Module 2 : Réseaux convolutionnels pour l'image
Formation Cdiscount

Nicolas Baskiotis
nicolas.baskiotis@lip6.fr
Benjamin Piwowarski
benjamin.piwowarski@lip6.fr

équipe MLIA, Laboratoire d'Informatique de Paris 6 (LIP6)
Sorbonne Université

Mercredi 25 Mai

Le domaine du *Computer Vision*

Pourquoi les MLP ne sont pas suffisants ?

- Une image est représentée en grille de pixels $\Rightarrow$ sémantique très pauvre
- Dans un fully-connected (MLP), chaque entrée a un rôle propre $\Rightarrow$ dépendance à la position du pixel

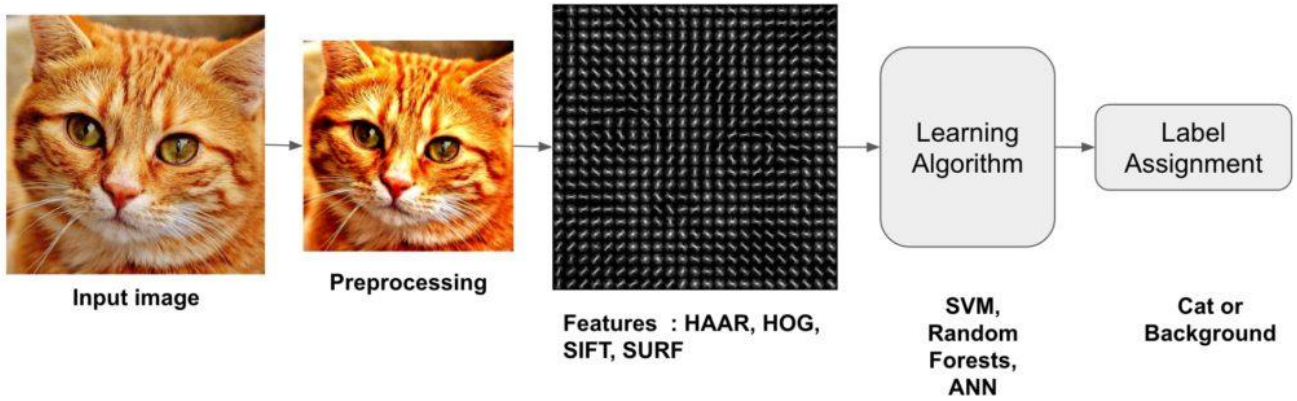
$\neq$ Dans une image, l'absolu n'a pas d'importance, le relatif est bien plus important (mais pas seulement)

Les *invariants* sont très importants en image car un objet peut prendre de multiples formes !

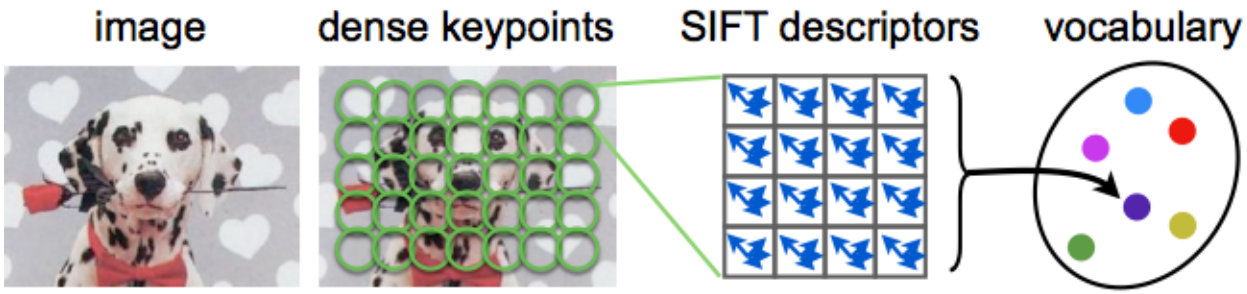
Avant le deep

La chaîne de traitement habituelle consistait à :

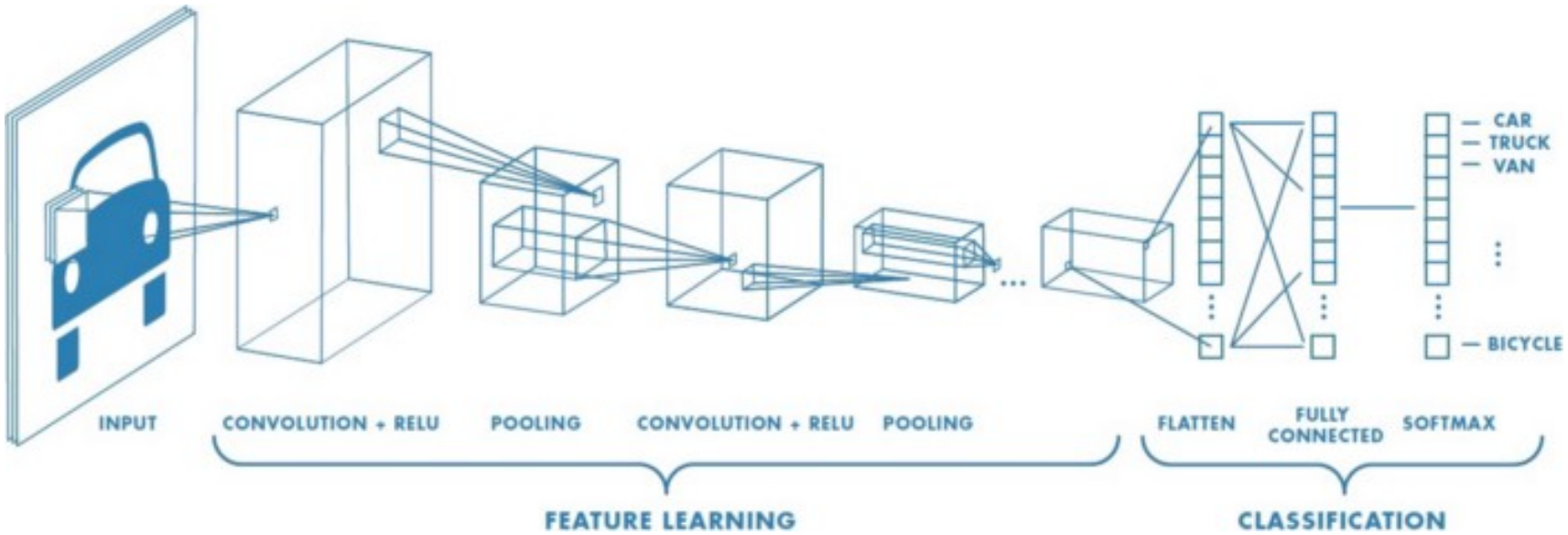
- extraire un dictionnaire de *features*
- agréger les features (représentation type *Bag Of Word*)
- utiliser un classifieur sur cette représentation pour classifier



