

Computational Linguistics: Language and Vision I

RAFFAELLA BERNARDI

Contents

1	Credits	7
2	What is (Computer) Vision	8
	2.1 Interdisciplinary	9
	2.2 How did it started?	10
	2.3 What is Computer Vision goal?	11
3	How to represent an image: Pixels	12
	3.1 How to represent an image: Keep all the pixels	13
	3.2 How to represent an image: Compute average pixel	14
	3.3 How to represent an image: Spatial grid of average pixel colors?	16
	3.4 Image representation challenges: Invariance	17
4	A CV sample task: Object Classification	18
	4.1 Object Classification	19
	4.2 Data Driven	20
	4.3 The image classification pipeline	21
	4.4 Nearest Neighbor Classifier	22
	4.5 Nearest Neighbor examples	23

4.6	Image distance	24
4.7	Evaluation	25
4.8	K-Nearest Neighbor Classifier	26
4.9	Validation dataset vs Test dataset	27
4.10	First problem: the classifier	28
4.11	Second problem: the Raw Pixel representation	29
5	Representation Problem: From pixel to feature	31
5.1	Two methods	32
5.2	Bag of Visual Words: Pipeline	33
5.3	Low-level Features extraction	34
5.4	Characteristics of good low-features	35
5.5	Example visual vocabulary	36
5.6	Image Representation	37
5.7	Summary: Images representation pipeline	38
5.8	From hand-crafted feature to feature learning	39
5.9	Convolutional Neural Network: transfer	40
5.10	Inspiration	41
5.11	Hierarchy of features	42
6	Classifier problem	43

6.1	Score and Loss functions: example	44
6.2	Score and Loss functions	45
6.3	Score function: Linear Classifier	46
6.4	Loss Function: Super Vector Machine	47
6.5	Linear Classifier: cartoon representation	48
6.6	non linear problems	49
7	Applications: CV exploits NLP and vice-versa	50
8	Computer Vision exploits language	51
8.1	Traditional CV task: Object recognition	52
8.2	Object recognition: methods	53
8.3	Corpora as KB source: Object recognition	54
8.4	Corpora as KB source: Action recognition	55
8.5	Caption generation	56
8.6	Caption generation: biblio	57
9	Visual Question Answering	60
10	NLP exploits vision	62
10.1	Lexical Preference	63
10.2	Translation	64
10.3	Co-reference Resolution	65

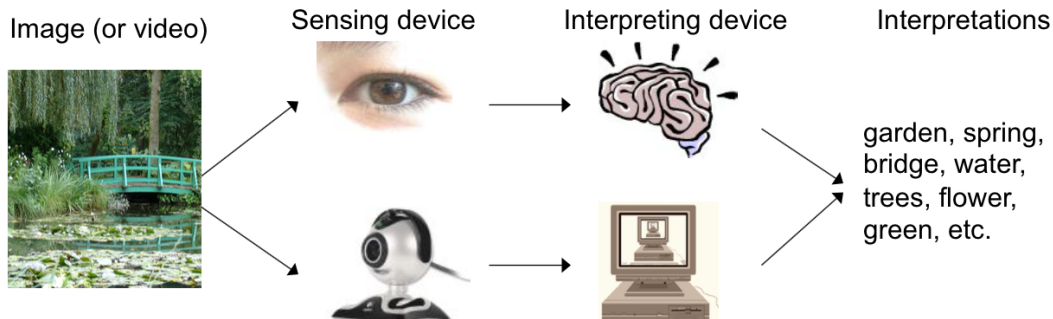
10.4	Co-reference Resolution	66
11	Summary: CV and NLP	67
12	Foundational: Grounding	68
13	Foundational: Reference	69
14	Data Set	70
14.1	CIFAR	71
14.2	ImageNet	72
14.3	VisA	73
14.4	SUN	74
15	Dataset for sentence-based image description	75
15.1	Online Caption?	76
15.2	Photo-sharing?	77
15.3	Photo-sharing?	78
15.4	IAPR-TC12 data set	79
15.5	ILLINOIS PASCAL data set	80
15.6	Crowdsourcing	81
15.7	Crowdsourcing results	82
15.8	LabelMe	83
16	Demos TBD	84

17	Softwares	85
18	Language and Vision Research Groups	86
19	Language and Vision	88
20	Other Useful Links	89

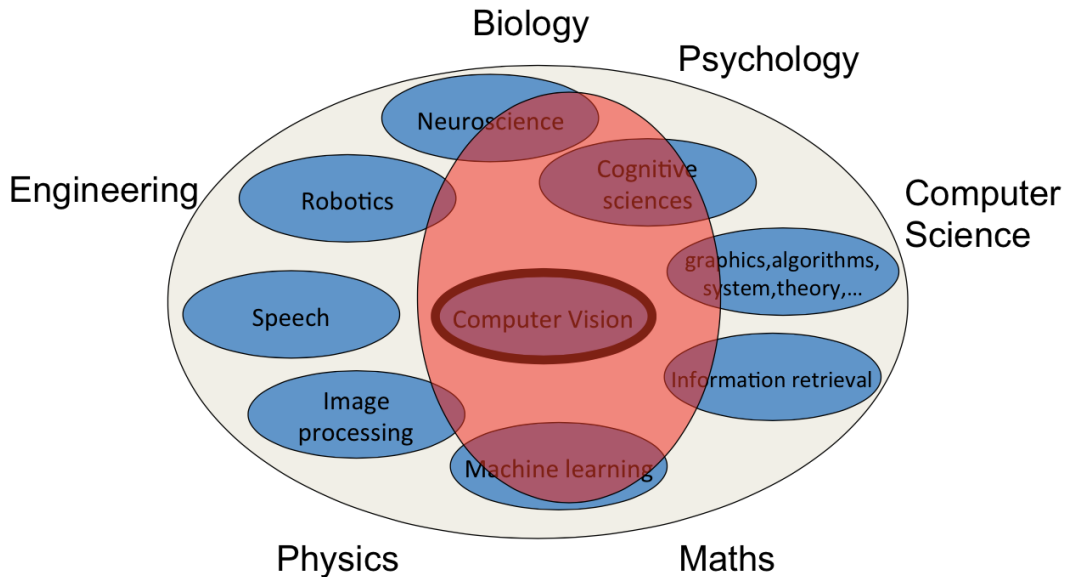
1. Credits

Honglak Lee, L. Fei Fei, Tamara Berg, Angeliki Lazaridou, Elia Bruni, Marco Baroni, Desmond Elliott, Douwe Kiela,

