Multimodal CL and NLP: Combining Language and Vision for (Computational) Semantics

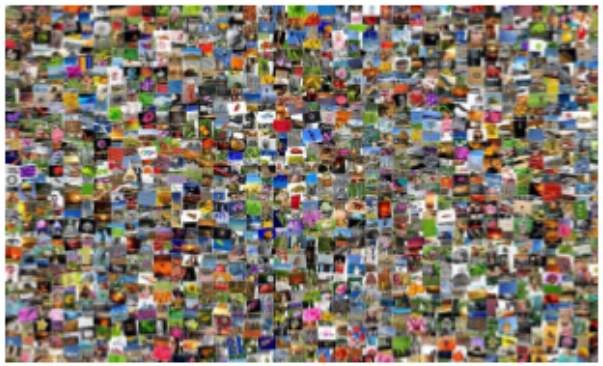
Carina Silberer

IMS, University of Stuttgart

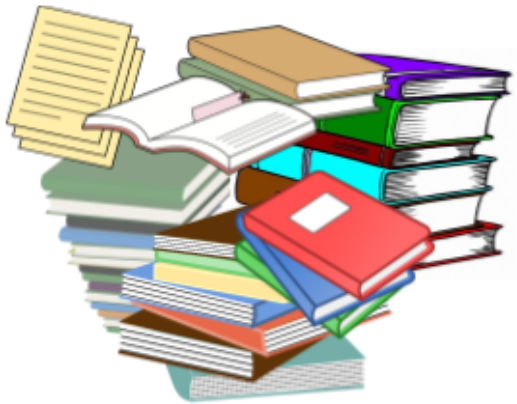
CL Fall School 2024, Passau
16 - 20 September 2024

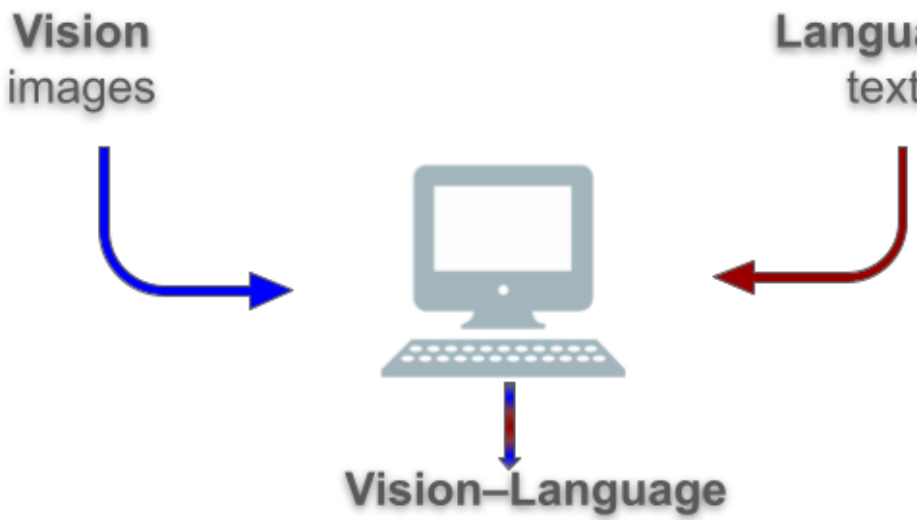
Outline

Multimodal (VL) Representations



https://pixabay.com/photos/photo-collection-pictures-photos-382018



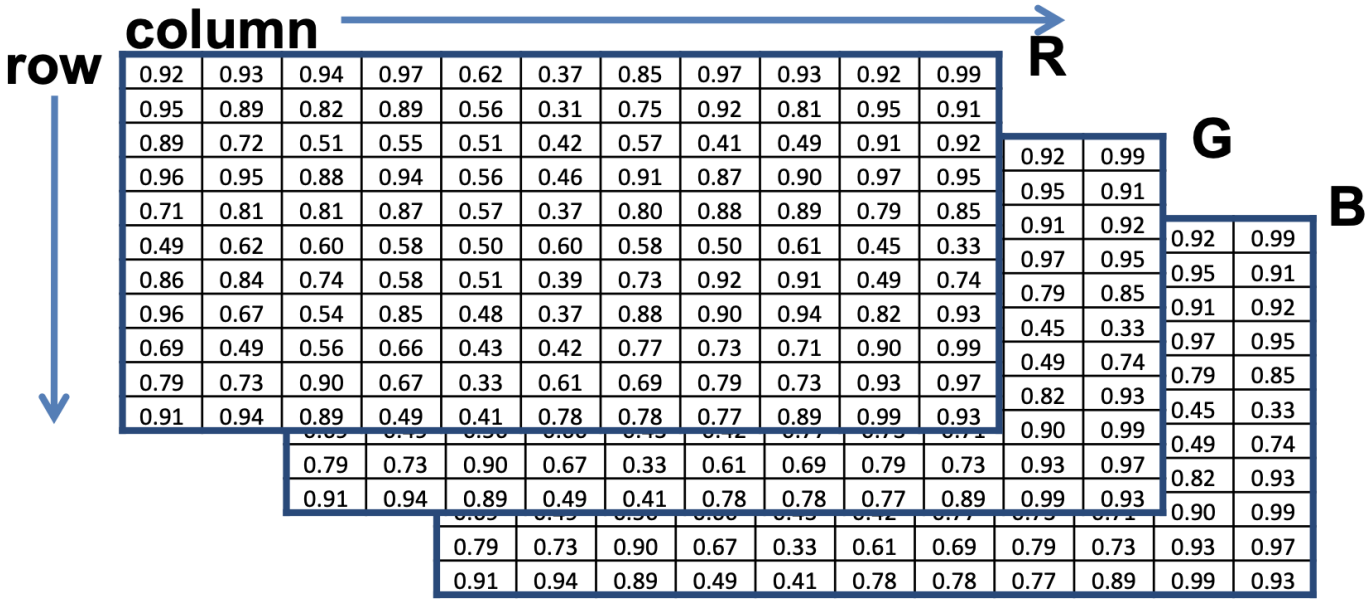


Outline

How are images encoded?

$N \times M \times 3$ Matrix

- 3 *channels* for encoding image in certain colour space
- E.g., RGB (Red, Green, Blue), HSV (Hue, Saturation, Value), L*a*b, ...
- N rows, M columns



How are images encoded?

$N \times M \times 3$ Matrix

- 3 *channels* for encoding image in certain colour space
- E.g., RGB (Red, Green, Blue), HSV (Hue, Saturation, Value), L*a*b, ...
- N rows, M columns

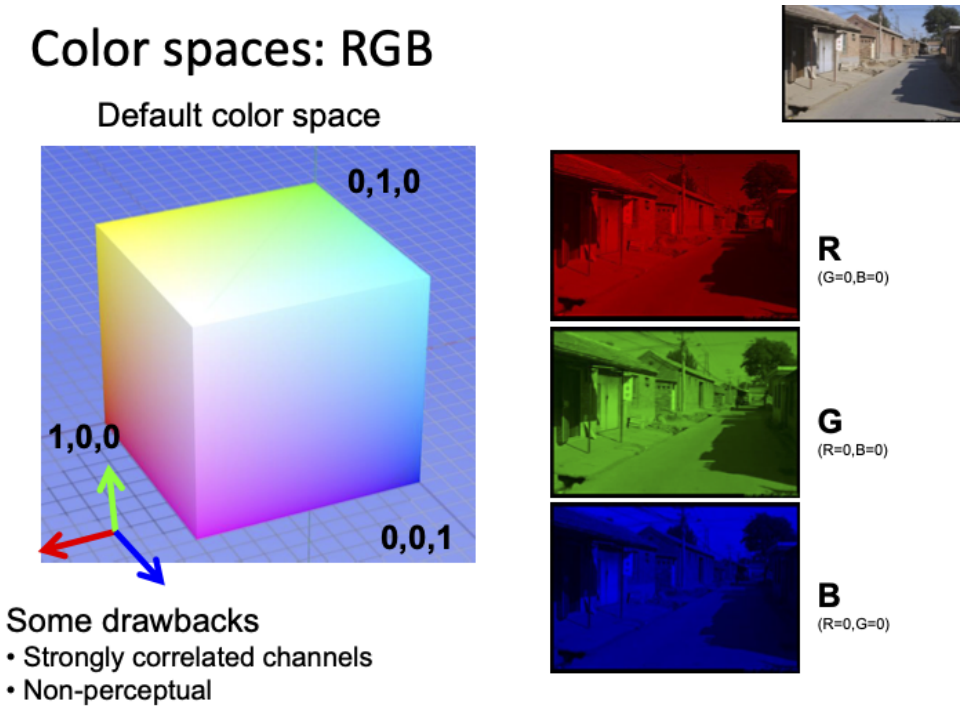


Image from: http://en.wikipedia.org/wiki/File:RGB_color_solid_cube.png

Visual Representations

Development of Visual Feature Representations

early	*handcrafted* feature vectors
2012+	Convolutional Neural Networks
2019+	Visual Transformers
Dataset	Huge manually labeled datasets required
2021+	Visual-Linguistic Models (transformer-based) learn in unsupervised way from image-text pairs