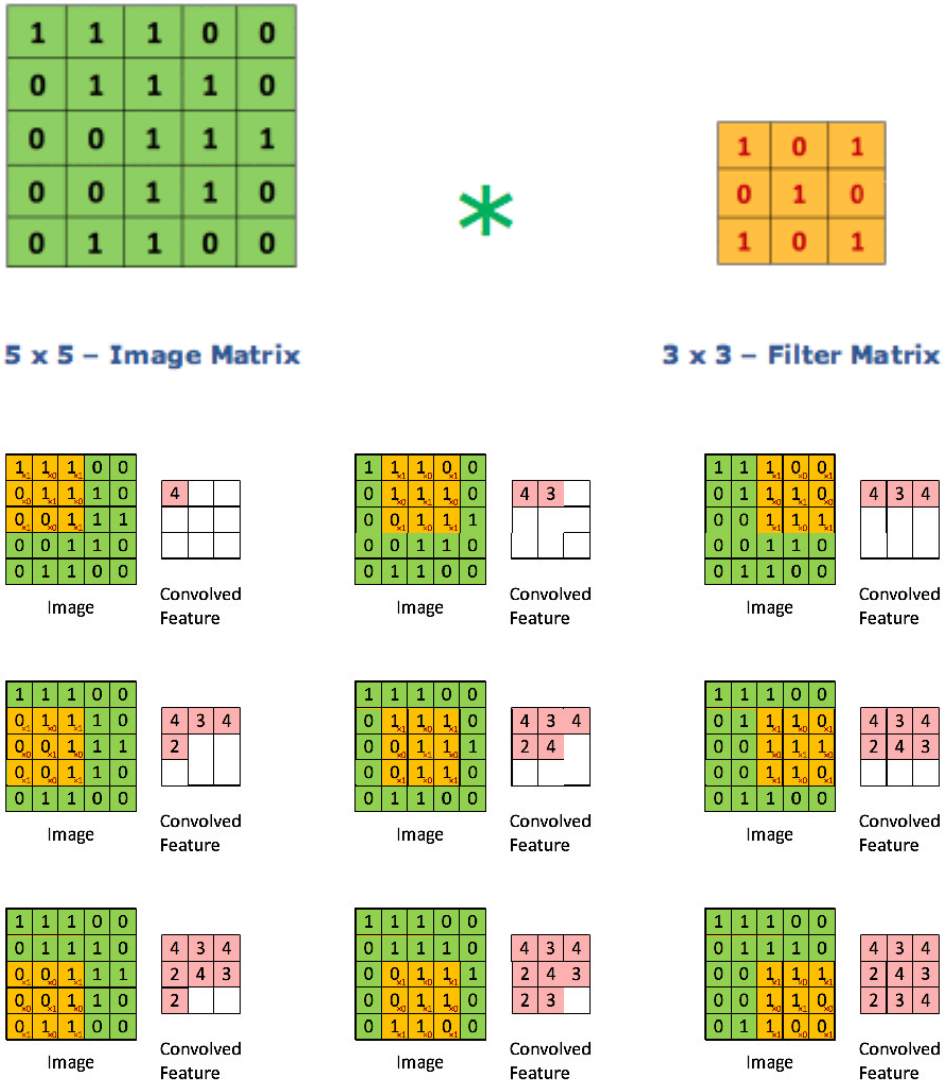
Le dictionnaire de features et son agrégation est très important : exemple pour SIFT :



Réseaux convolutifs



Une convolution



Application d'une transformation linéaire sur toutes les régions de l'image

Couche de *Pooling*

Pooling (ou subsampling) : Réduire la dimensionalité de sortie

- Max Pooling : on prend le max sur une fenêtre
- Average Pooling : on fait la moyenne
- Sum Pooling : la somme

...