2. What is (Computer) Vision

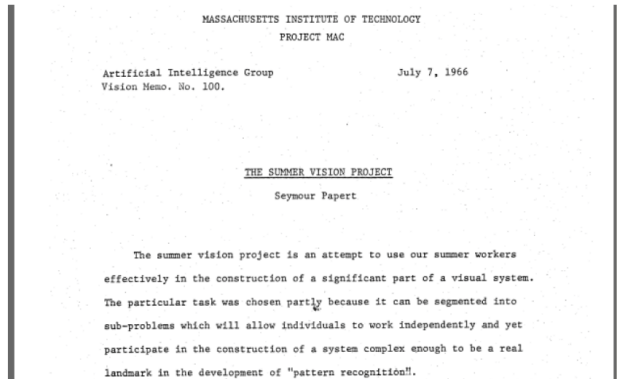


2.1. Interdisciplinary



2.2. How did it started?

Origins of computer vision: an MIT undergraduate summer project



2.3. What is Computer Vision goal?

- To bridge the gap between pixels and “meaning”



What we see

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

What a computer sees

Source: S. Narasimhan

3. How to represent an image: Pixels

Raw images representation consists of pixels (a pixel is the minimum element of an image).

Pixels, identified by their physical coordinates, are stored as numbers encoding their color intensity. For instance, a black and white image is a 1-D representation of the pixel brightness); a colored image is a 3-D arity of intensity values:

$$f(x, y) = \begin{bmatrix} red(x, y), \\ green(x, y), \\ blue(x, y) \end{bmatrix}$$

where $color(x,y)$ is the intensity of that color (x) at position (y).

If we want to retrieve images similar to a given one, or we want to recognize the object in an image or perform other tasks, pixel representations are not suitable, we need to have an abstract representation of the image.

3.1. How to represent an image: Keep all the pixels



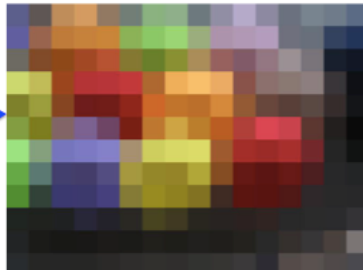
3.2. How to represent an image: Compute average pixel



3.3. How to represent an image: Spatial grid of average pixel colors?



Photo by: [marielito](#)



3.4. Image representation challenges: Invariance

Viewpoint variation



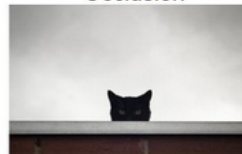
Scale variation



Deformation



Occlusion



Illumination conditions



Background clutter



Intra-class variation

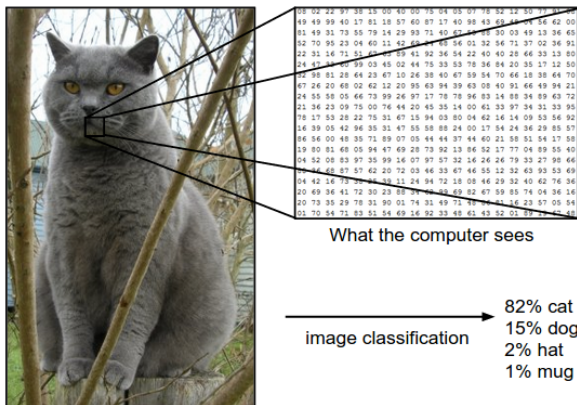


4. A CV sample task: Object Classification

Slides taken from

<http://cs231n.github.io/classification/>

4.1. Object Classification



The task in Image Classification is to predict a single label (or a distribution over labels as shown here to indicates our confidence) for a given image. Images are 3-dimensional arrays of integers from 0 to 255, of size Width x Height x 3. The 3 is due to the three color channels Red, Green, Blue.

4.2. Data Driven

Data-driven approach : it relies on first accumulating a training dataset of labeled images.



An example training set for four visual categories. In practice we may have thousands of categories and hundreds of thousands of images for each category.

4.3. The image classification pipeline

- ▶ **Input.** Our input consists of a set of N images, each labeled with one of K different classes. We refer to this data as the **training set**.
- ▶ **Learning.** Our task is to use the training set to learn what every one of the classes looks like. We refer to this step as **training a classifier**, or learning a model.
- ▶ **Evaluation.** In the end, we evaluate the quality of the classifier by asking it to **predict labels** for a new set of images that it has never seen before (**test set**). We will then compare the true labels of these images to the ones predicted by the classifier. Intuitively, we're hoping that a lot of the predictions match up with the true answers (which we call the ground truth).

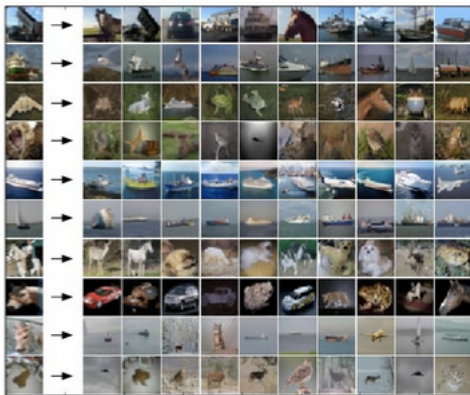
4.4. Nearest Neighbor Classifier

The nearest neighbor (NN) classifier will

1. take a test image,
2. compare it to every single one of the training images, and
3. predict the label of the closest training image.

4.5. Nearest Neighbor examples

In only about 3 out of 10 examples an image of the same class is retrieved, while in the other 7 examples this is not the case. For example, in the 8th row the nearest training image to the horse head is a red car, presumably due to the strong black background. As a result, this image of a horse would in this case be mislabeled as a car.



4.6. Image distance

The difference (or the familiar cosine similarity)

test image				training image				pixel-wise absolute value differences			
56	32	10	18	10	20	24	17	46	12	14	1
90	23	128	133	8	10	89	100	82	13	39	33
24	26	178	200	12	16	178	170	12	10	0	30
2	0	255	220	4	32	233	112	2	32	22	108

→ 456

An example of using pixel-wise differences to compare two images with L1 distance (for one color channel in this example). Two images are subtracted elementwise and then all differences are added up to a single number. If two images are identical the result will be zero. But if the images are very different the result will be large.

