DEEPLOMATICS Project: Kick-off Meeting

Nicolas Thome - Cnam Paris - CEDRIC / MSDMA

February 11, 2019

Context: Big Data

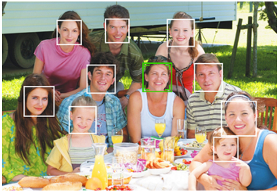
- Superabundance of data: images, videos, audio, text, user traces, *etc*

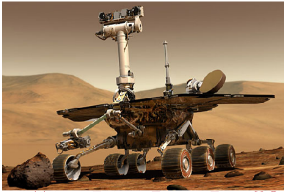
BBC: 2.4M videos

Social media,
e.g. Facebook: 1B each day

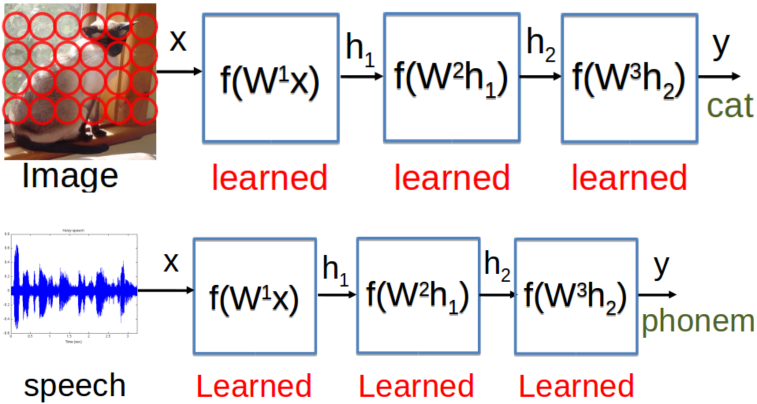
100M monitoring cameras

- Need to access, search, or classify these data: **Recognition**
- Huge number of applications: mobile visual search, robotics, autonomous driving, augmented reality, medical imaging *etc*



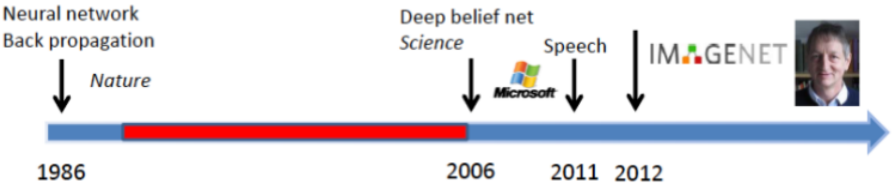


Deep Learning (DL) & Recognition of low-level signals



- **DL: learning intermediate representations**
  - vs handcrafted features
  - Filling the semantic gap
  - Disentangling data manifold

Deep Learning Success since 2010



- **2012: ImageNet ILSVRC Challenge** (Stanford)
  - Up to 2012, leading approaches: BoW + SVM
  - **ILSVRC'12: the deep revolution** ⇒ outstanding success of ConvNets [Krizhevsky et al., 2012]

Rank	Name	Error rate	Description
1	**U. Toronto**	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.26979	
4	Xerox/INRIA	0.27058	

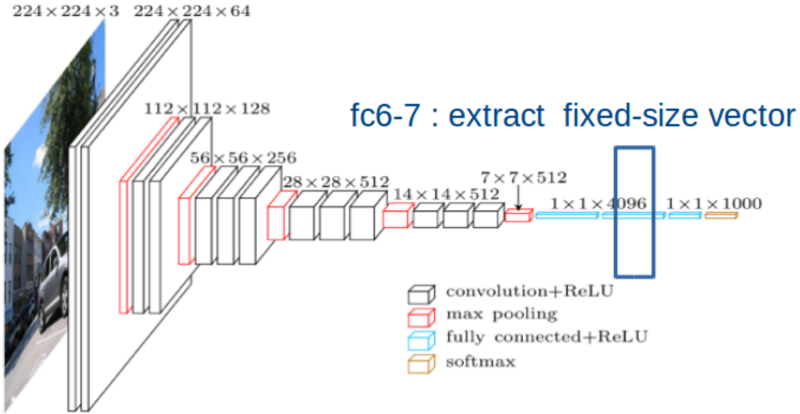
Deep ConvNet success at ILSVRC'12

Two main practical reasons:

1. Huge number of labeled images (10^6 images)
    - Possible to train very large models without over-fitting
    - Larger models enables to learn rich (semantic) features hierarchies
2. GPU implementation for training
    - Relatively cheap and fast GPU
    - Training time reduced to 1-2 weeks (up to 50x speed up)

Transferring Representations learned from ImageNet

- Deep ConvNets require large-scale annotated datasets
- **BUT:** Extract layer ⇒ fixed-size vector: **"Deep Features" (DF)**



- **Now state-of-the-art for any visual recognition task** [Azizpour et al., 2016]
    - Fine-tuning potentially improves performances

Deeplomatics Project: Task 3

- Deep Learning for drone recognition and tracking
  - Using RBG + optronic cameras
- Cnam, CEDRIC Lab, MSDMA Team (N. Thome)
  - Task 3.1: Supervised object detection (R. Fournier)
  - Task 3.2: Weakly supervised localization (N. Thome)
  - Task 3.3: Multi-modal detection (V. Audigier, A. Bar-Hen)

	Partenaires					Semestre du projet (M0-6, M6-M12, M12-M18, M18-M24, M24-M30, M30-M36)
	LMSSC	CEDRIC	ISL	ROBOOST		
TÂCHE 0	100%					MANAGEMENT DE PROJET
TÂCHE 1	73%	15%	11%	1%		CONSTITUTION ET AUGMENTATION DE BASE DE DONNÉES MULTIMODALES
					1.1	*Campagnes de mesures acoustiques et optroniques*
					1.2	*Augmentation de données et acquisitions sur capteurs compacts par synthèse de champ physique*
					1.3	*Annotation et gestion des bases de données*
TÂCHE 2	95%	4%		1%		LOCALISATION ET IDENTIFICATION ACOUSTIQUE SUR ANTENNES COMPACTES INTELLIGENTES
					2.1	*Conception et tests antennes*
					2.2	*Développement et évaluation de réseaux de neurones profonds pour la localisation et l'identification*
					2.3	*Transfert d'IA pré-entraînées*
TÂCHE 3		99%	0.5%	0.5%		SUIVI ET RECONNAISSANCE DE CIBLES PAR DEEP LEARNING VIDEO
					3.1	*Deep Learning supervisé sur images clés classiques et infrarouges*
					3.2	*Deep Learning faiblement supervisé*
					3.3	*Aprentissage sur données hétérogènes, données manquantes*
TÂCHE 4		4%	95%	1%		OPTIMISATION ET MOTORISATION ASSERVIE DU SYSTÈME OPTRONIQUE
					4.1	*Accrochage de cible et motorisation*
					4.2	*Méthodes complémentaires*
					4.3	*Poursuite de cible*
TÂCHE 5	32%	32%	35%	1%		FUSION DE DONNÉES MULTIMODALES ET MULTICAPTEURS
					3.1	*Spécification des données à fusionner*
					3.2	*Fusion de données multicapteurs et multimodales*

Outline

Deep Features for Localization

- Core (simple) idea: deep features for local information in image regions
  - Crop given image sub-area
  - Rescale → ImageNet input size, *e.g.* 224×224

Localization with Region-CNN [Girshick et al., 2014]

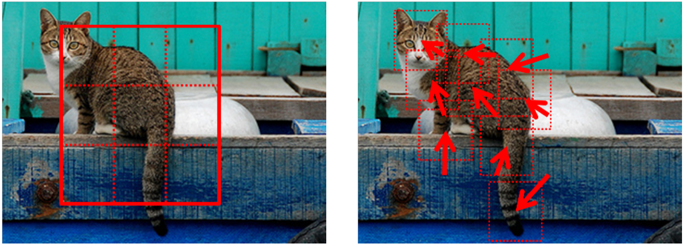
1. R-CNN, 1^{st} step: extract a set of region proposal candidates
   - Goal: pre-select candidates based on their "objectness"
   - Low-level, unsupervised
   - Many approaches,*e.g.* selective search [Uijlings et al., 2013]

Localization with Region-CNN [Girshick et al., 2014]