Visual Representations

Development of Visual Feature Representations

early	*handcrafted* feature vectors
2012+	Convolutional Neural Networks
2019+	Visual Transformers
Dataset	Huge manually labeled datasets required
2021+	**Visual-Linguistic Models (transformer-based) learn in unsupervised way from image-text pairs**

Visually Grounded Semantic Models: Overview

Visual-Linguistic Representation Learning: Visual Representations

- Early Works: Linguistic Symbols
  - Human-elicited Feature Norms [McRae et al., 2005]
    [Andrews et al., 2009, Silberer and Lapata, 2012]
    [Johns and Jones, 2012, Roller and Im Walde, 2013]
  - Image Tags [Bruni et al., 2012, Hill et al., 2014]
- Pre-CNN Works: Image features based on visual descriptors + visual words (human-defined filters)
  [Bruni et al., 2012, Silberer et al., 2013, Kiela et al., 2014]
- Nowadays standard: Image features extracted from convolutional NNs or transformers (automatically learnt filters)
  [Kiela and Bottou, 2014, Lazaridou et al., 2015, Silberer et al., 2017]

Visually Grounded Semantic Models: Early Models

Visual-Linguistic Static Representations Learning:

Visual Representations

Visually Grounded Semantic Models: Early Models

Perceptual Representations: Feature Norms [McRae et al., 2005]

- Human-elicited, normed properties of concepts

Features	table	dog	apple
has_4_legs	.28	.60	0
used_for_eating	.50	0	0
a_pet	0	.40	0
is_brown	0	0	0
is_crunchy	0	0	.58
is_round	.22	0	.42
has_fangs	0	0	0

Table 1: Feature norms for the nouns *table*, *dog*, and *apple* shown as a distribution.

source: [Silberer and Lapata, 2012]

Visually Grounded Semantic Models: Early Models

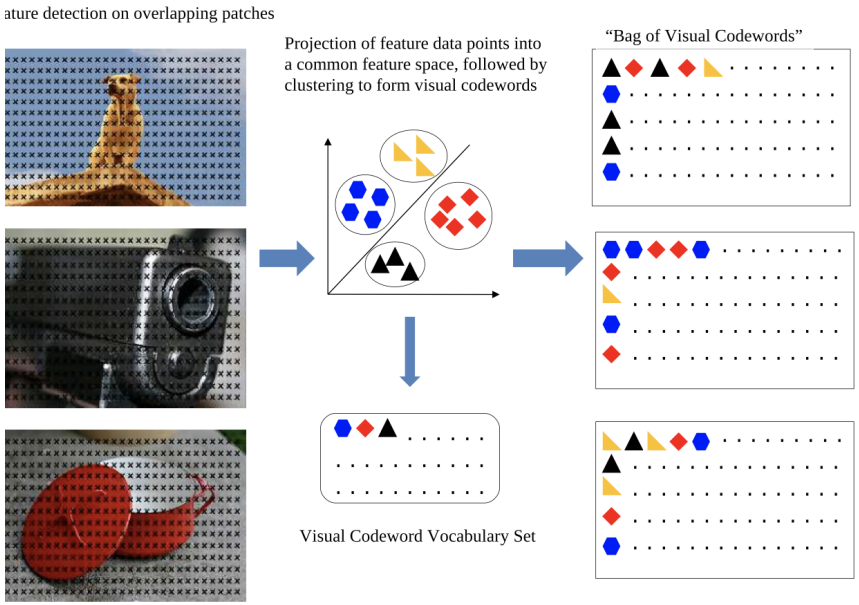
Visual Representations: Bag of Visual Words (BoVW)

- Visual features are extracted from image collection (*feature detection and description*; descriptors: numerical vectors)
- Cluster features into *visual codewords* (centers of clusters)

$\Rightarrow$ Visual vocabulary

- Represent each image as a bag of visual codewords

$\Leftrightarrow$ Distributional method to represent images

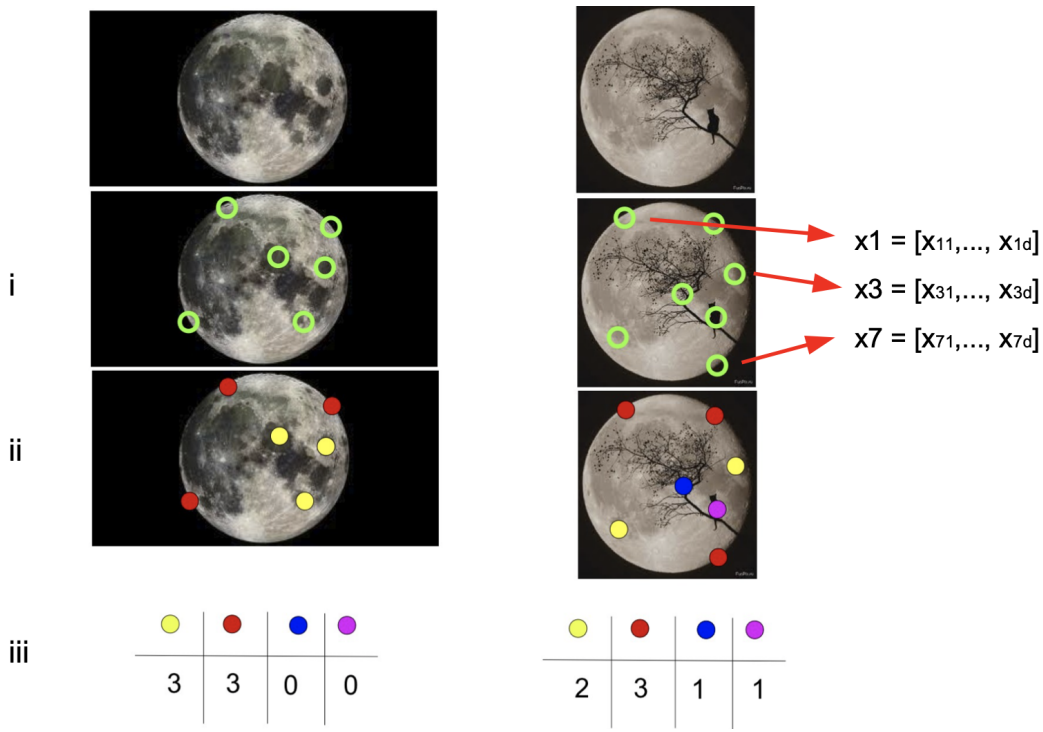


gure 1: An illustration of the process of generat-

Visually Grounded Semantic Models: Early Models

Visual Representations of **Words**: BoVW Approach

- Feature extraction and description
- *Visual codeword* generation
- BoVW approach: Distributional representation of images