4.7. Evaluation

The CIFAR-10 training set of 50,000 images (5,000 images for every one of the labels), and we wish to label the remaining 10,000.

The NN Classifier based on raw pixel representation and the image distance measure above reaches 38.6 % accuracy (vs. upper bound: 94% – human).

State of the art classifier, Convolutional Neural Network reaches 95%.

4.8. K-Nearest Neighbor Classifier

You may have noticed that it is strange to only use the label of the nearest image when we wish to make a prediction. Indeed, it is almost always the case that one can do better by using what's called a k-Nearest Neighbor Classifier. The idea is very simple: instead of finding the single closest image in the training set, we will find the top k closest images, and have them vote on the label of the test image.

Which is the best k?

K is an hyperparameter. There are others too.

4.9. Validation dataset vs Test dataset

They can be trained/learned.

- ▶ Training data-set: to train the classifier.
- ▶ Train data-set (or development dataset or validation dataset): to tune the parameters.
- ▶ Test data-set. To test the classifier.

4.10. First problem: the classifier

NN Classifier: pro and contra.

- ▶ the classifier takes no time to train, since all that is required is to store and possibly index the training data.
- ▶ However, we pay that computational cost at test time, since classifying a test example requires a comparison to every single training example.
- ▶ This is backwards, since in practice we often care about the test time efficiency much more than the efficiency at training time.

In CV it's better to use other classifiers.

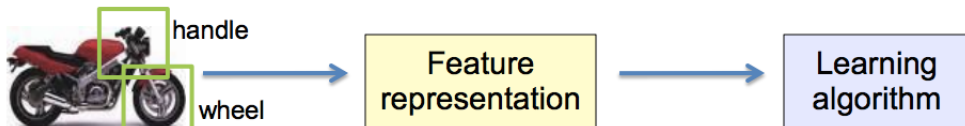
State of the art: Deep neural networks are very expensive to train, but once the training is finished it is very cheap to classify a new test example. This mode of operation is much more desirable in practice.

4.11. Second problem: the Raw Pixel representation

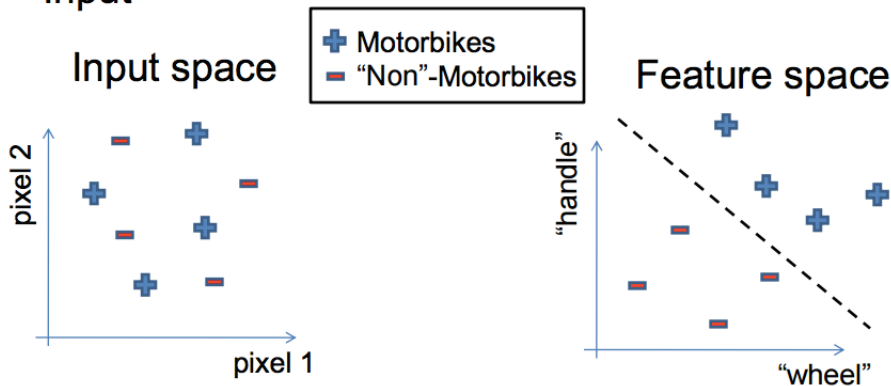
- ▶ Using the NN classifier over the raw pixel representation images that are nearby each other are much more a function of the general color distribution of the images, or the type of background rather than their semantic identity.
- ▶ For example, a dog can be seen very near a frog since both happen to be on white background.
- ▶ Ideally we would like images of all of the 10 classes to form their own clusters, so that images of the same class are nearby to each other regardless of irrelevant characteristics and variations (such as the background).

To get this property we will have to go beyond raw pixels.

5. Representation Problem: From pixel to feature



Input



5.1. Two methods

- ▶ Bag of visual words (Sivic and Zisserman, 2003)
- ▶ Convolutional neural network (LeCun et al., 1998)

5.2. Bag of Visual Words: Pipeline

1. Low-level feature extraction

- ▶ Identify keypoints
- ▶ Get local feature descriptors: change of intensity of each point is computed (“gradients”)

2. Bag of Visual words

- ▶ Cluster local descriptors
- ▶ Quantize

5.3. Low-level Features extraction

Keypoints detectors To locate interesting points/content, various kinds of low-level features detectors exist:

- ▶ edge detection: the lines we would draw – encode shape info
- ▶ corner detection
- ▶ blob detection

Local description The identified interesting points are then described: clustered into regions and transformed into vectors representing the region. Several local descriptors exist, e.g:

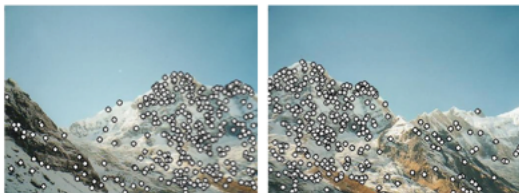
- ▶ SIFT: Scale-invariant feature transform (Lowe '99) – edge based features.
- ▶ Textons (Leung and Malik '01)
- ▶ HoG (Dalal and Triggs '05)

The low-level features can capture eg. Color, Texture, Shape,

(Note on Image gradients: <http://www.cs.umd.edu/~djacobs/CMSC426/ImageGradients.pdf>)

Software for feature extraction MATLAB and others.

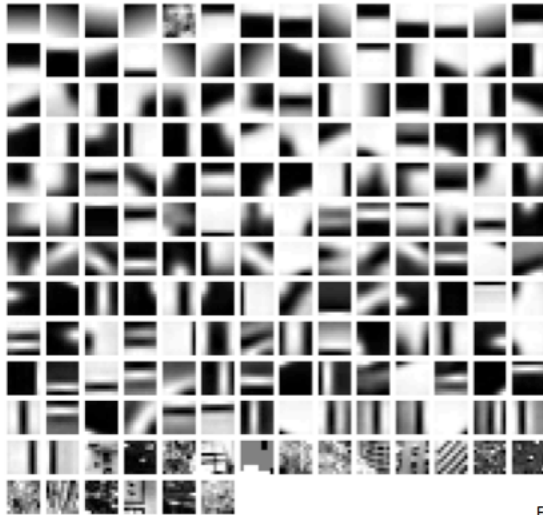
5.4. Characteristics of good low-features