2. R-CNN, 2^{nd} step: classifiy each regions proposal
    - Rescale proposal & extract deep feature
    - Add transfer layer with $K+1$ classes
        - +BB regression, *i.e.* remap proposal (red) $\rightarrow$ GT BB (green)

Part-based Representations [Mordan et al., 2018a]

Part-based representations better adapt to objects than boxes



Goal: boost spatial invariance of ConvNets, without additional annotations

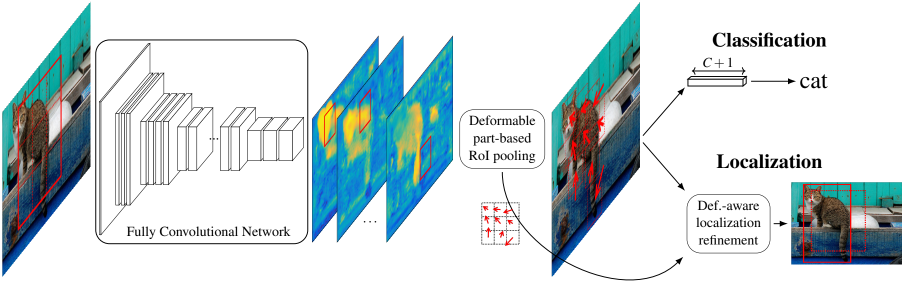
Contribution: efficient end-to-end learning of deep part-based features

→ Improving both recognition and localization

- Idea PoC at BMVC'17 [Mordan et al., 2017b]
- Extended version at IJCV'18 [Mordan et al., 2018a]

DP-FCN Global Architecture [Mordan et al., 2018a]

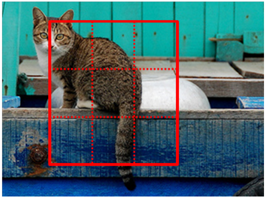
Exploits **deformable parts** in **region-based deep ConvNets**

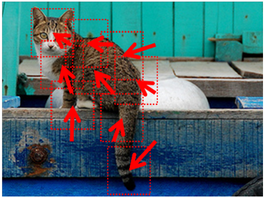


3 main blocks:

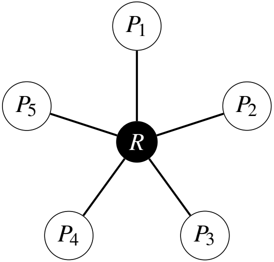
- FCN backbone architecture ⟶ higher efficiency
- **Deformable part-based RoI pooling ⟶ better recognition**
- **Def.-aware localization refinement ⟶ finer localization**

Deformable Part-based RoI Pooling [Mordan et al., 2018a]



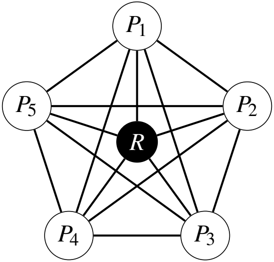


Independent deformations
(Star model, *c.f.* DPM)



Simple and light optimization

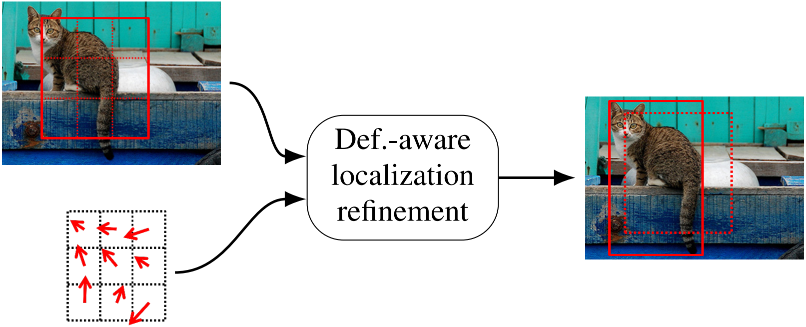
Joint deformations
(Fully connected model)



Heavy but fine optimization

Deformation-aware Localization Refinement [Mordan et al., 2018a]

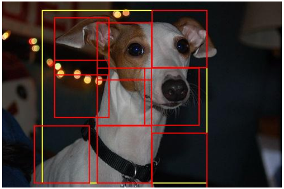
Spatial layout of parts → **geometric information** for localization

