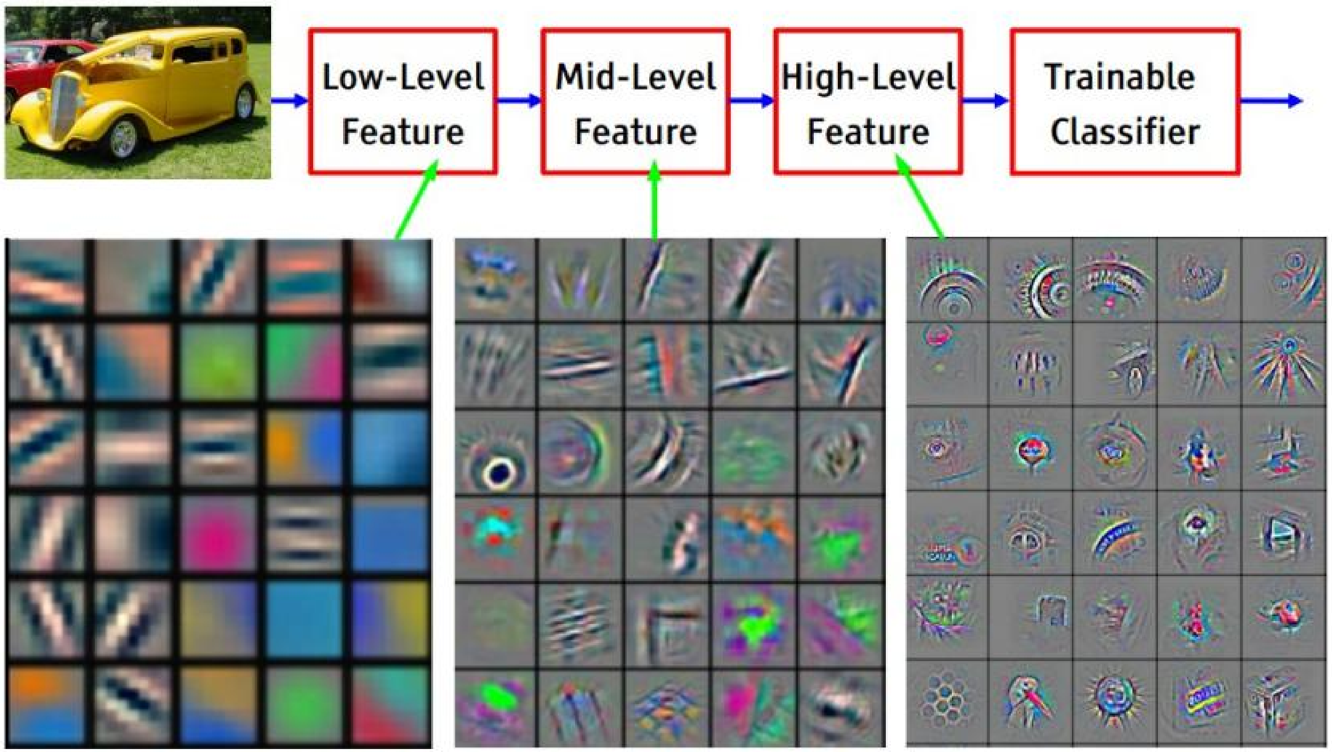
2.Methodology

Max Pooling Layer



2.Methodology

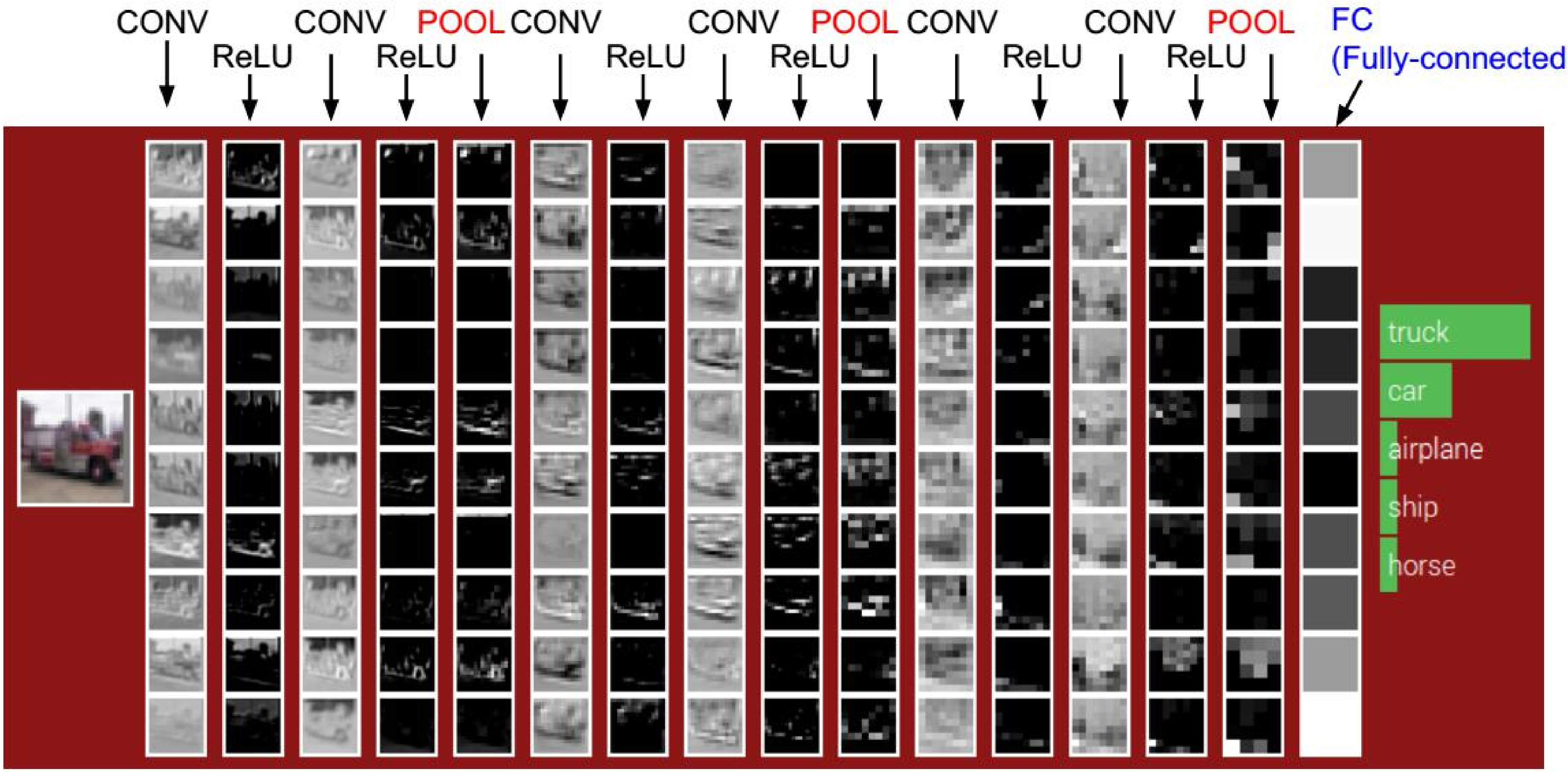
What do the neurons learn?