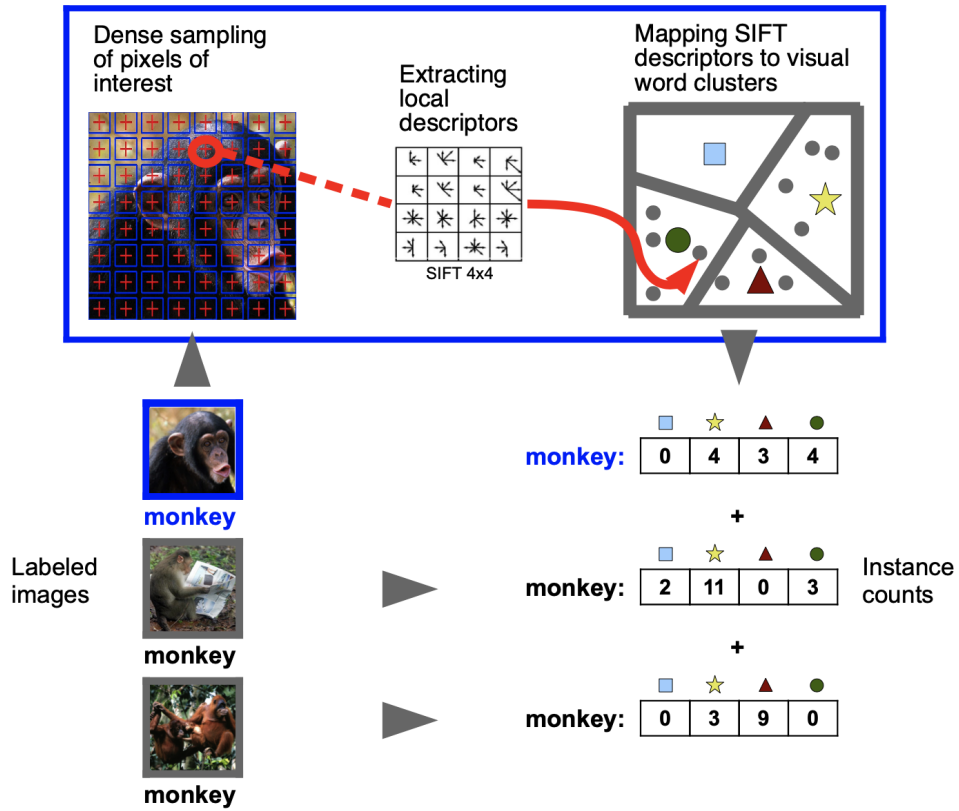
source: [Bruni et al., 2014]

Visually Grounded Semantic Models: Early Models

Visual Representations of **Words**: BoVW Approach

- Feature extraction and description
- *Visual codeword* generation
- BoVW approach: Distributional representation of images

$\Rightarrow$ Distributional representation of words associated with images



Visually Grounded Semantic Models: Early Models

Visual Representations: Visual Global Features

- Histograms of colours (e.g., HSV,RGB,LAB)
- Textures (Gabor and Haar wavelets; Makadia et al., 2018)

Visual Representations: Visual Local Features (Fei-Fei and Perona, 2005)

- Key points
- Robust: SIFT (scale-invariant feature transform; Lowe, 2004)

Visually Grounded Semantic Models: Early Models

Visual Representations: Visual Attribute Predictions

[Silberer et al., 2013]

- Represent each image by means of extracted visual feature vectors (local features, BoVW)
- Train attribute classifiers: given image, predict the visual attributes occurring in image
- Apply attribute classifiers to new images, each associated with word

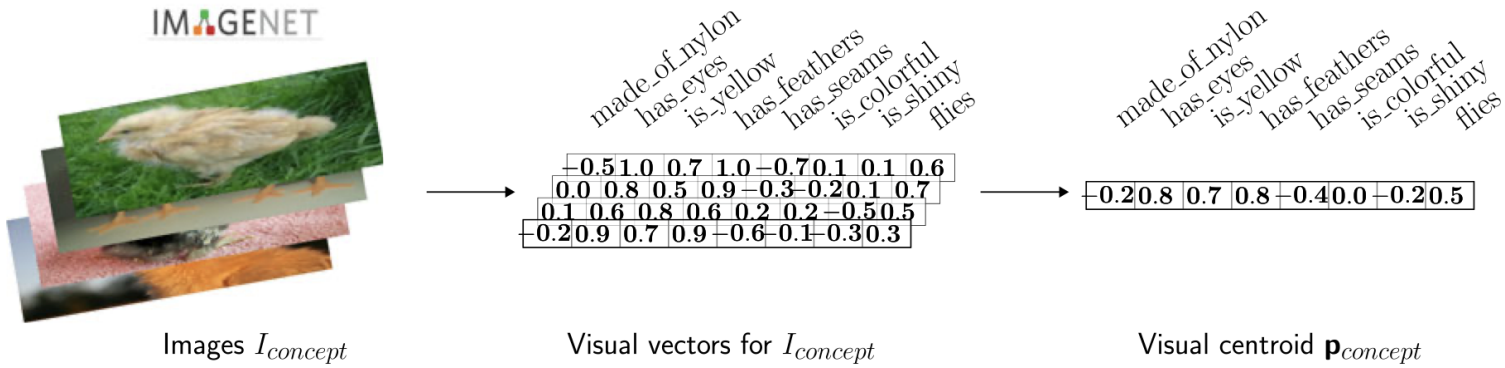


Fig. 5. Visual representation for concept *chick*. Attribute classifiers predict attributes for example images depicting *chicks*. These prediction scores are then converted into vectors (first arrow). To compute a single visual attribute vector for a concept, all vectors are aggregated into $\mathbf{p}_{chick}$.

source: [Silberer et al., 2013]

Visually Grounded Semantic Models: Early Models

Visual Representations: Visual Attribute Predictions

[Silberer et al., 2013]

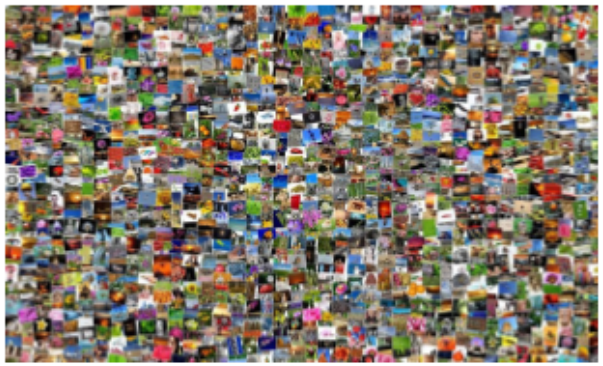
- Represent each image by means of extracted visual feature vectors (local features, BoVW)
- Train attribute classifiers: given image, predict the visual attributes occurring in image
- Apply attribute classifiers to new images, each associated with word

⇒ Attribute-based representation of words associated with images

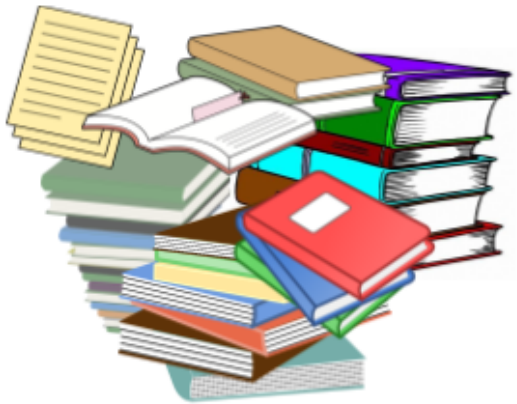
		eats_seeds	has_beak	has_claws	has_handlebar	has_wheels	has_wings	is_yellow	made_of_wood
Visual	*canary*	0.05	0.24	0.15	0.00	–0.10	0.19	0.34	0.00
	trolley	0.00	0.00	0.00	0.30	0.32	0.00	0.00	0.25

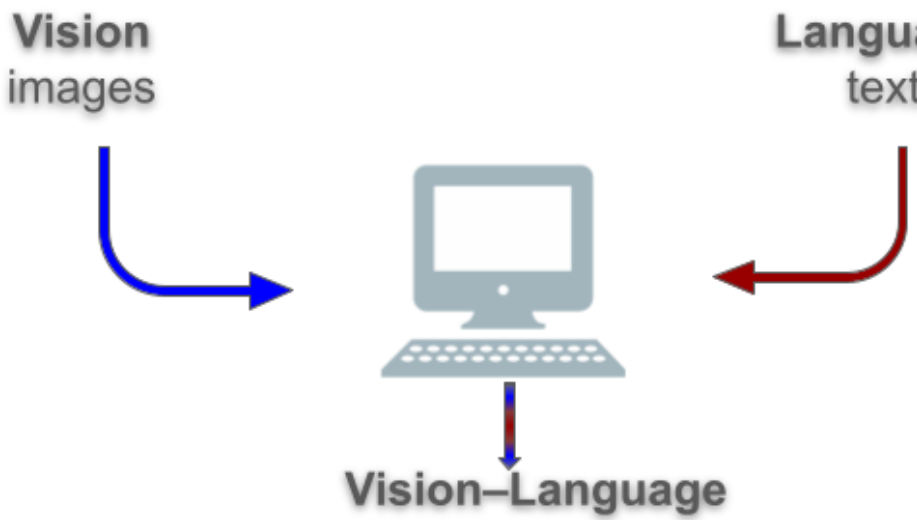
source: [Silberer et al., 2013]

Multimodal (VL) Representations



https://pixabay.com/photos/photo-collection-pictures-photos-382018





How to obtain useful image features?