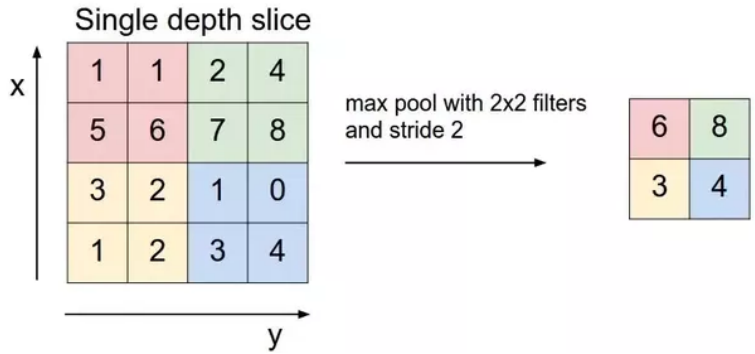
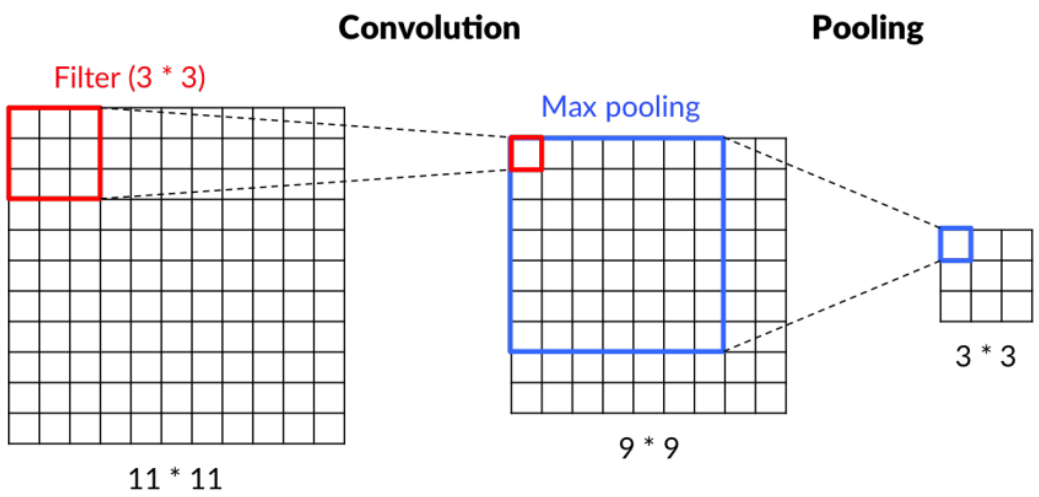
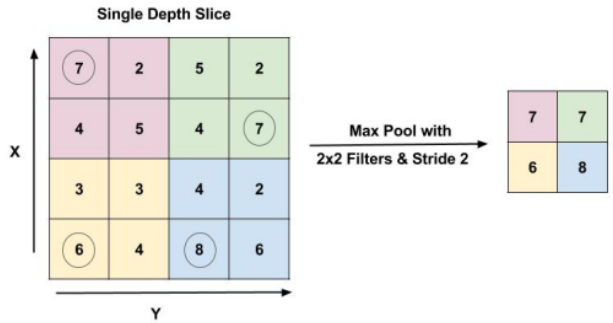
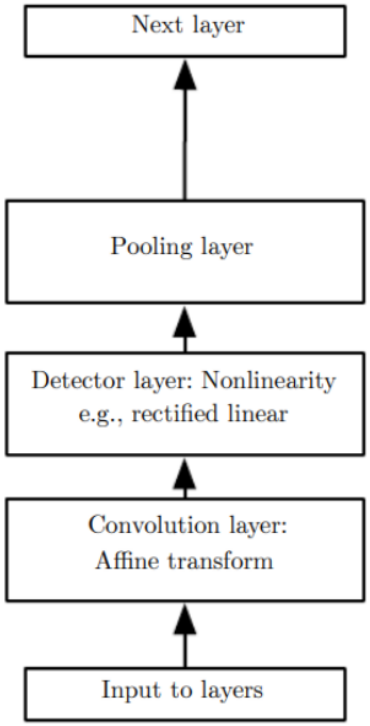


illustration : https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148

Couche convolutionnelle usuelle







Exemple

Reconnaissance de caractères [Duda et al 00]