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

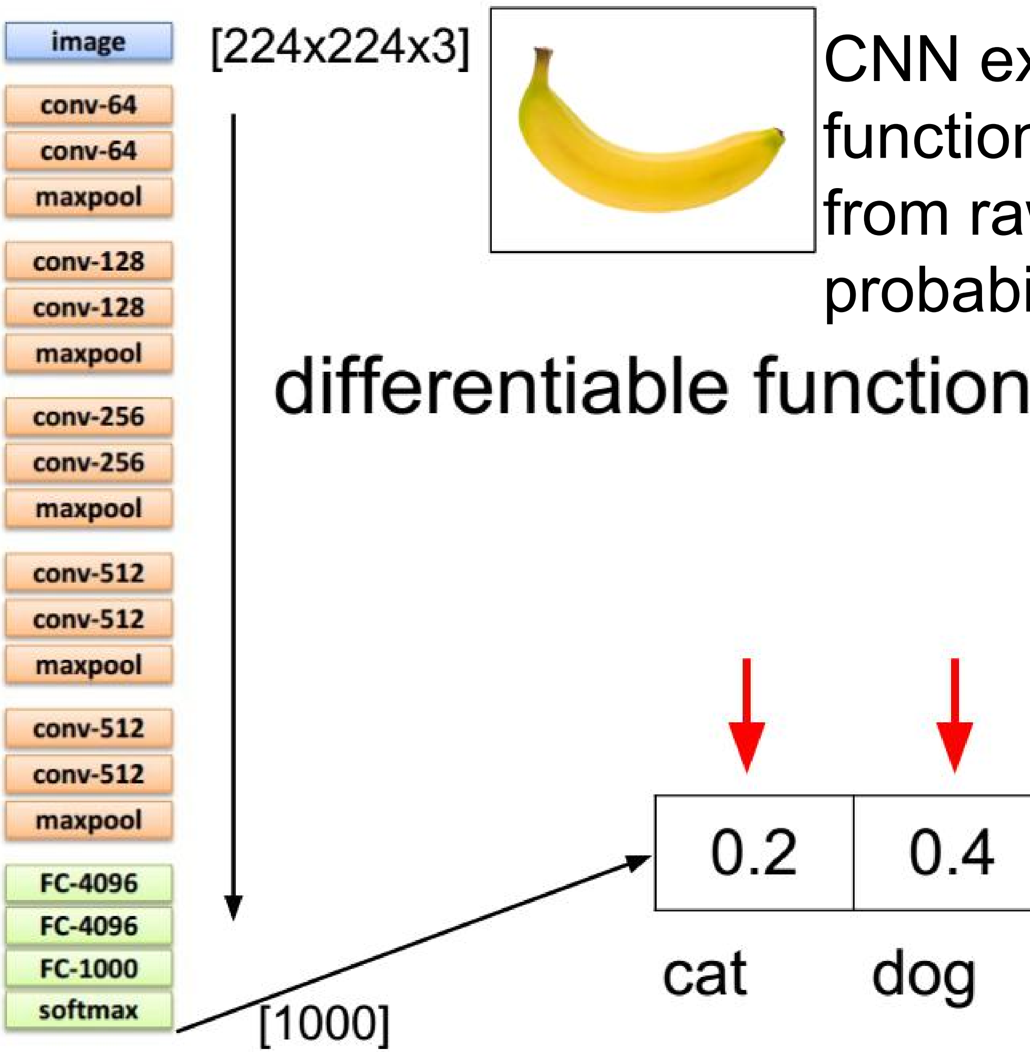
2.Methodology

Example activation maps



2.Methodology

CNN express a single differentiable function from raw image pixel values from raw image pixel values to class probabilities

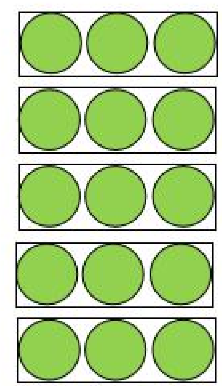


2.Methodology

2) Sentence fragments

Dependency Relations
"A red tanker caught fire with smoke on the highway."
(red, tanker, highway) →
(black, fire, smoke) →
(on, highway, smoke) →
(fire, a, with) →
(tanker, caught, fire) →

$\{w_1, w_2\}$



Embedding space: $s = f(W_R \begin{bmatrix} W_e w_1 \\ W_e w_2 \end{bmatrix} + b_R).$

2.Methodology

2) Sentence fragments

Where W_e is a 200X400,000 matrix that encodes a 1-of-k vector into a 200-dimentional word vector representation.