Pre-Deep Learning Era

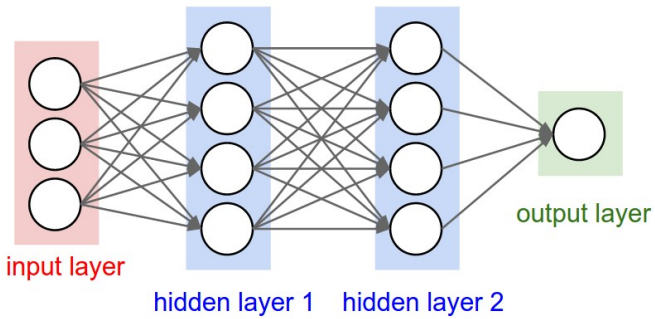
- Use of Handcrafted images filters / descriptors to extract features from images
- Use of simple machine learning approaches
  e.g., k-nearest neighbours, support vector machines, ...

How to obtain useful image features?

Pre-Deep Learning Era

- Use of Handcrafted images filters / descriptors to extract features from images
- Use of simple machine learning approaches
  e.g., k-nearest neighbours, support vector machines, ...

Regular NNs vs. Convolutional NNs (Fukushima, 1988)



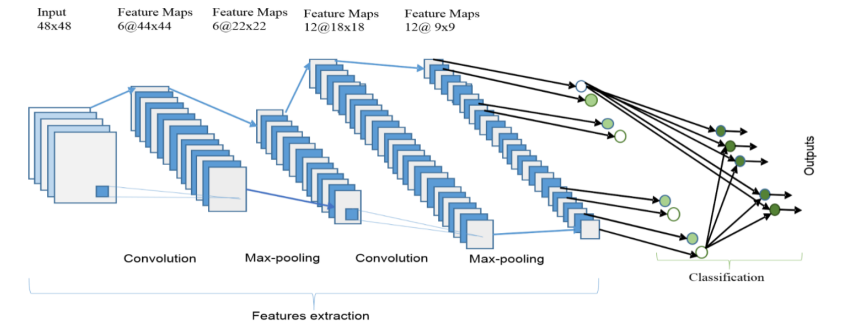


Fig. 11. The overall architecture of the CNN includes an input layer, multiple alternating convolution and max-pooling layers, one fully-connected layer and one classification layer.

source: *CS321n*

Language & Vision: Machine Learning

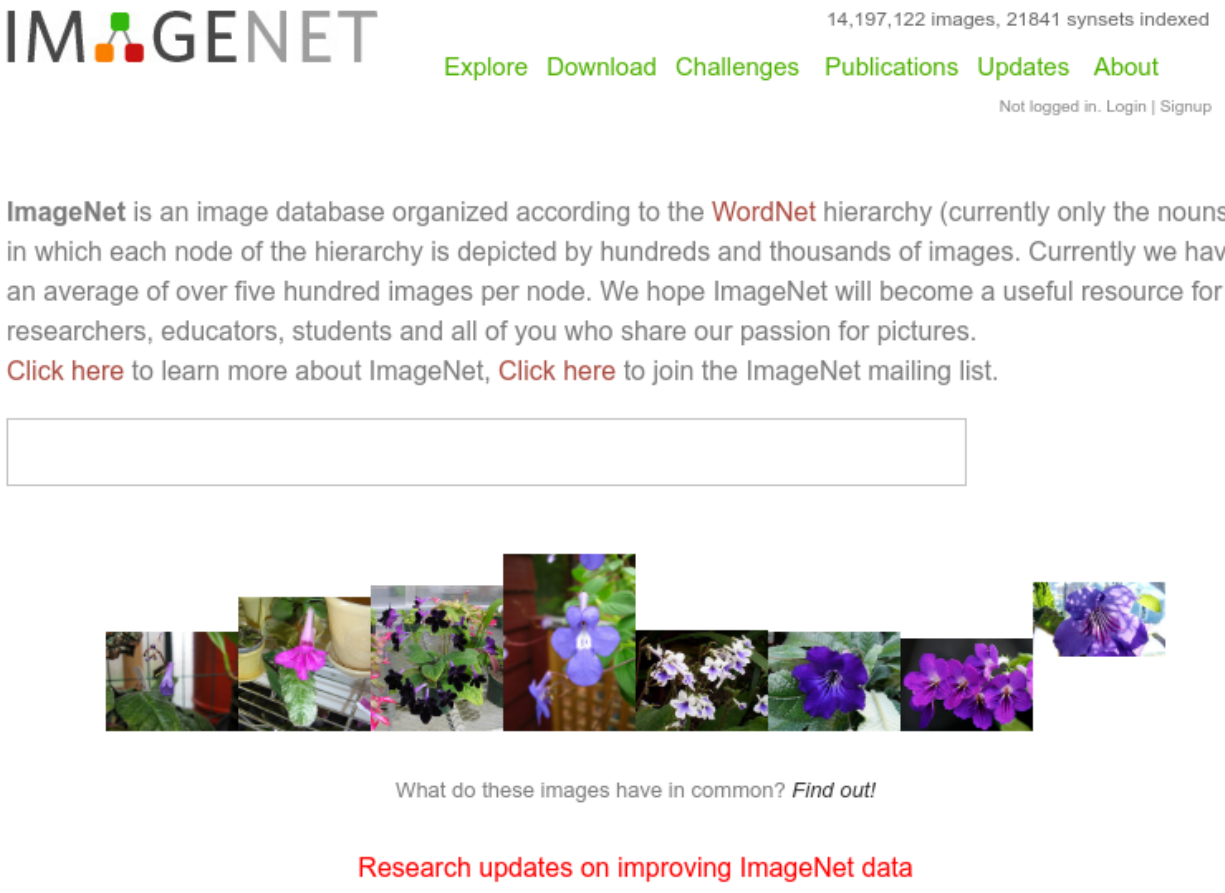
CV Methods: (Deep) Learning Roadmap

1. **Given a dataset** of images, each associated with, e.g., object class(es) or categories (or descriptions, cf. ViLBERT)
2. Possibly also **annotated with bounding boxes** of shown objects
3. **Train a deep learning model**: feed the model with many images; it will learn features along with how to solve the task at hand (e.g., object detection, segmentation)
4. Evaluate the model on images not in the training data
   $\Rightarrow$ Test model (measure, e.g., accuracy)
5. **Transfer trained model**:
   Use the trained model **as *feature extractor*** to represent visual data for other tasks, or for L&V methods
   $\Rightarrow$The model is called to be *pre-trained*

Computer Vision Datasets (also L&V datasets)

ImageNet: https://image-net.org

- localisation, detection, image classification, segmentation, ...
- Standard dataset for pre-training models