- **Repeatability**
 - The same feature can be found in several images despite geometric and photometric transformations
- **Saliency**
 - Each feature has a distinctive description
- **Compactness and efficiency**
 - Many fewer features than image pixels
- **Locality**
 - A feature occupies a relatively small area of the image; robust to clutter and occlusion

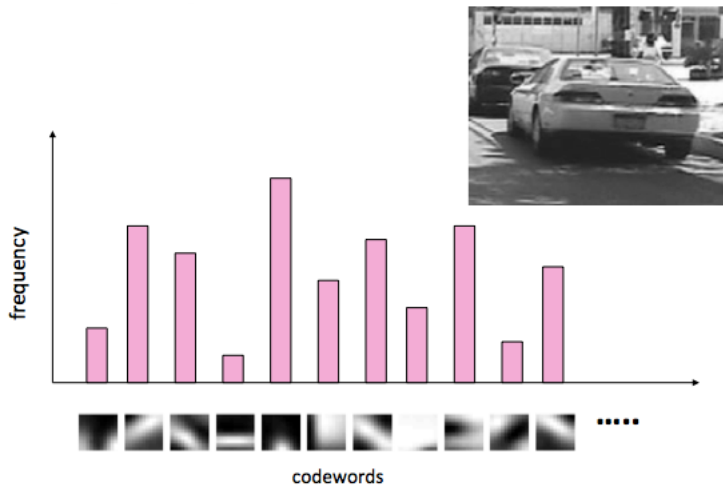
source: [Svetlana Lazebnik](#)

5.5. Example visual vocabulary



151
Fei-Fei et al. 2005

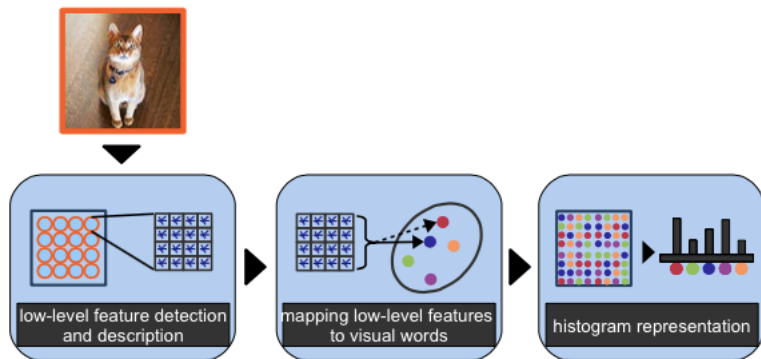
5.6. Image Representation



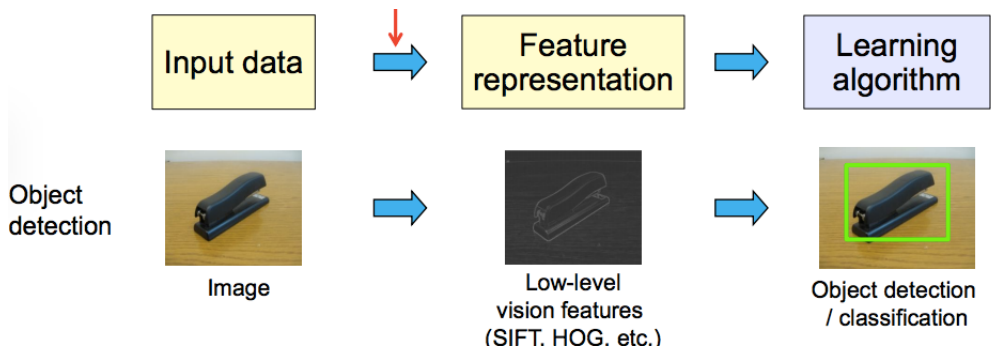
153

source: Svetlana Lazebnik

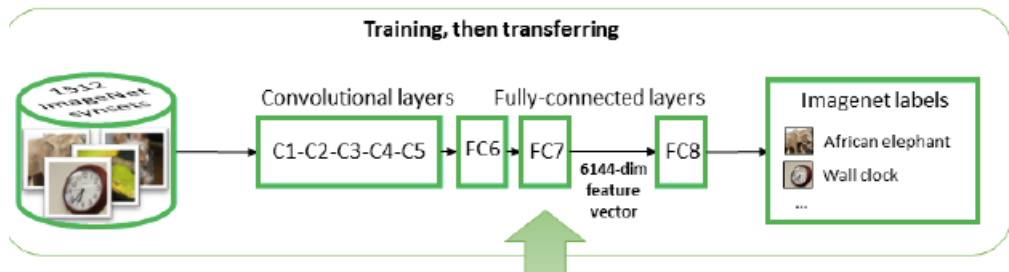
5.7. Summary: Images representation pipeline



5.8. From hand-crafted feature to feature learning

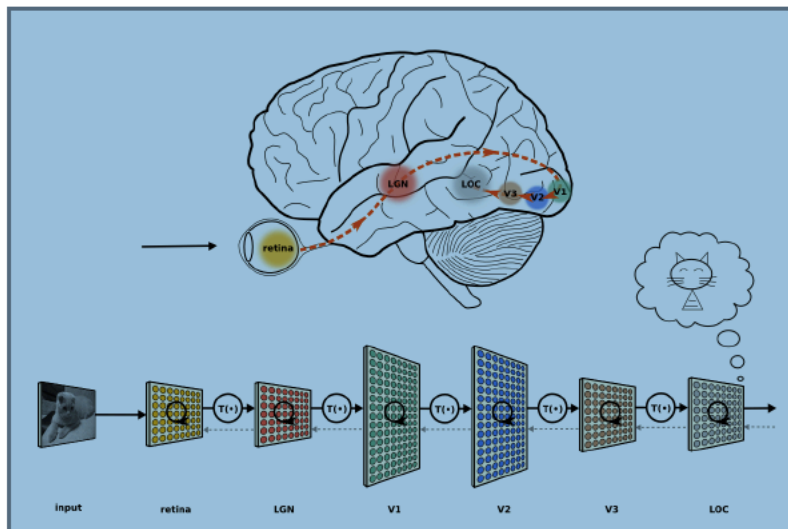


5.9. Convolutional Neural Network: transfer



- 1 Train a **convolutional neural network** on a vision task
e.g. AlexNet [Krizhevsky et al., 2012b] on ILSVRC
[Russakovsky et al., 2015]
- 2 Do a **forward pass** given an image input
- 3 **Transfer** one or more layers (e.g. FC₇, or CONV₅)