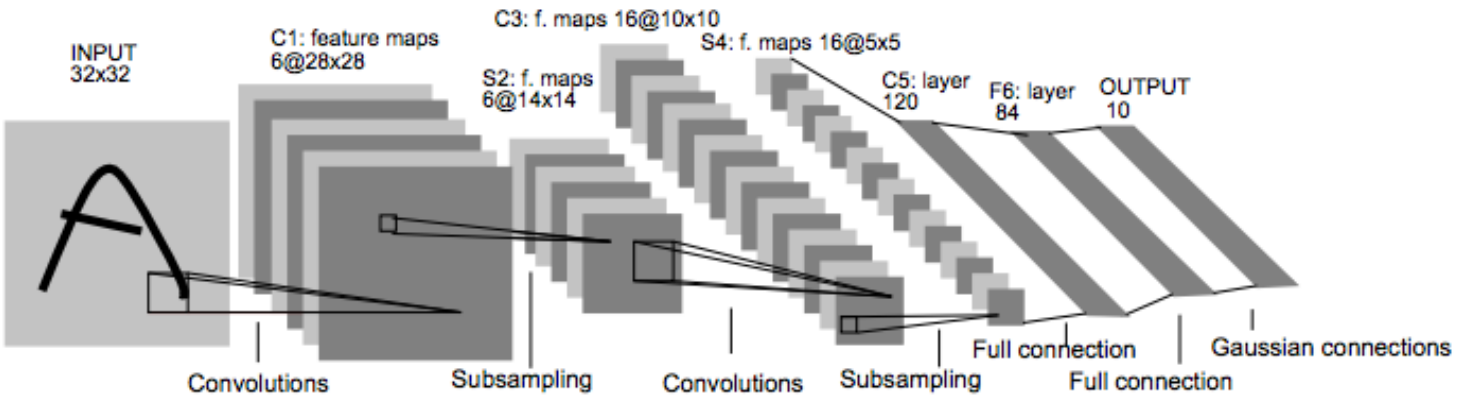
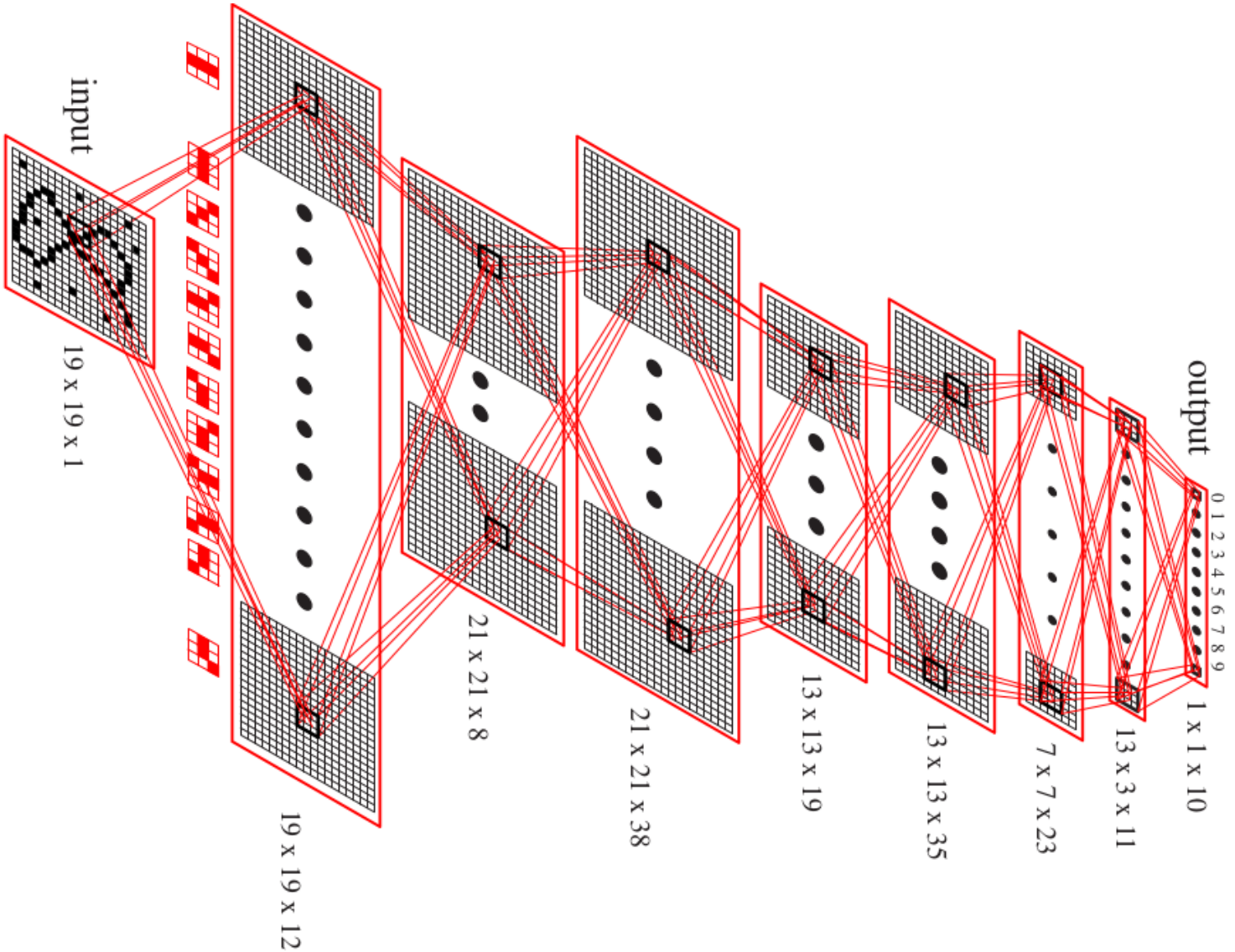


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Exemple

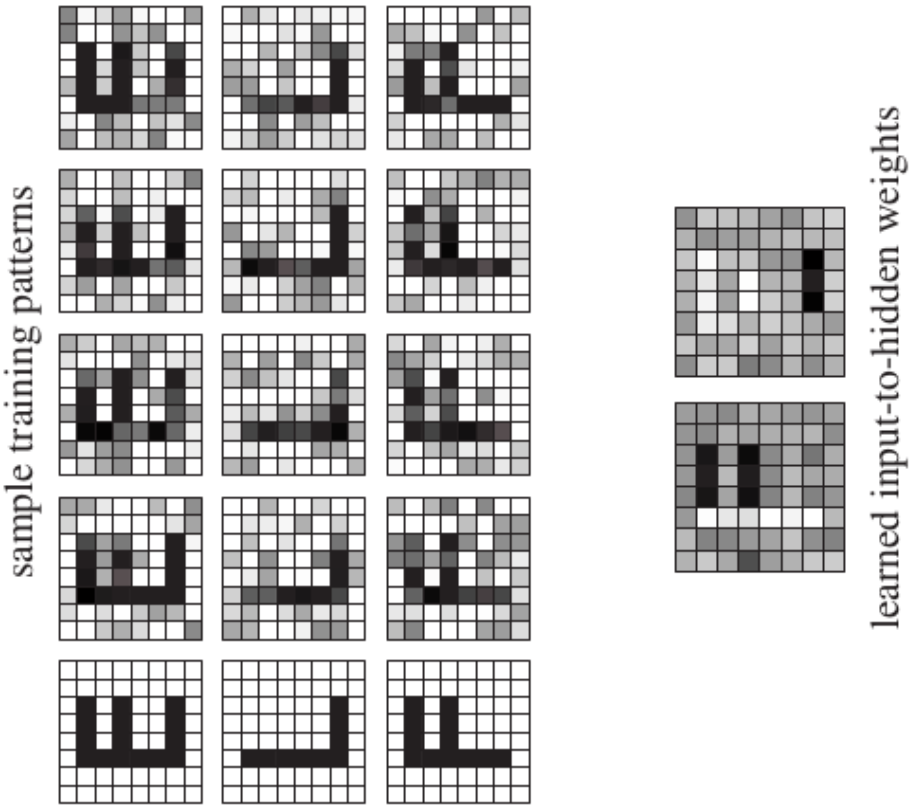
Reconnaissance de caractères

[Duda et al 00]