Transfer Learning: CNN-based Visual Features

Basic Layers of a CNN

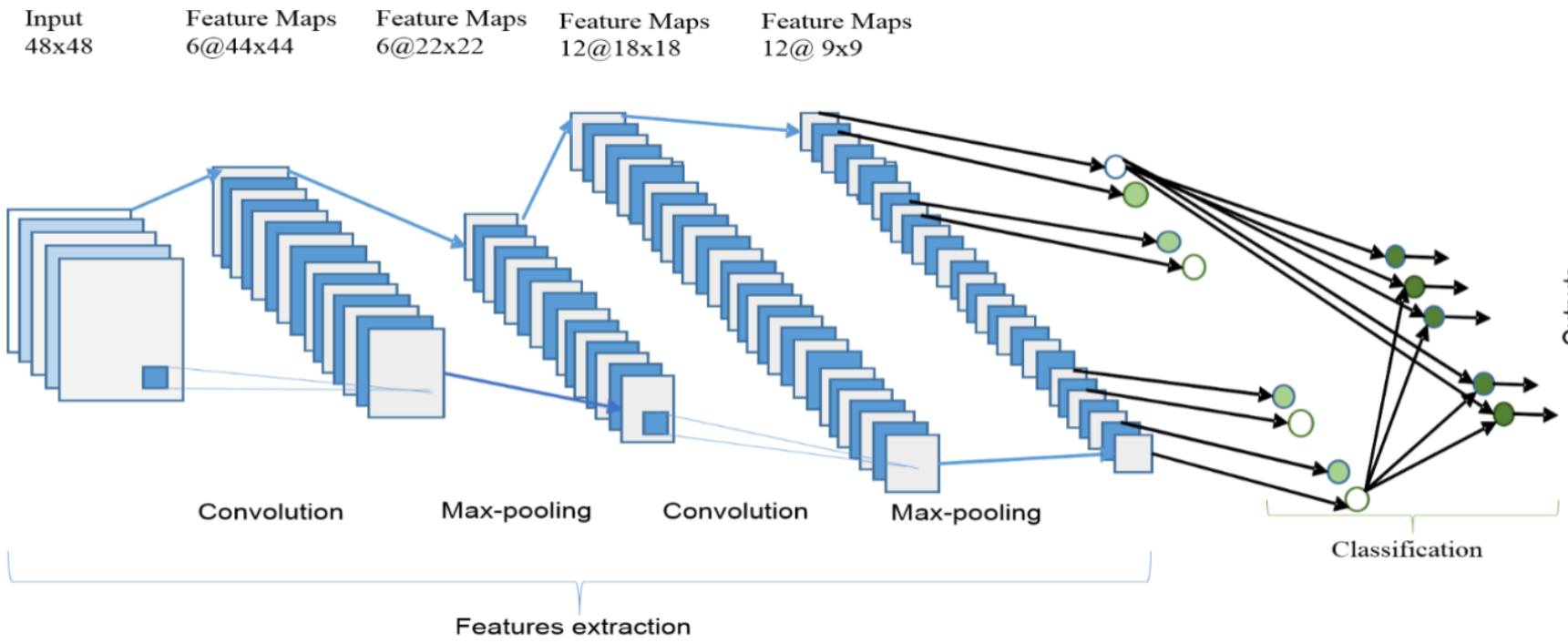
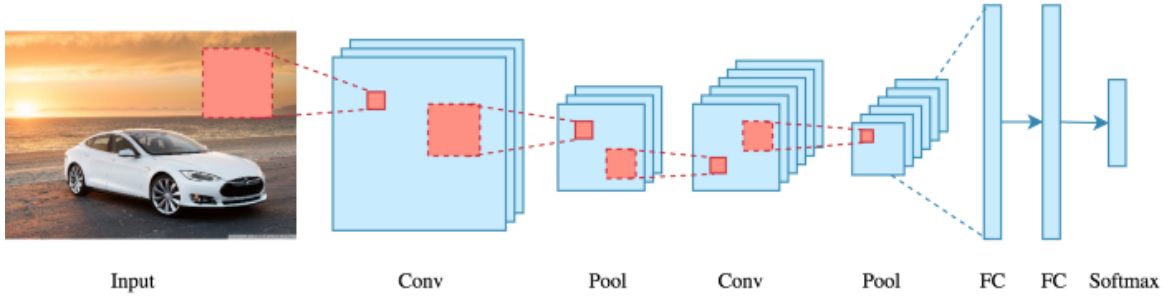


Fig. 11. The overall architecture of the CNN includes an input layer, multiple alternating convolution and max-pooling layers, one fully-connecte layer and one classification layer.

source: *[?]*

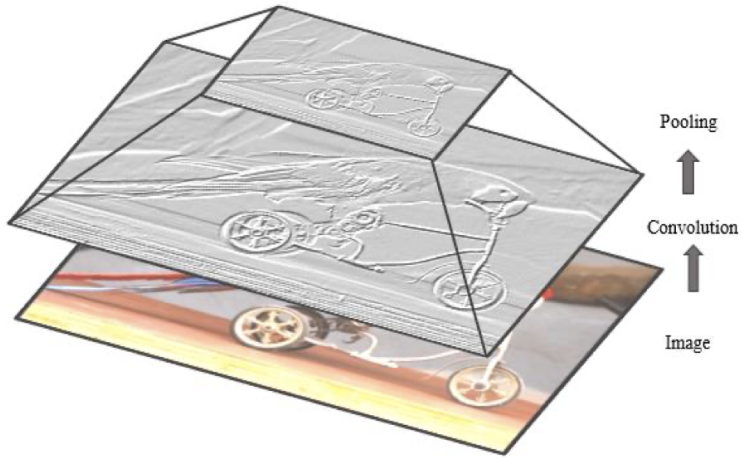
Deep Learning: CNN-based Approaches

Basic Layers of a CNN



source: *towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks*

- Convolutional Layer
- Pooling Layer
- Fully Connected Layer / Output Layer



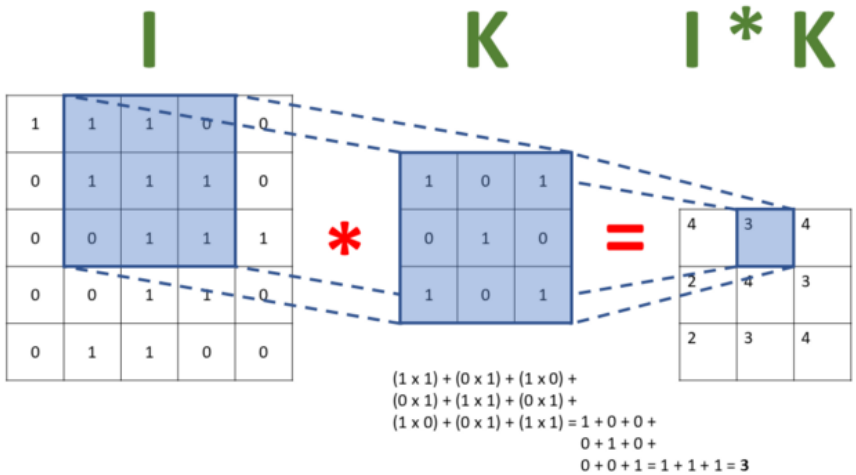
Basic Layers of a CNN

Input Layer

- Holds the raw pixel values of the image
- Example: $[32 \times 32 \times 3]$ pixel values: width 32, height 32, and with three color channels R,G,B

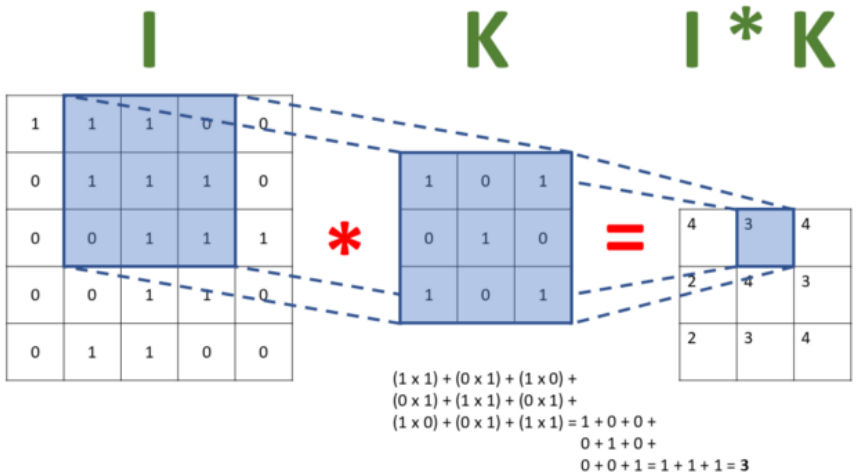
Basic Layers of a CNN

Convolutional Layer



Basic Layers of a CNN

Convolutional Layer



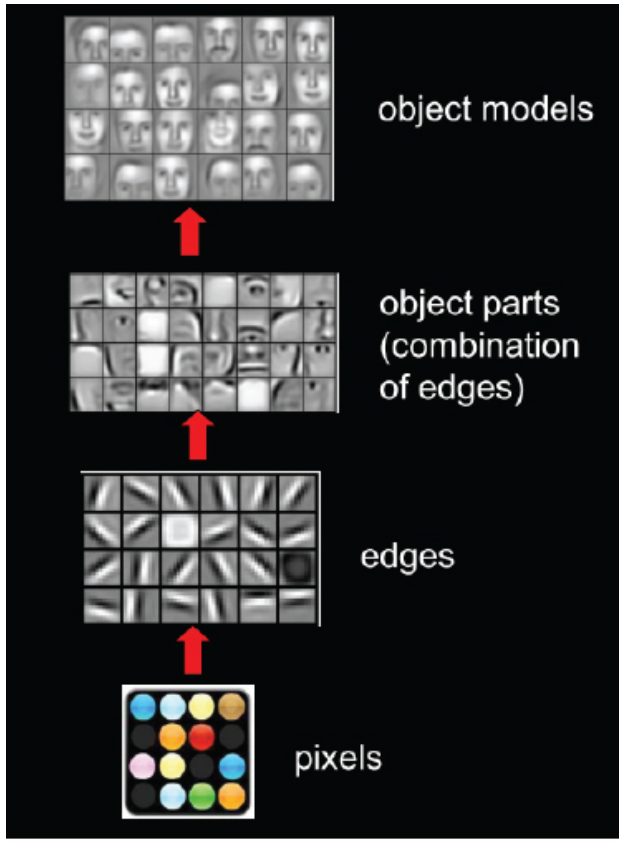
- Filters detect valuable object features in image I
- A filter is a 3D matrix with dimensions $h \times w \times d$
- Dot product between filter and image patch