I_b

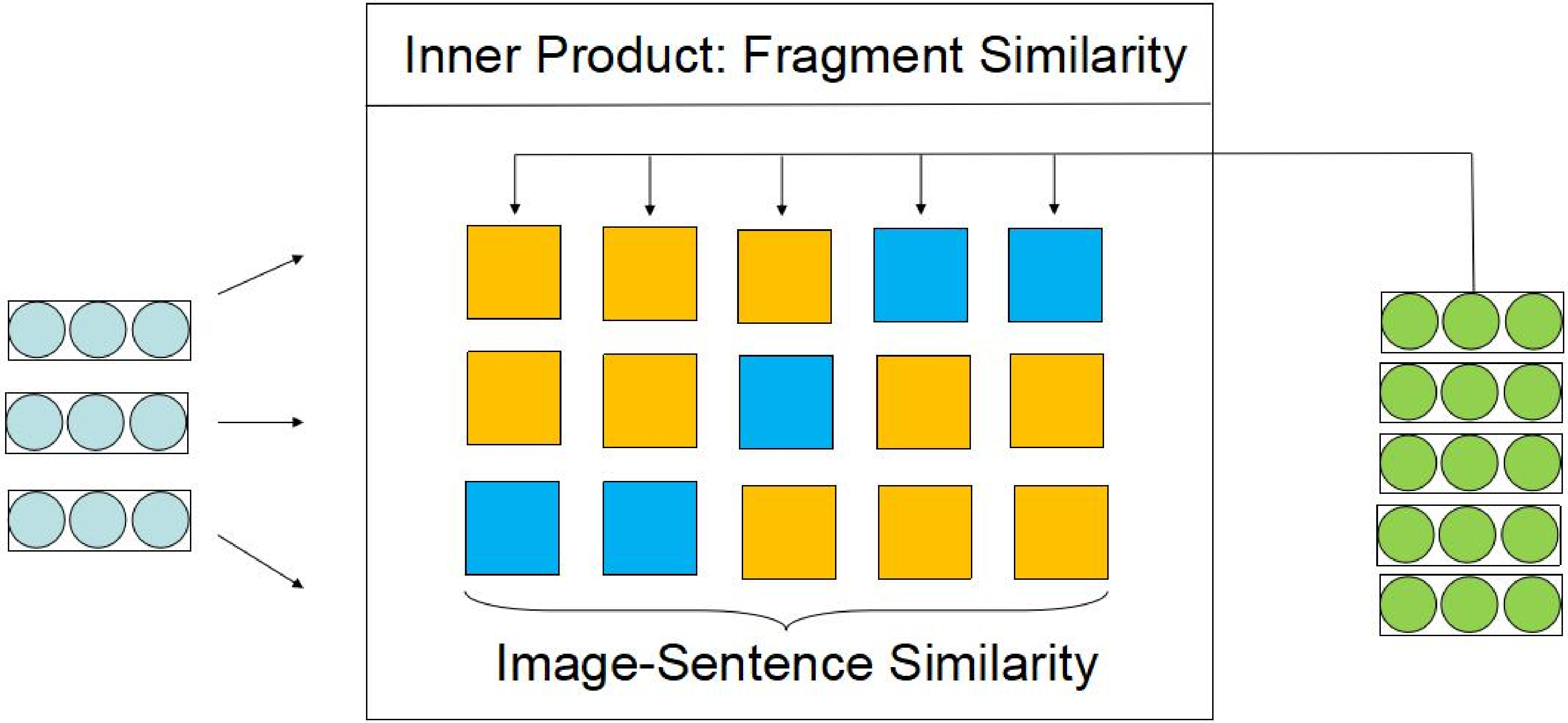
Every relation R can have its own set of weights and biases.

f(.) to be the Rectified Linear Unit, x from max(0,x).

$$s = f\left(W_R \begin{bmatrix} W_e w_1 \\ W_e w_2 \end{bmatrix} + b_R\right).$$

2.Methodology

3) Fragment alignment



Objective function:

$$\Gamma(\theta) = \Gamma_F(\theta) + \beta\Gamma_G(\theta) + \alpha||\theta||_2^2, \theta = \{W_e, W_R, b_R, W_m, b_m, \theta_c\}.$$

2.Methodology

3) Fragment alignment

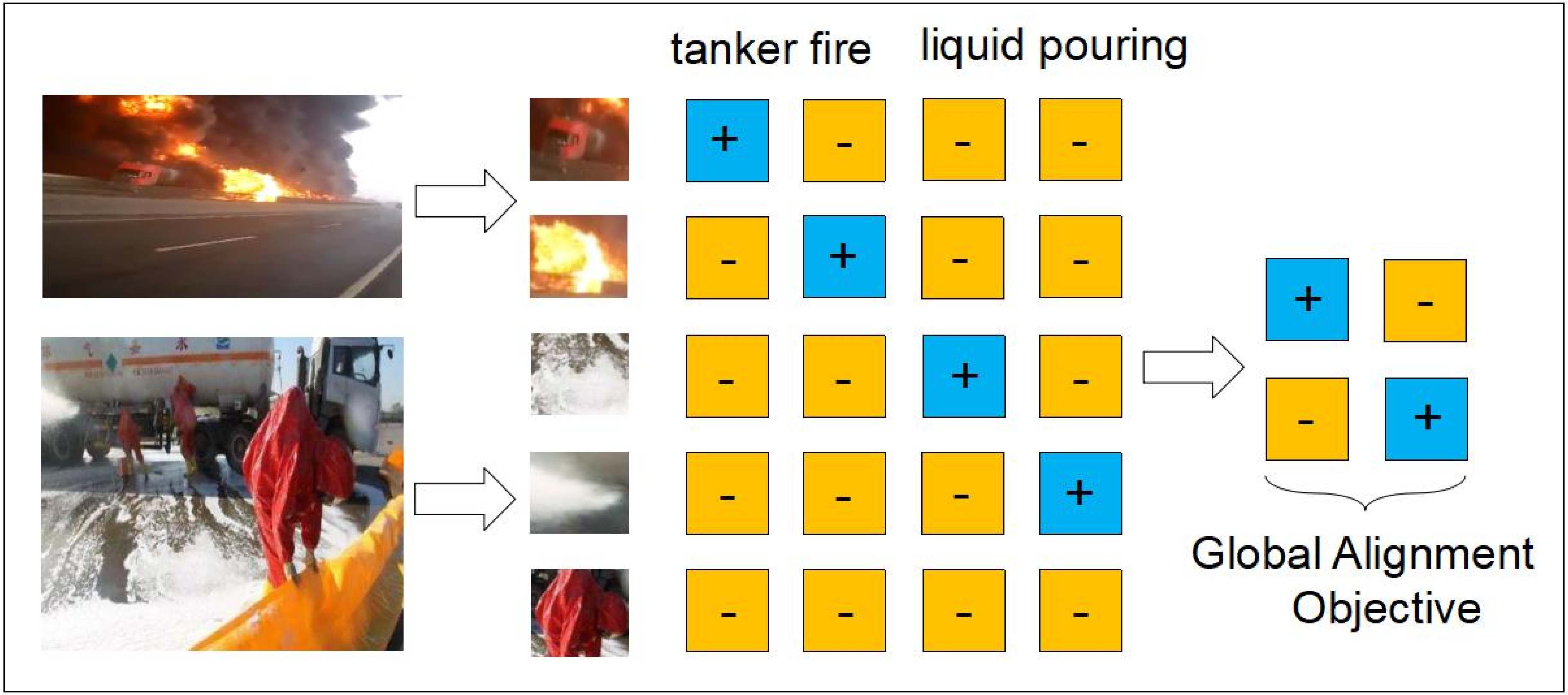
$$\Gamma(\theta) = \Gamma_F(\theta) + \beta \Gamma_G(\theta) + \alpha ||\theta||_2^2,$$
$$\theta = \{W_e, W_R, b_R, W_m, b_m, \theta_c\}.$$