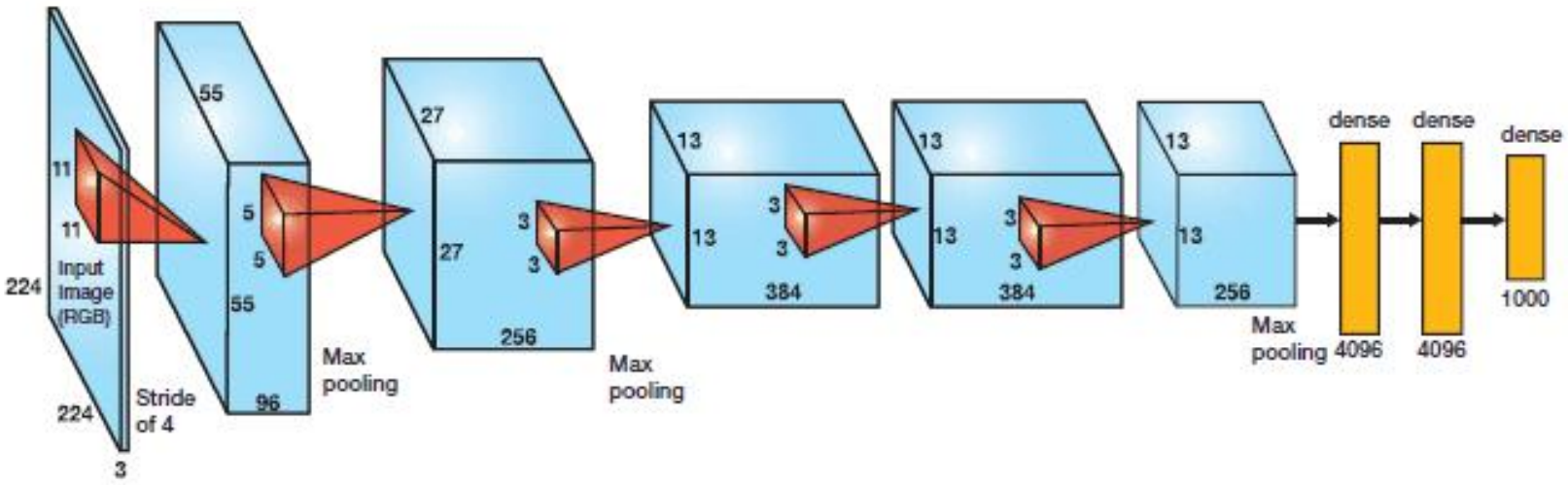


Exemple

Reconnaissance de caractères [Duda et al 00]

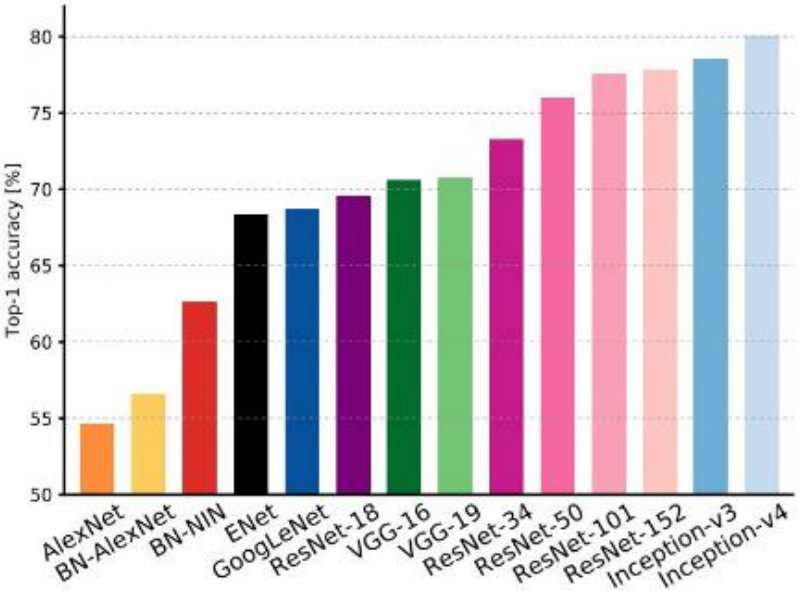


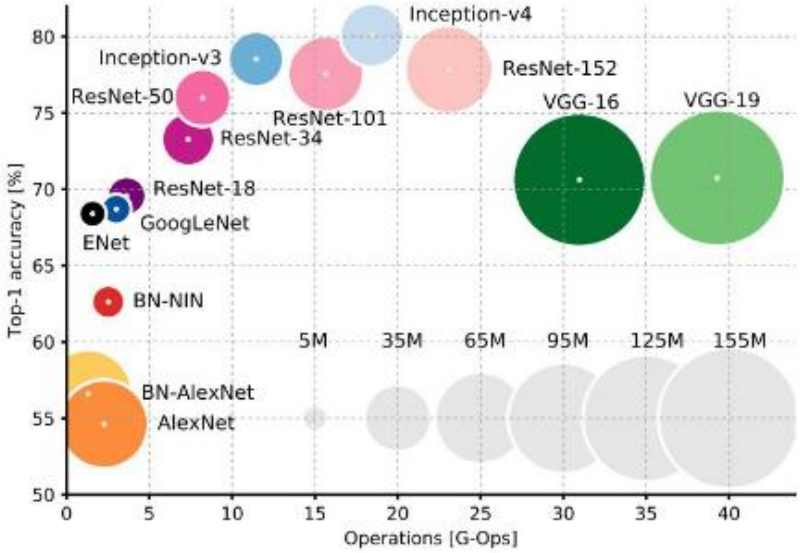
AlexNet (2012)