$\Rightarrow$ Feature / Activation map

blog: `towardsdatascience.com/how-convolutional-neural-network-works-cdb58d992363`

We want to detect different patterns

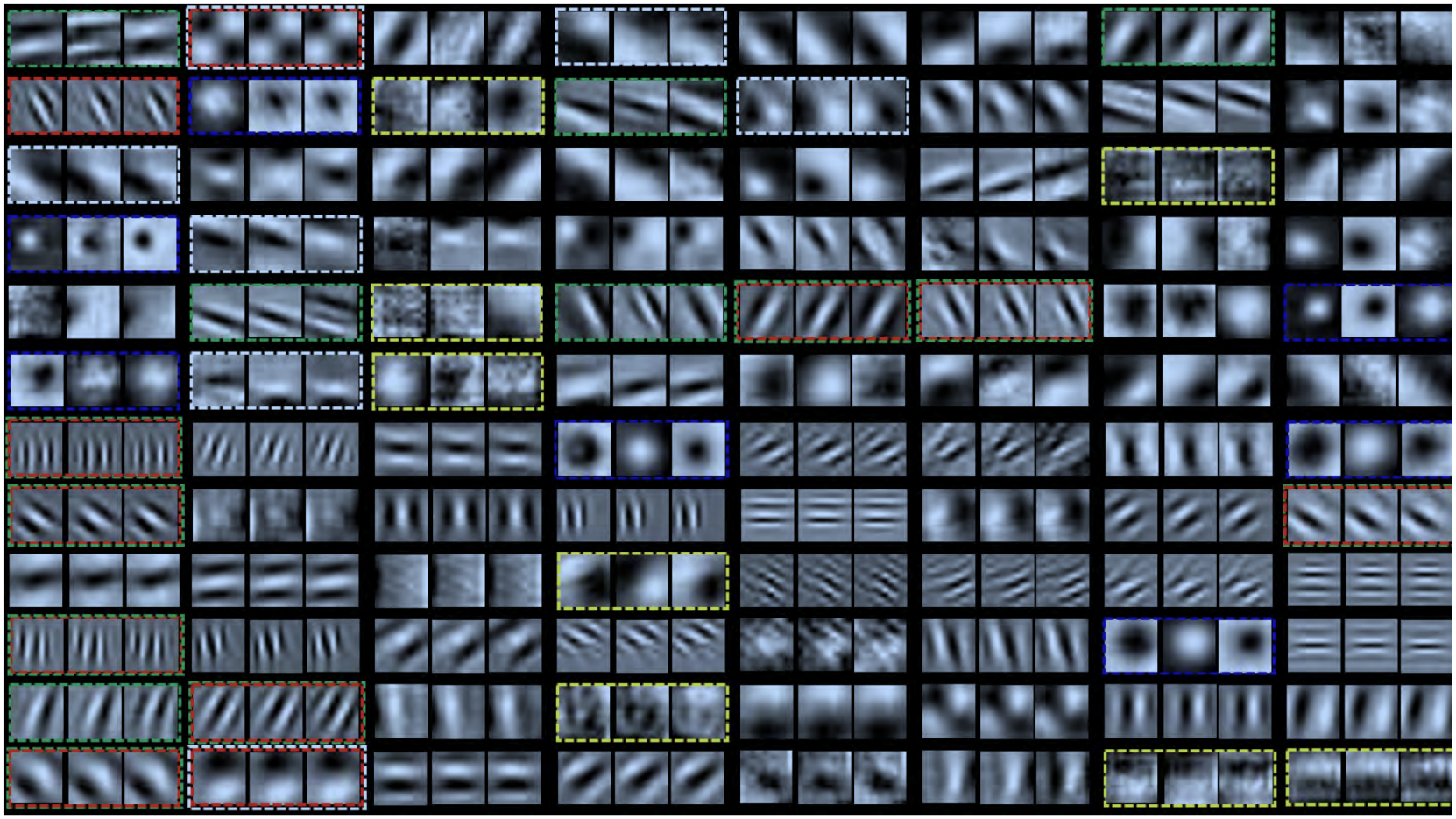
Visualisation of CNNs filters & outputs



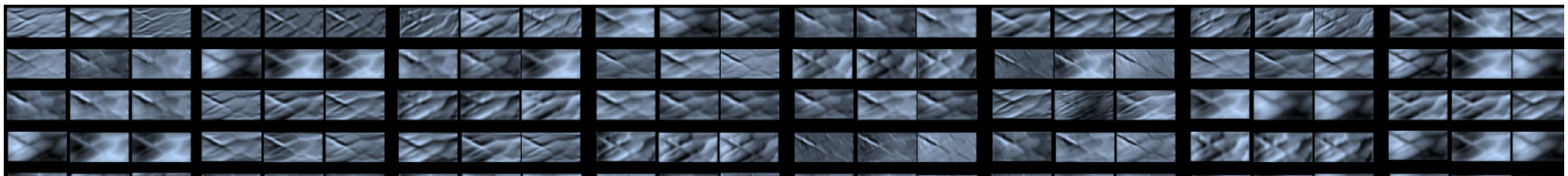
source: *datascience.stackexchange.com/questions/15671/how-are-per-layer-detected-patterns-in-a-trained-cnn-plotted*

We want to detect different patterns

Visualisation of CNNs filters & outputs [?]

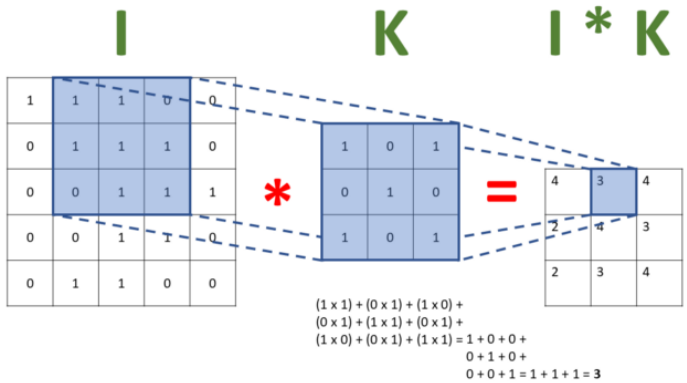


(a)



Basic Layers of a CNN

Convolutional Layer



- Filters detect valuable object features in image I
- A filter is a 3D matrix with dimensions $h \times w \times d$
- It has d kernels, each kernel K is of size $h \times w$
- Dot product between filter and image patch