5.10. Inspiration



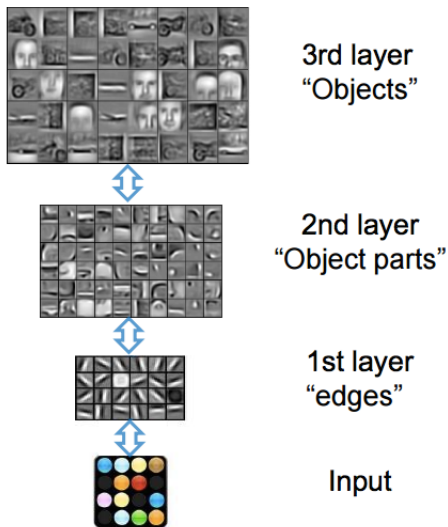
(c) Jonas Kubilius



5.11. Hierarchy of features

Deep Learning

- Deep architectures can be representationally efficient.
- Natural progression from low level to high level structures.
- Can share the lower-level representations for multiple tasks.



6. Classifier problem

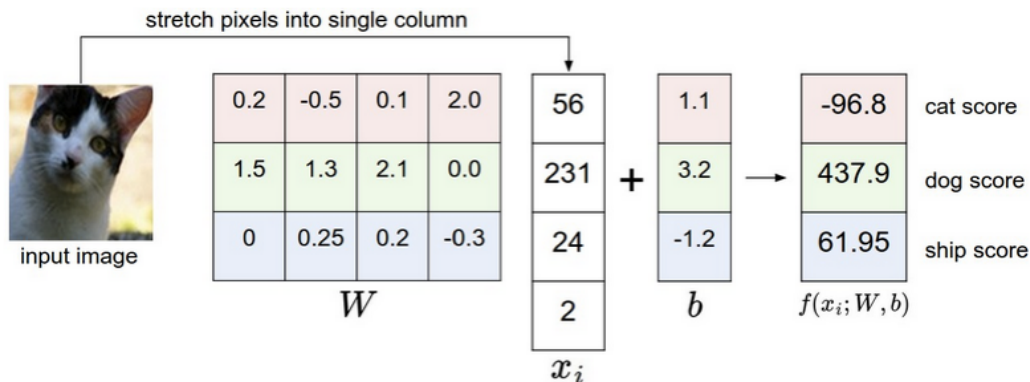
Before we saw that kNN classifier are not suitable for CV tasks since:

- ▶ The classifier must remember all of the training data and store it for future comparisons with the test data. This is space inefficient because datasets may easily be gigabytes in size.
- ▶ Classifying a test image is expensive since it requires a comparison to all training images.

People use a **parametric approach**, since

- ▶ we learn the parameters we can discard the training data.
- ▶ Additionally, the prediction for a new test image is fast since it requires a single mathematical operation, not an exhaustive comparison to every single training example.

6.1. Score and Loss functions: example



An example of mapping an image to class scores. For the sake of visualization, we assume the image only has 4 pixels and that we have 3 classes (red, blue, green class). We stretch the image pixels into a column and perform matrix multiplication to get the scores for each class. Note that this particular set of weights W is not good at all: the weights assign our cat image a very low cat score. In particular, this set of weights seems convinced that it's looking at a dog.

6.2. Score and Loss functions

- ▶ a **score function** that maps the raw data to class scores, and
- ▶ a **loss function** (alternative names: cost function or the objective) that quantifies the agreement between the predicted scores and the ground truth labels

This can be casted as an **optimization problem** in which the loss function is minimized with respect to the parameters (weights) of the score function.

6.3. Score function: Linear Classifier

Given Images x of dimensions D to be classified against K classes: $x : D \times 1$, $W : K \times D$, $b : K \times 1$, we can use the score function:

$$f(x_i, W, b) = Wx_i + b$$

where x_i is an image, W is a matrix whose values are called **weights** and b is called a **bias vector** – it influences the result without interacting with the actual data. The score function is based on a linear combination of the matrices weights and the input (multiplication) – hence linear classifier.

Each row of W

- ▶ is a classifier of a specific category.
- ▶ can be also seen as a “prototype” of the class, and the inner product as a way to compare the prototype with the test image.

(The W and b parameters are usually put together by extending W with an extra dimension (the b values), and the image with a dimension with 1.)

6.4. Loss Function: Super Vector Machine

The SVM **loss** is set up so that the SVM "wants" the correct class for each image to have a score higher than the incorrect classes by some fixed margin Δ .

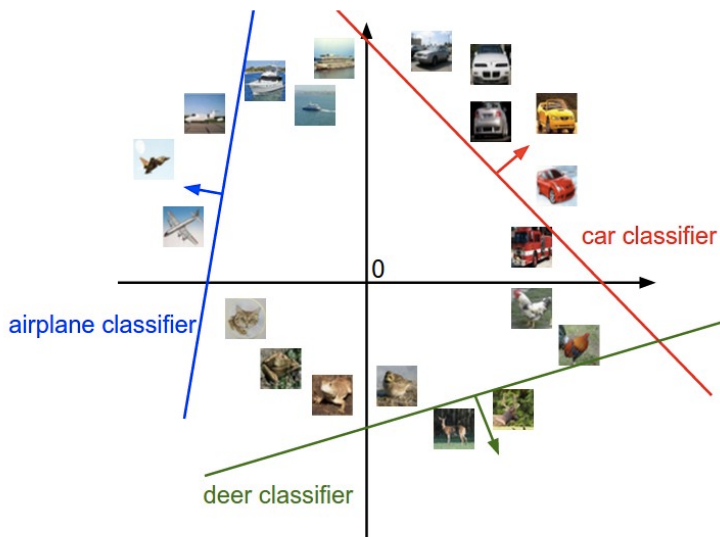
The loss function quantifies our unhappiness with predictions on the training set



The Multiclass Support Vector Machine "wants" the score of the correct class to be higher than all other scores by at least a margin of Δ . If any class has a score inside the red region (or higher), then there will be accumulated loss. Otherwise the loss will be zero. Our objective will be to find the weights that will simultaneously satisfy this constraint for all examples in the training data and give a total loss that is as low as possible.