- $11 \times 11, 5 \times 5, 3 \times 3$ convolutions
- Max-pooling, ReLU activations
- Dropout et Data-augmentation

Un grand nombre d'architectures

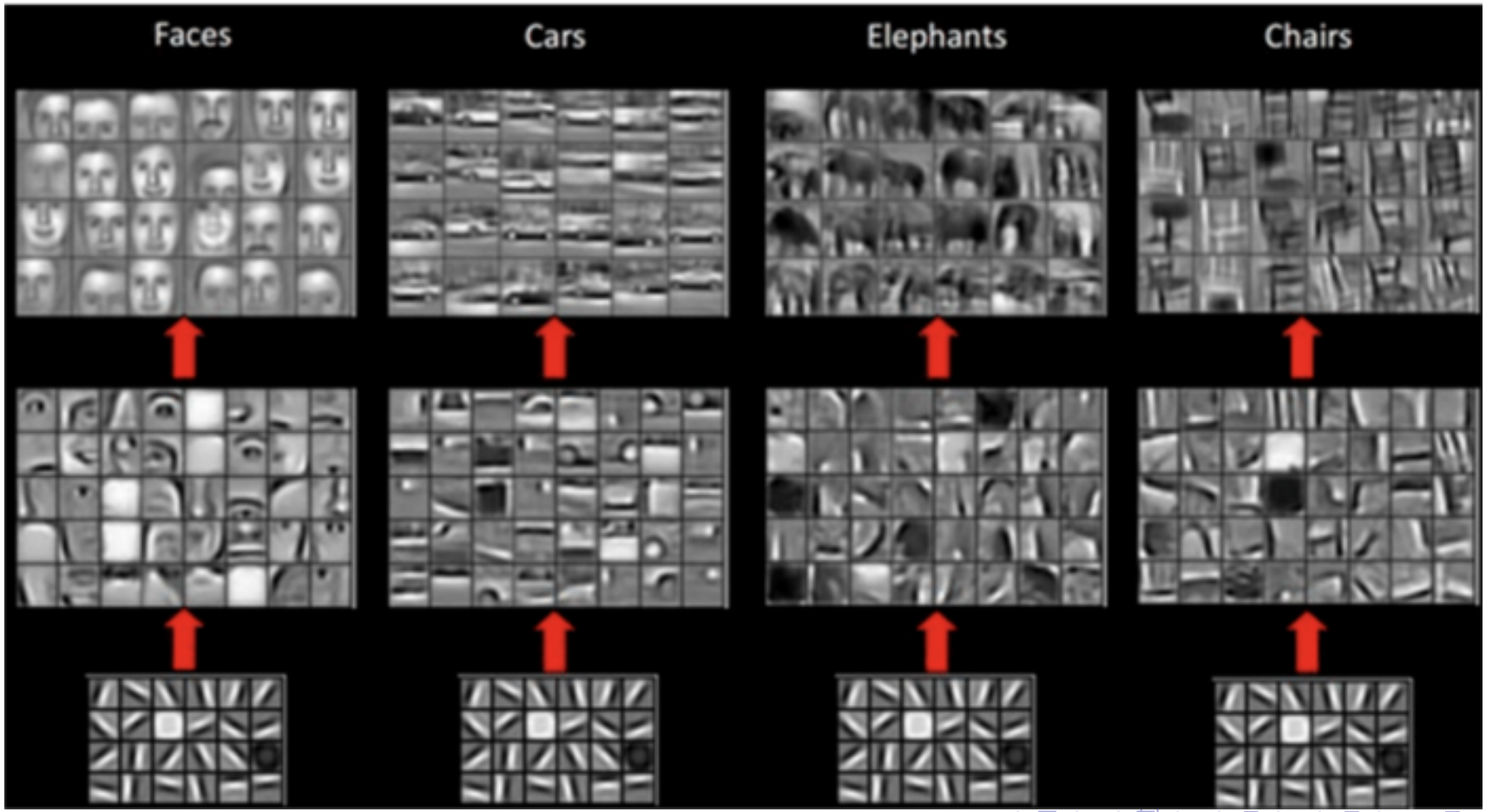




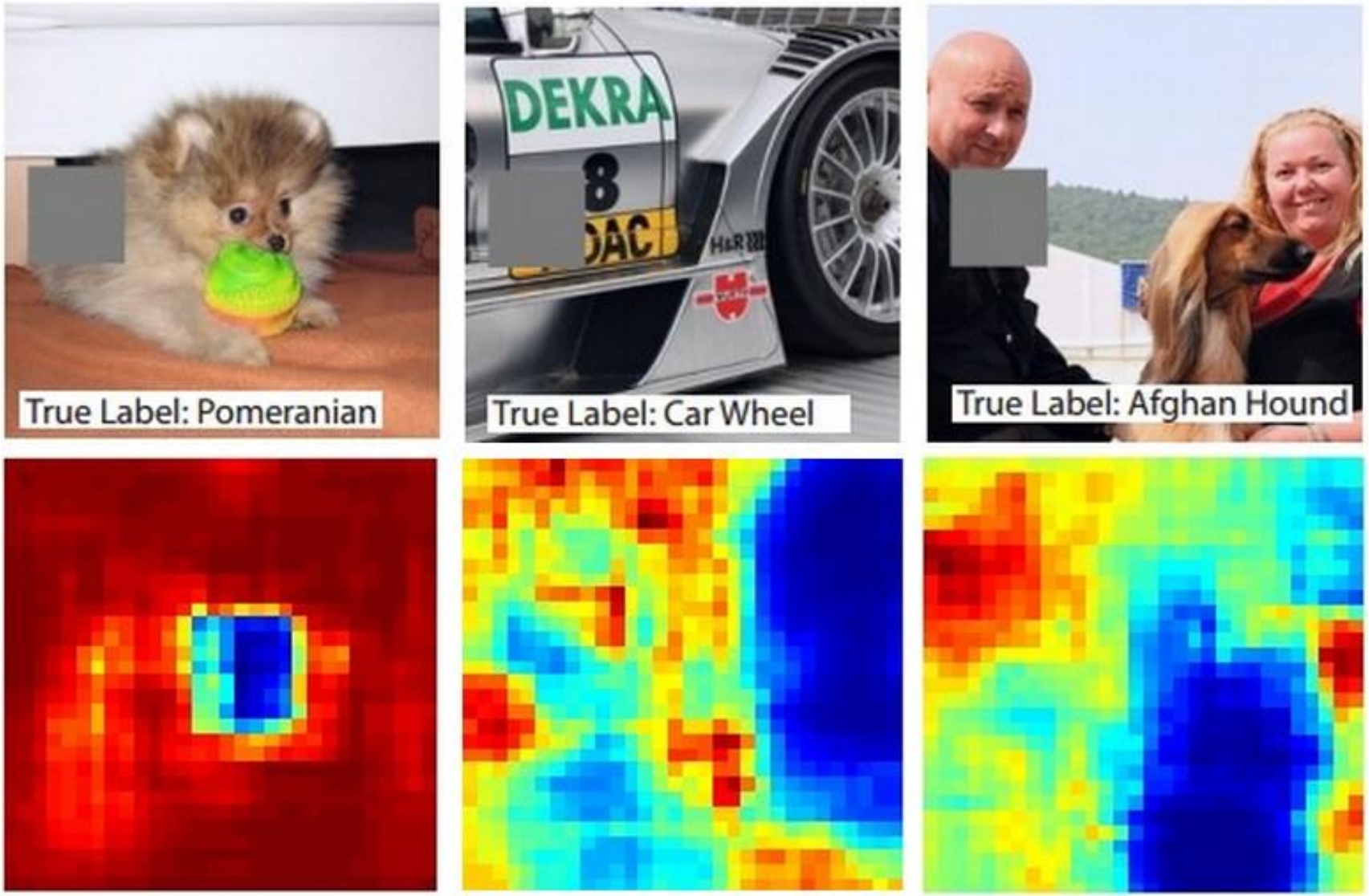
An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Introspection d'un CNN