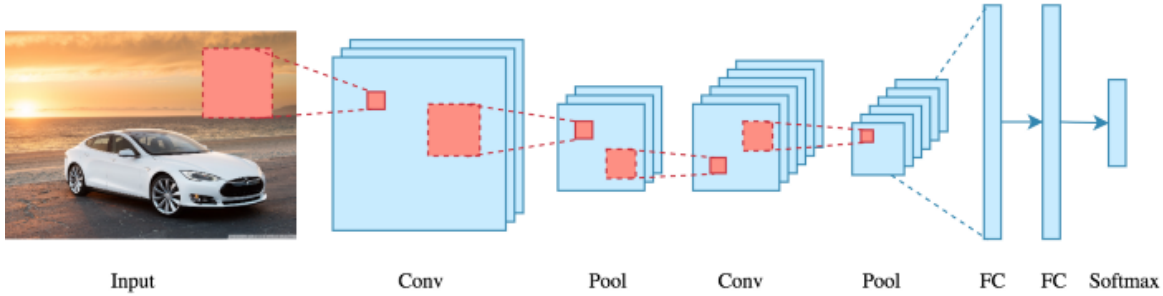
$\Rightarrow$ Feature / Activation map

- Multiple filters are used in one conv layer, each detecting different features $\leftrightarrow$ cf. multiple attention heads per layer

source: *blog: https://towardsdatascience.com/*

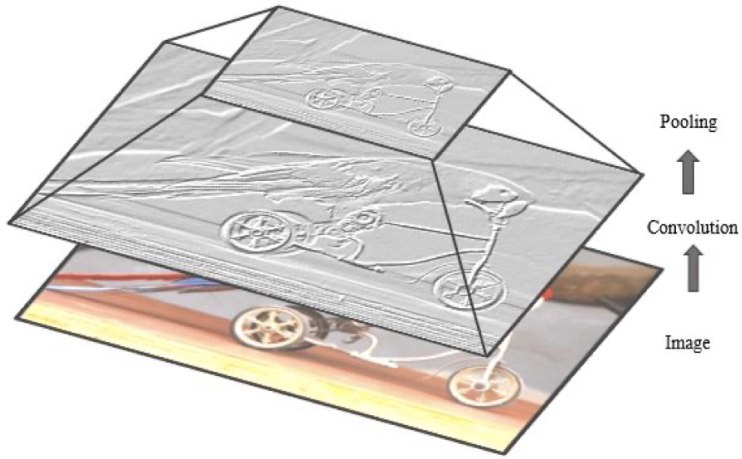
Deep Learning: CNN-based Approaches

Basic Layers of a CNN



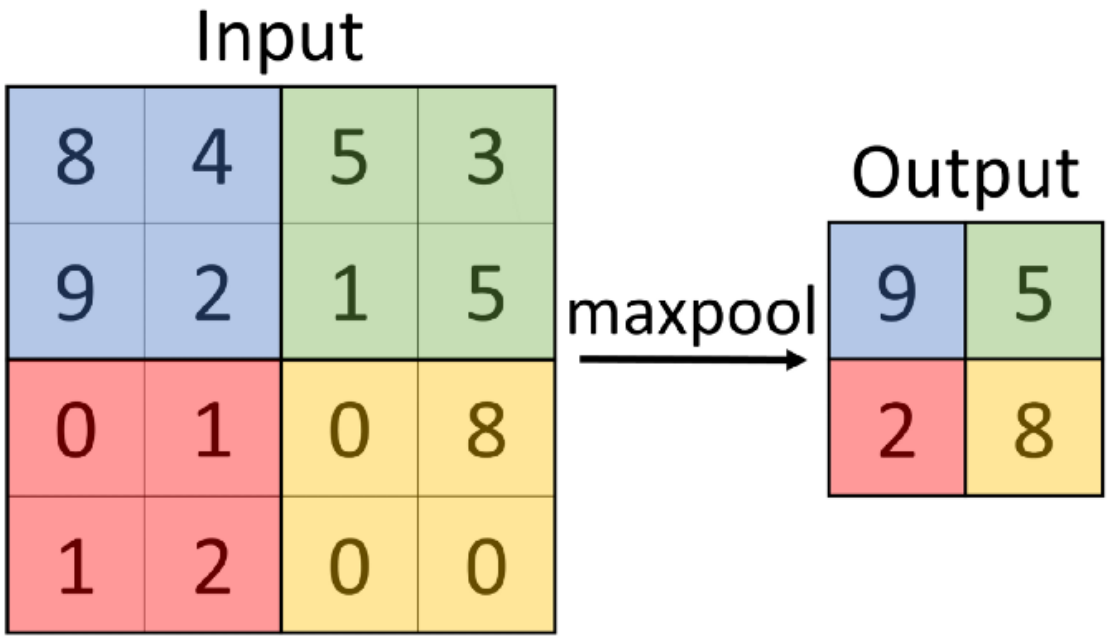
source: *towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks*

- Convolutional Layer
- Pooling Layer
- Fully Connected Layer / Output Layer



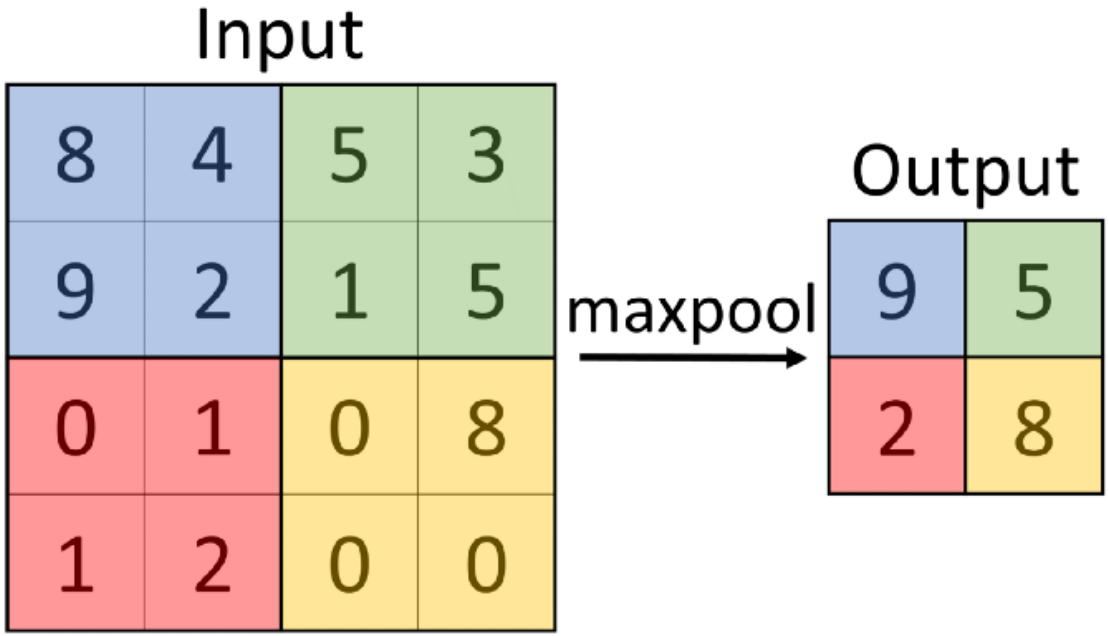
Basic Layers of a CNN

Pooling Layer



Basic Layers of a CNN

Pooling Layer

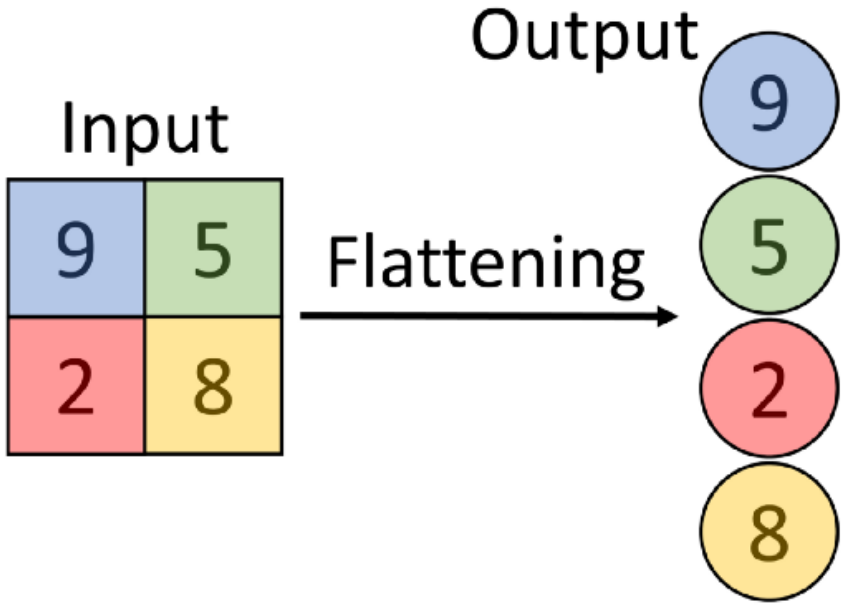


- Goal: Determine most valuable features
- Subsamples the output from the Convolutional Layer
- Max Pooling: Return maximum aggregate value
  (also: Average Pooling)

source: *blog: https://towardsdatascience.com/how-convolutional-neural-network-works-cdb58d992363*