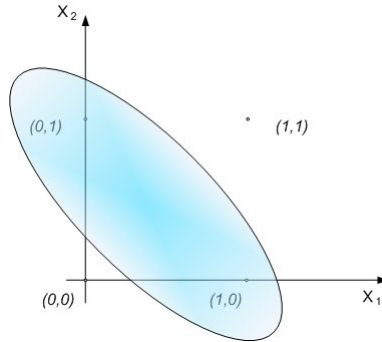
A loss function can be optimized with iterative refinement, where we start with a random set of weights and refine them step by step until the **loss is minimized**.

6.5. Linear Classifier: cartoon representation



Cartoon representation of the image space, where each image is a single point, and three classifiers are visualized. Using the example of the car classifier (in red), the red line shows all points in the space that get a score of zero for the car class. The red arrow shows the direction of increase, so all points to the right of the red line have positive (and linearly increasing) scores, and all points to the left have a negative (and linearly decreasing) scores.

6.6. non linear problems



XOR (exclusive or) function cannot be implemented by a linear classifier.

7. Applications: CV exploits NLP and vice-versa

- ▶ Computer Visions tasks:
 - ▷ Recognition: object, scene, events, action, people ...
 - ▷ Image Annotation
 - ▷ Image Retrieval
 - ▷ Image Generation
- ▶ NLP tasks:
 - ▷ Lexical Preferences
 - ▷ Machine Translation
 - ▷ Question Answering,
 - ▷ Information Retrieval,
 - ▷ Textual Entailment

Use a Multi-modal knowledge to improve CV and/or NLP tasks.

8. Computer Vision exploits language

Tasks : Old one, e.g, Object recognition, New one: e.g, Caption generation

Language sources More and more CV people are looking into ways to exploit prior knowlege obtained from language models built from

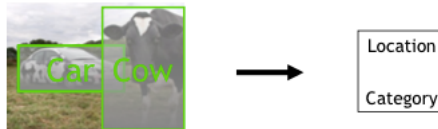
- ▶ image tags
- ▶ image captions
- ▶ corpora

8.1. Traditional CV task: Object recognition

Image classification: assigning a label to the image.



Object localization: define the location and the category.



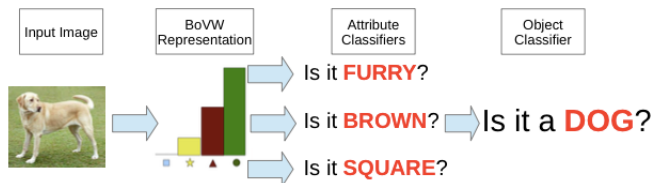
Similarly, scene recognition.

8.2. Object recognition: methods

Traditional pipe-line:

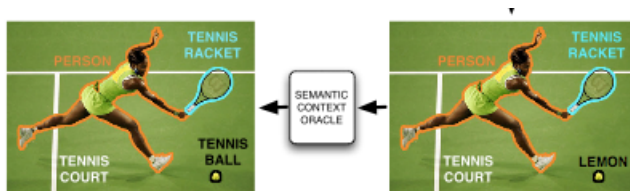


Later (before the deep-learning revolution) proposed pipe-line:



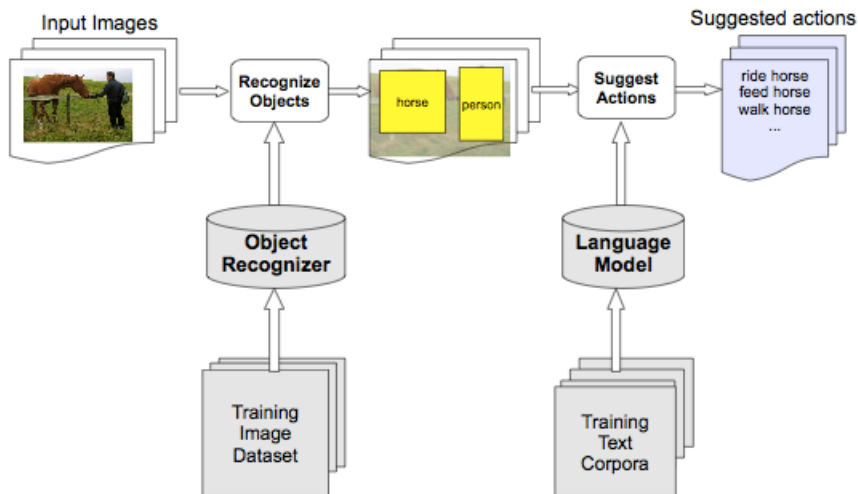
8.3. Corpora as KB source: Object recognition

A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, S. Belongie (ICCV 2007)
Objects in Context.



Not a Lemon, it's more probable a Tennis Ball. Info come from a KB (word similarity list, extracted from internet – Google Sets).

8.4. Corpora as KB source: Action recognition

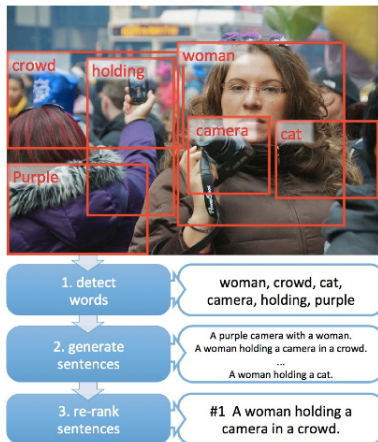


Le Dieu Thu's PhD Thesis (DISI)

8.5. Caption generation

Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, Mitchell, Platt, Zitnick, Zweig, (CVPR 2014)

From captions to visual concepts and back



8.6. Caption generation: biblio

- ▶ X. Chen and C. L. Zitnick, Learning a Recurrent Visual Representation for Image Caption Generation (2014). GOOD. RNN, bi-directional.
- ▶ BabyTalk pipeline.
- ▶ R. Socher and L. Fei-Fei. (CVPR 2010)
Connecting Modalities: Semi-supervised Segmentation and Annotation of Images Using Unaligned Text Corpora.
- ▶ R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, A. Y. Ng. (NIPS 2013)
Grounded Compositional Semantics for Finding and Describing Images with Sentences.
- ▶ J. Thmason, S. Venugopalan, S. Guardarrama, K. Saenko, R. Mooney. (COLING 2014)
Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild.