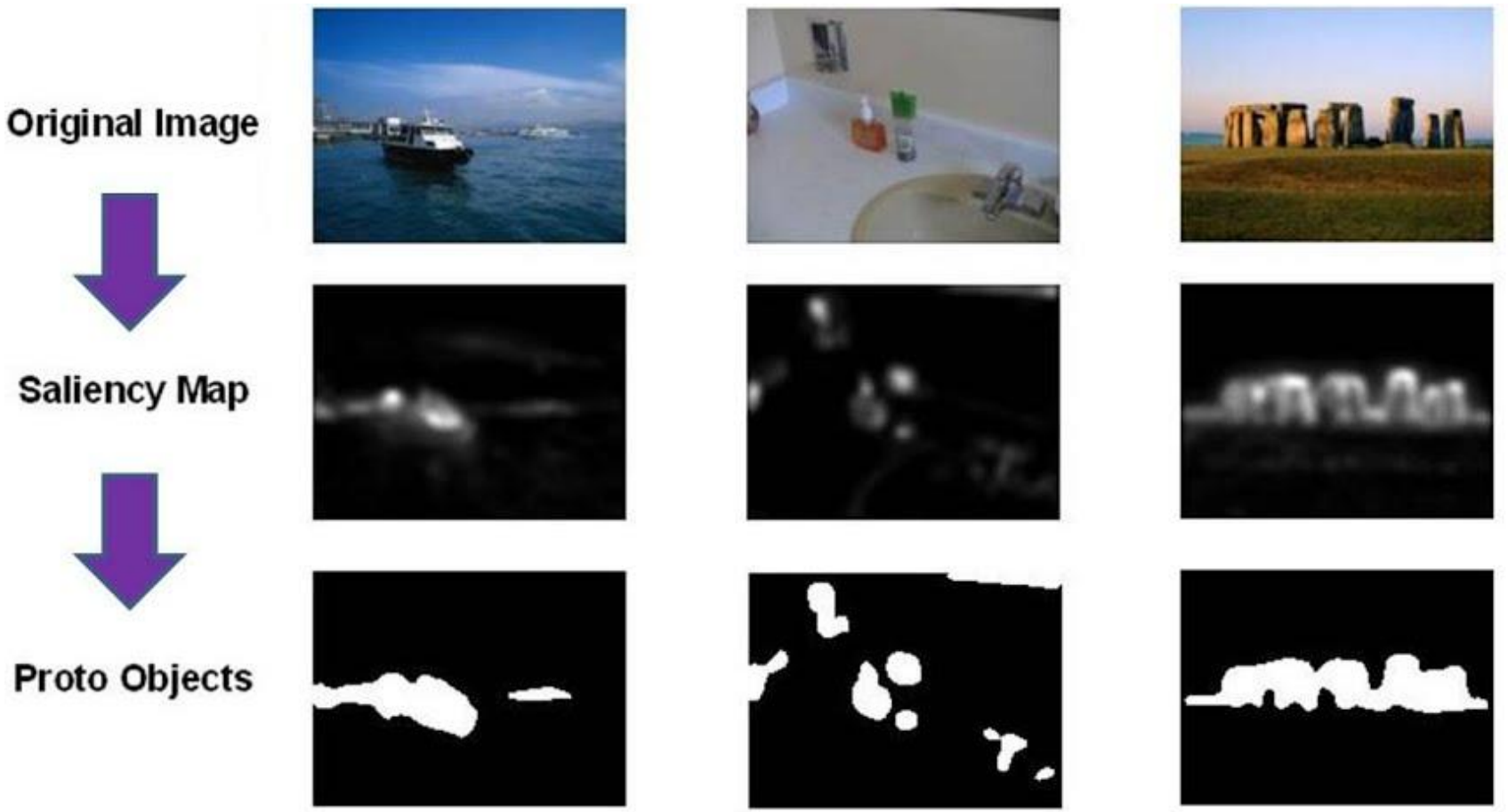
- Comprendre la classification : quelle région de l'image active la classification, quels filtres sont les plus importants . . .
- Les filtres sont de plus en plus abstraits, produisant des mots élémentaires visuels
- Les couches conservent les informations topologiques