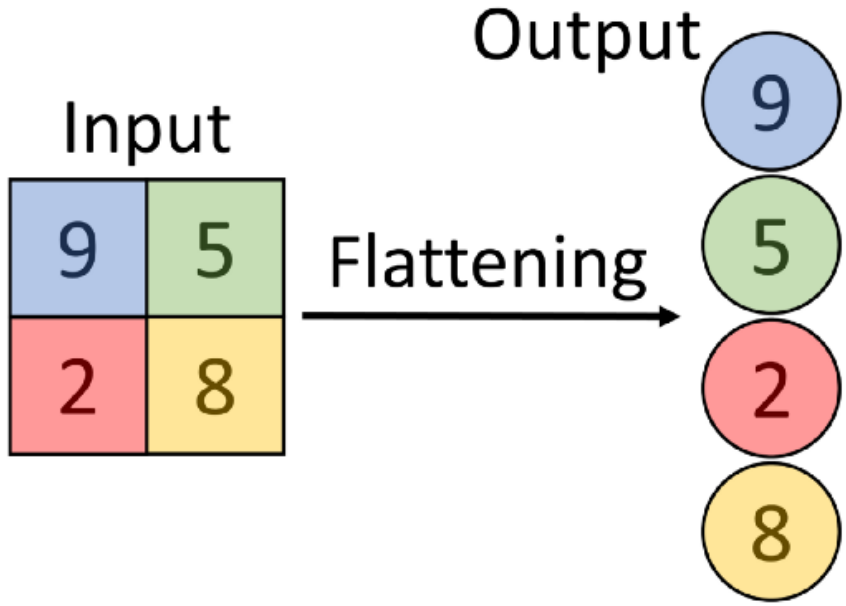
Basic Layers of a CNN

Fully Connected Layer (FC)



Basic Layers of a CNN

Fully Connected Layer (FC)

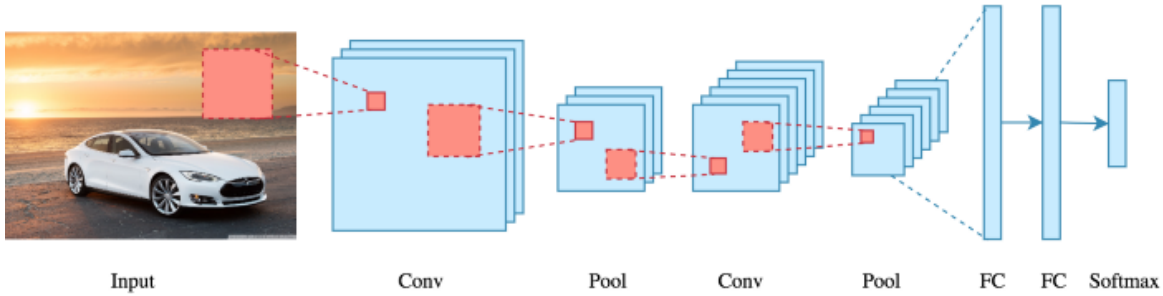


- Goal: Flattening
- Multiple fully connected layers (FCs)
- Last FC is output layer

source: *blog: https://towardsdatascience.com/how-convolutional-neural-network-works-cdb58d992363*

Deep Learning: CNN-based Approaches

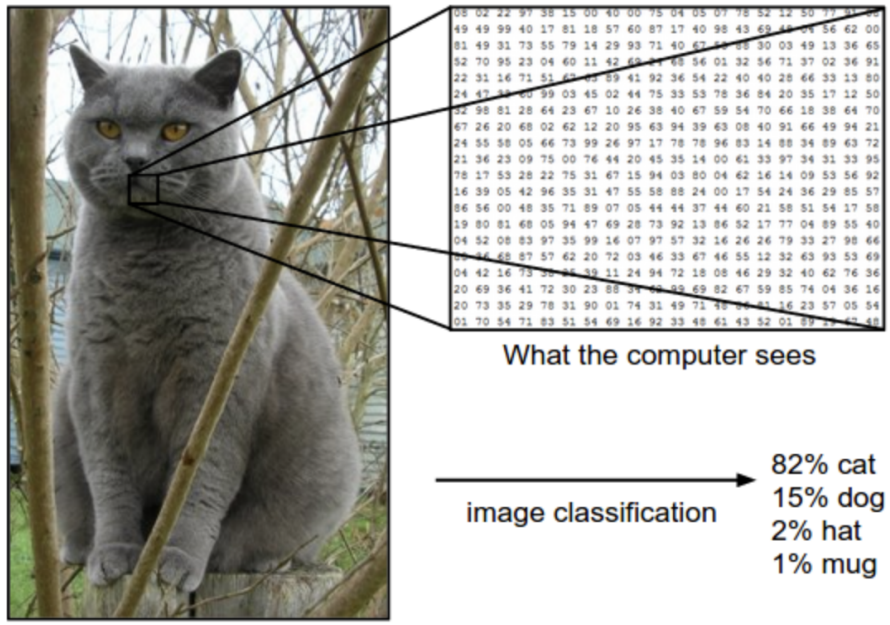
Basic Layers of a CNN



source: *towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks*

- Convolutional Layer
- Pooling Layer
- Fully Connected Layer / Output Layer
- Softmax for classification

Deep NN Architectures for Image Classification



source: *He et al.'15*

Deep NN Architectures for Image Classification

source: *He et al.'15*

- VGG — Simonyan and A. Zisserman, '15
- ResNet — He et al.,2015
- Inception — [?, ?]

Language & Vision: Machine Learning

Computer Vision Methods and Visual Representations

(Deep) Neural Architectures for Computer Vision

- Image classification: AlexNet, VGGNet, GoogLeNet, ResNet, etc.
- Object detection: region-based CNN (R-CNN), Fast R-CNN, Faster R-CNN, YOLO (and many more)
  - Object detector use pre-trained image classifier as backbone (i.e., initialise with pretrained classifier)

$\Rightarrow$ **Extract feature vectors from hidden layers which represent the input image, or image regions of it (e.g., objects)**

- Now: Transformers
  [Carion et al., 2020, Dosovitskiy et al., 2021, Zhu et al., 2021]
- (Image generation: GAN (generative adversarial networks) and Stable Diffusion)

References I

Andrews, M., Vigliocco, G., and Vinson, D. (2009).
Integrating experiential and distributional data to learn semantic representations.
Psychological review, 116(3):463.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012).
Distributional semantics in technicolor.
In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020).
End-to-end object detection with transformers.
In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham. Springer International Publishing.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021).
An image is worth 16x16 words: Transformers for image recognition at scale.
In *International Conference on Learning Representations*.

Hill, F., Reichart, R., and Korhonen, A. (2014).
Multi-modal models for concrete and abstract concept meaning.
Transactions of the Association for Computational Linguistics, 2:285–296.

Johns, B. T. and Jones, M. N. (2012).
Perceptual inference through global lexical similarity.
Topics in Cognitive Science, 4(1):103–120.

References II

Kiela, D. and Bottou, L. (2014).

Learning image embeddings using convolutional neural networks for improved multi-modal semantics.

In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 36–45. ACL.

Kiela, D., Hill, F., Korhonen, A., and Clark, S. (2014).

Improving multi-modal representations using image dispersion: Why less is sometimes more.

In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 835–841.

Lazaridou, A., Baroni, M., et al. (2015).

Combining language and vision with a multimodal skip-gram model.

In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005).

Semantic feature production norms for a large set of living and nonliving things.

Behavior research methods, 37(4):547–559.

Roller, S. and Im Walde, S. S. (2013).

A multimodal lda model integrating textual, cognitive and visual modalities.

In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157.

Silberer, C., Ferrari, V., and Lapata, M. (2013).

Models of semantic representation with visual attributes.

In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–582, Sofia, Bulgaria. Association for Computational Linguistics.

References III

Silberer, C., Ferrari, V., and Lapata, M. (2017).

Visually grounded meaning representations.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11):2284–2297.

Silberer, C. and Lapata, M. (2012).

Grounded models of semantic representation.

In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, Jeju Island, Korea. Association for Computational Linguistics.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021).

Deformable {detr}: Deformable transformers for end-to-end object detection.

In *International Conference on Learning Representations*.