- ▶ A. Karpathy and Li Fei-Fei. (CVPR 2015)
Deep Visual-Semantic Alignments for Generating Image Descriptions

9. Visual Question Answering



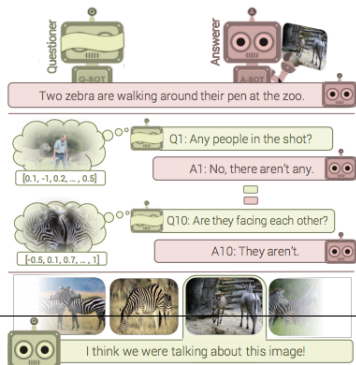
VQA 2015

What colour is the moustache made of?

Who is wearing glasses?
man woman



VQA2 2017



Questioner Answerer

Two zebra are walking around their pen at the zoo.

Q1: Any people in the shot?
A1: No, there aren't any.

Q10: Are they facing each other?
A10: They aren't.

I think we were talking about this image!

IVQA new

More at: <http://www.visualqa.org/>

10. NLP exploits vision

Examples:

- ▶ Selectional Preference
- ▶ Translation

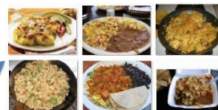
10.1. Lexical Preference

S. Bergsma and R. Goebel. (RANLP 2011).

Using visual information to predict lexical preference.

Is the argument
migas a
plausible object
for the
predicate *eat*?

Can you *eat*
“*migas*”?



Can you *eat*
“*carillon*”?



Can you *eat*
“*mamey*”?



5

Difference: concrete (visual space helps) vs. abstract nouns (visual space does not.)

10.2. Translation

S. Bergsma, B. Van Durme, (IJCAI 2011)

Learning Bilingual Lexicons using the Visual Similarity of Labeled Web Images,

Images for "candle" (English)

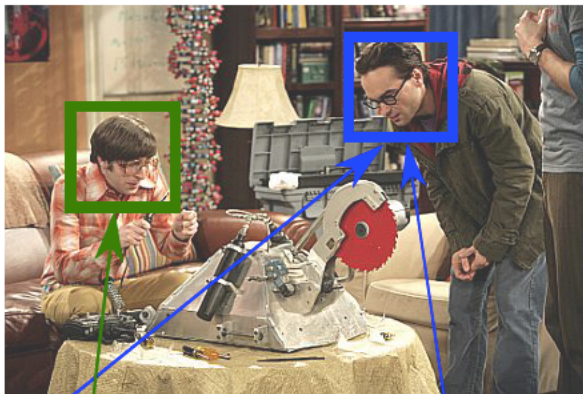


Images for "vela" (Spanish)

Figure 1: Matching words through their images: Images retrieved from the web for the English word *candle* (top) and the Spanish word *vela* (bottom). The matching between detected SIFT keypoints is shown for a pair of images.

10.3. Co-reference Resolution

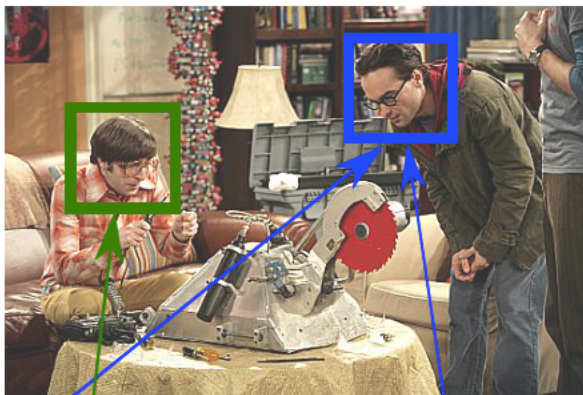
Ramanathan et al. 2014



Leonard looks at the robot, while the only
engineer in the room fixes it. He is amused.

10.4. Co-reference Resolution

Ramanathan et al. 2014



Leonard looks at the robot, while the only engineer in the room fixes it. He is amused.

11. Summary: CV and NLP

No Integration of info. One in support of the other as KB?

12. Foundational: Grounding

Multimodal knowledge

A sheep

- ▶ McRae et al 2005's norms: **is white, has wool, has 4 legs, ...**
- ▶ Text-generated description of Baroni et al 2010: **needs a shepherd, might suffer of scapie, grazes, in a farm ...**

Kelly et al 2010: use large corpora, weak supervision, lexico-syntactic patterns, achieve max 24% precision 48% recall at guessing McRae-subject-generated properties.

We acquire knowledge from several modalities, not only language.



Current corpus based models lack grounding on other modalities, e.g. vision.

13. Foundational: Reference



14. Data Set

Several annotated image datasets exists.

14.1. CIFAR

<http://www.cs.toronto.edu/~kriz/cifar.html>

The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

The classes are completely mutually exclusive: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck.

The test batch contains exactly 1000 randomly-selected images from each class. The training batches contain the remaining images in random order,

The CIFAR-100 consists of 100 classes with 600 image each. Info is given about classes (eg. beaver, dolphin, otter, seal, whale) and superclasses (e.g. aquatic mammals)

14.2. ImageNet

<http://www.image-net.org/>

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node.

Shared Task organized each year since 2010: Large Scale Visual Recognition Challenge

Tasks: Detection, Classification, Localization.

14.3. VisA

<http://homepages.inf.ed.ac.uk/s1151656/resources.html>

This dataset contains visual attribute annotations for over 500 concrete (animate and inanimate) concepts.

All concepts are represented in ImageNet and the feature production norms of McRae et al. (2005). Each concept is annotated with visual attributes based on a taxonomy of 636 attributes. See Silberer et al. (2013) for details.

14.4. SUN

<http://groups.csail.mit.edu/vision/SUN/>

A comprehensive collection of annotated images covering a large variety of environmental scenes, places and the objects within.

To build the core of the dataset, the authors counted all the entries that corresponded to names of scenes, places and environments (any concrete noun which could reasonably complete the phrase I am in a place, or Lets go to the place), using WordNet English dictionary.

Once they established a vocabulary for scenes, they collected images belonging to each scene category using online image search engines by quering for each scene category term, and annotate the objects in the images manually.

15. Dataset for sentence-based image description

Credits: Julia Julia Hockenmaier. (EACL Tutorial)

Using captioned images from the web (news, photo-sharing sites):

- ▶ Advantage: Size, 'natural' captions
- ▶ Disadvantage: Online captions may not describe images
- ▶ Example: SBU Captioned Photo dataset; BBC dataset

Using images with purposely created captions:

- ▶ Advantage: Sentence describe the images
- ▶ Disadvantage: Smaller size, 'unnatural'.
- ▶ Examples: IAPR=TC, Illinois Pascal dataset; Flickr 8K, etc.