- Where $\Gamma_F(\theta)$ is the Fragment Alignment Objective, $\Gamma_G(\theta)$ is the Global Ranking Objective, θ is a shorthand for parameters of the neural network. α, β are hyperparameters that we cross-validate.

I_b

2.Methodology

3) Fragment alignment



2.Methodology

3) Fragment alignment

Incomplete alignment objective:

Incomplete alignment objective:

$$\Gamma_0(\theta) = \sum_i \sum_j \max(0, 1 - y_{ij} v_i^T s_j)$$

- $v_i^T s_j$: alignment score of visual fragment and sentence fragment. y_{ij} =+1 if occur together,
- y_{ij} =-1 otherwise.

3.Experiment

Microsoft COCO
[Tsung-Yi Lin et al. 2014]
mscoco.org

Linux 16.04
GeForce 10 series
Caffe

we use 5,000 images for both validation and testing.

currently:
~120K images
~5 sentences each

Data Preprocessing: Convert all sentences to lower-case,discard non-alphanumeric characters, and filter words to those that occur at least 5 times in the training set, which results 8791 words for MSCOCO.

4. Result

a group of people standing around a room with remotes
logprob: -9.17

a young boy is holding a baseball bat
logprob: -7.61

a cow is standing in the middle of a street
logprob: -8.84

4. Result

a cat is sitting on a toilet seat
`logprob: -7.79`

a display case filled with lots of different types of donuts
`logprob: -7.78`

a group of people sitting at a table with wine glasses
`logprob: -6.71`

a baby laying on a bed with a stuffed bear
`logprob: -8.66`

a table with a plate of food and a cup of coffee
`logprob: -9.93`

a young boy is playing frisbee in the park
`logprob: -9.52`

5. Conclusion

- We introduce a model that generates natural uage descriptions of image regions based on weak labels in form of a dataset of images and sentences.
- Our approach features a novel ranking model that aligned parts of visual and language modalitie through a common, multimodal embedding.
- we evaluated its performance on both fullframe and region-level experiments.

6. References

- 1.Addrej.Feifei Li. Deep Visual-Semantic Alignments for Generation Image Descriptions, CVPR,2015
- 2.R.Girshick,Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR,2014
- 3.A.Krizhky,Hinton. Imagenet classification with deep CNN. NIPS,2012
- 4.J.Mao. Explain images with multimodal RNN, CVPR,2014
- 5.Zisserman. Very deep CNN for Large-scale visual recognition,NIPS,2014

7. Acknowledgements

- I gratefully acknowledge Dr.Dagli for always hearty teaching, comments,discussion, and guidance on the class of “SysEng 5212/EE5370 Neural Network”