Occlusion

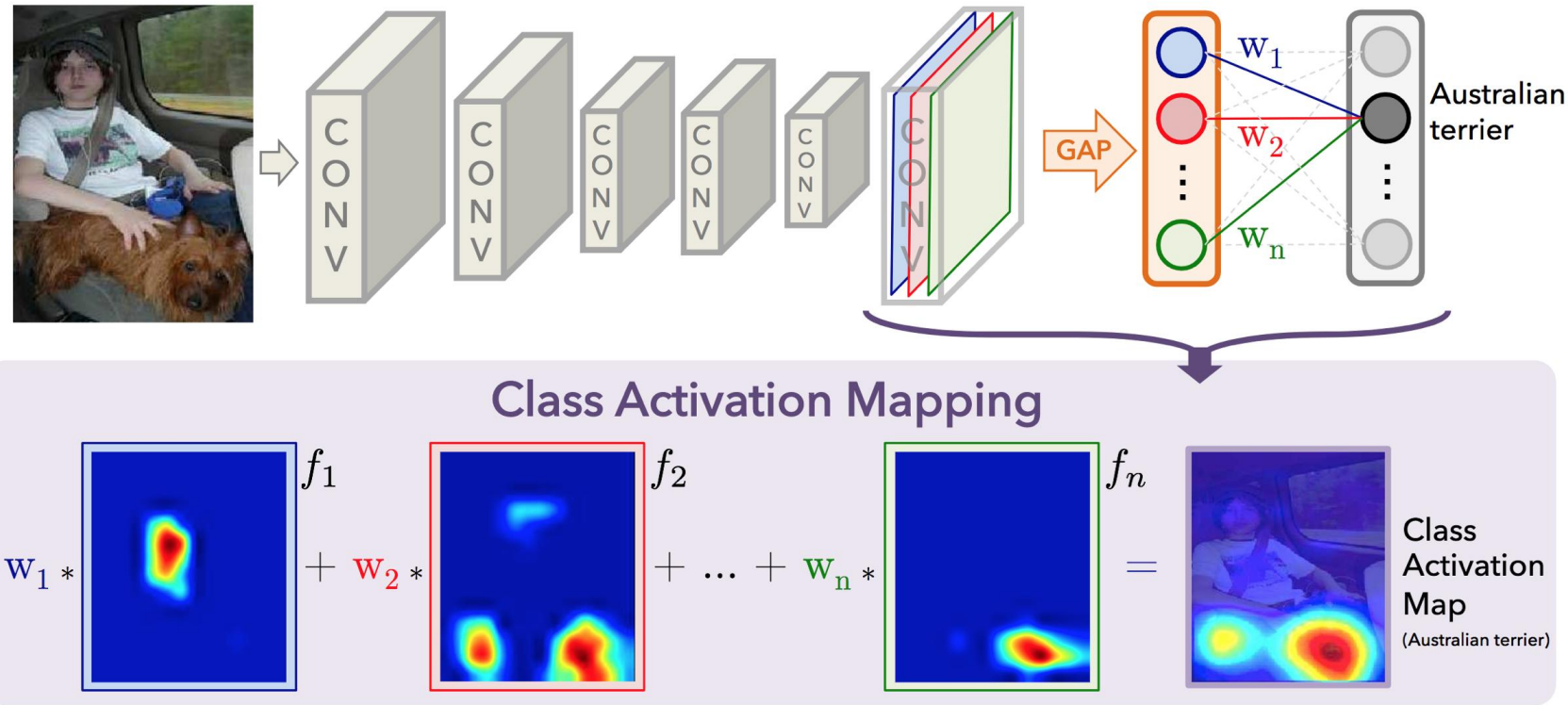

True Label: Pomeranian
DEKRA
8
DAC
True Label: Car Wheel
True Label: Afghan Hound

Carte de saillance



Elles sont calculées en prenant le gradient de la sortie par rapport à l'image d'entrée. Elles permettent de mettre en avant les pixels auxquels la classification est la plus sensible.

Class Activation Map



Class Activation Maps :

- Global Average Pooling (1 filtre = 1 feature)et pondération
- Entraînement d'un linéaire pour identifier les poids de chaque couche pour une entrée.
- Agrégation pondérée des filtres pour indiquer les régions d'intérêts.