15.1. Online Caption?

News sites often use images
just to embellish their stories

Drinking over recommended limit 'raises cancer risk'

COMMENTS (348)

Drinking more than a pint of beer a day can substantially increase the risk of some cancers, research suggests.

The Europe-wide study of 363,988 people reported in the British Medical Journal found one in 10 of all cancers in men and one in 33 in women were caused by past or current alcohol intake.

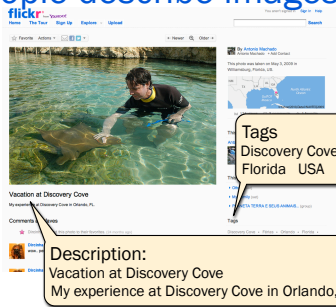
More than 18% of alcohol-related cancers in men and about 4% in women were linked to



Many people do not know that drinking alcohol can increase their cancer risk.

15.2. Photo-sharing?

On photo-sharing sites,
people describe images...



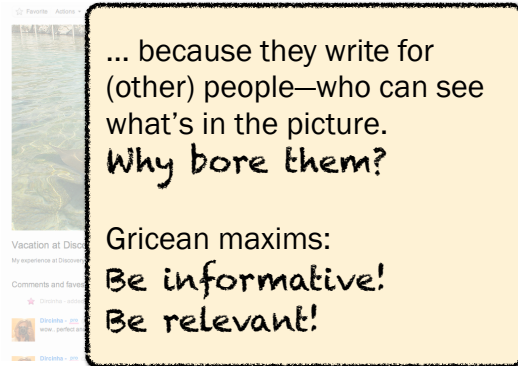
The image shows a screenshot of a Flickr photo page. The main photo is a man in a black shirt and yellow life vest kneeling in shallow, clear water, reaching out to touch a large sea turtle. The page includes a title, location, and tags. Two callout boxes are overlaid on the page:

- Tags:** Discovery Cove Férias Orlando Florida USA EUA Vacations
- Description:** Vacation at Discovery Cove
My experience at Discovery Cove in Orlando, FL

21

15.3. Photo-sharing?

... but they don't provide
conceptual descriptions...



15.4. IAPR-TC12 data set



Horse-Riding at the pampas

six people are riding on brown and white horses in a green, flat meadow in the foreground; cows behind them; white and grey clouds in a light blue sky in the background;

Buenos Aires, Argentina
9 December 2004



Panoramic View of the Iguazu Waterfalls

a cascading waterfall in the middle of the jungle;
front view with pool of dirty water in the foreground;
this picture was taken from the Brazilian side;

Foz do Iguaçu, Brazil
March 2002

20,000 manually annotated and segmented images
Grubinger et al. 2006; Escalante et al. 2010

15.5. ILLINOIS PASCAL data set



A grounded passage plane in a terminal.
An Air Pacific airplane sitting on the tarmac.
Large white commercial airliner parked on runway.
The back and right side of a parked passenger jet.
The passenger plane is sitting at the airport.




A hand holding bird seed and a small bird.
A person holding a small bluebird.
A person holds a bird and seeds.
A small bird is sitting on a person's hand that has bird seed in it.
A small black, white, and brown bird perched on and eating out of a man's hand.

1,000 images from the PASCAL VOC 2008 challenge
(20 object categories) with 5 crowdsourced captions
Rashtchian et al. 2010

15.6. Crowdsource

Image 1 / 10:



Please describe the image in one complete but simple sentence.

[Next →](#)

Instructions:

Describe the objects and actions; Use adjectives; be brief
5 captions per image

15.7. Crowdsourced results



Four basketball players in action.
Young men playing basketball in a competition.
Four men playing basketball, two from each team.
Two boys in green and white uniforms play
basketball with two boys in blue and white uniforms.
A player from the white and green highschool team
dribbles down court defended by a player from the other
team.

15.8. LabelMe

<http://labelme.csail.mit.edu/Release3.0/>

The goal of LabelMe is to provide an online annotation tool to build image databases for computer vision research.

16. Demos TBD

Image Caption generation: (<http://deeplearning.cs.toronto.edu/i2t>)

More at: <http://deeplearning.net/demos/>

17. Softwares

Some user-friendly ones:

- ▶ SIFT etc: <http://www.vlfeat.org/>
- ▶ CNN features: <http://www.vlfeat.org/matconvnet/>
- ▶ CNN features from another group: <http://caffe.berkeleyvision.org/>

18. Language and Vision Research Groups

- ▶ Stanford Vision Lab – Le Fei Fei <http://vision.stanford.edu/>
- ▶ MIT: Antonio Torralba <http://web.mit.edu/torralba/www/>
- ▶ University of North Carolina – Tamara Berg <http://www.tamaraberg.com/>
- ▶ Virginia University – Devi Parikh <https://filebox.ece.vt.edu/~parikh/CVL.html>
- ▶ CLIC <http://clic.cimec.unitn.it/lavi/> – Us.
- ▶ Center for Cognition, Vision, and Learning – Alan L. Yuille <http://ccvl.stat.ucla.edu/>
- ▶ Edinburgh University (M. Lapata, F. Keller)
- ▶ Cognitive Systems Research Institute <http://www.csri.gr/en/>
- ▶ University of Leuven <http://hci.cs.kuleuven.be/>

- ▶ More on the iV&L Net Cost Action http://www.cost.eu/COST_Actions/ict/Actions/IC1307

19. Language and Vision

- ▶ CVPR: language and vision workshop. This year 2nd edition.
- ▶ ACL/EACL: language and vision workshop: 2017 7th edition. ACL'17 topic of the CfP.

20. Other Useful Links

<http://nlp.cs.illinois.edu/HockenmaierGroup/EACLTutorial2014/index.html>

http://info.usherbrooke.ca/hlarochelle/neural_networks/content.html

<http://www.iro.umontreal.ca/~bengioy/dlbook/>

<http://www.vlfeat.org/matconvnet/matconvnet-manual.pdf>

Vision and Language Summer Schools: 2nd edition 2016 (Malta)

Blog posts: <http://colah.github.io/>

Multimodal Learning and Reasoning, Desmond Elliott, Douwe Kielay, and Angeliki Lazaridou (Tutorial at ACL 2016) http://acl2016.org/index.php?article_id=59