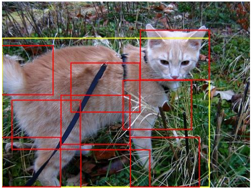


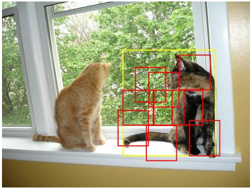
- **Coarse shapes** of objects with positions of parts
- Final localization: combination of
  - deep visual features at deformed locations
  - geometric displacements of parts

Visualizations of Deformations and Detections

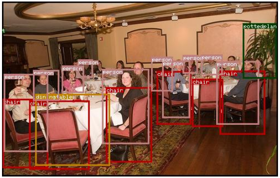
Deformation of parts (3×3 parts):





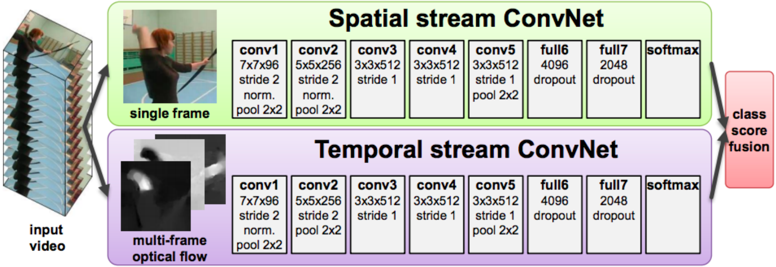


Example detections (on VOC07+12):



Object Detection in Videos: Task 3.1

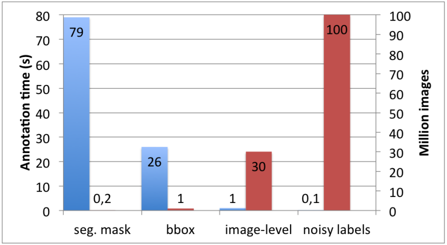
- Requires Bounding Box Annotation
  - Good to have at least a sub-set for evaluating localization quality (Task 3.2)
  - Starting by task 3.2?
- Extension to videos
  - 2-stream, Flow+image [Simonyan and Zisserman, 2014]
  - Detection + tracking
  - Recurrent Networks

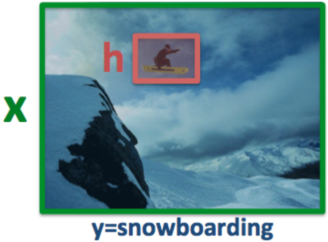


Outline

How to use deep architecture on complex scenes?

- Using full (precise) annotation, *e.g.* BB or segmentation masks
- **BUT:** full annotations expensive [Bearman et al., 2016]
  $\Rightarrow$ **training with weak supervision**

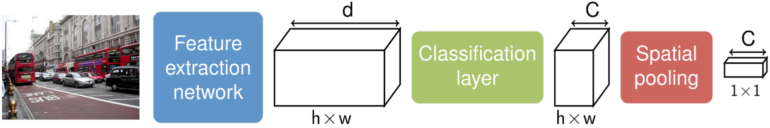




Variable	Notation	Space	Train	Test
Input	$\mathbf{x}$	$\mathcal{X}$	observed	observed
Output	$\mathbf{y}$	$\mathcal{Y}$	observed	unobserved
Latent	$\mathbf{h}$	$\mathcal{H}$	unobserved	unobserved

Weakly supervised learning

- Make learning and recognition more challenging
- Adapt deep architecture

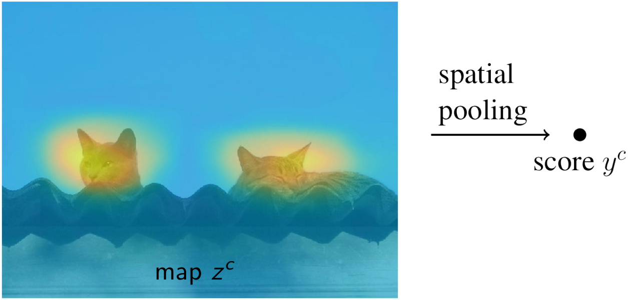


- $h \times w \times C$ tensor: Class Activation Maps (CAM)

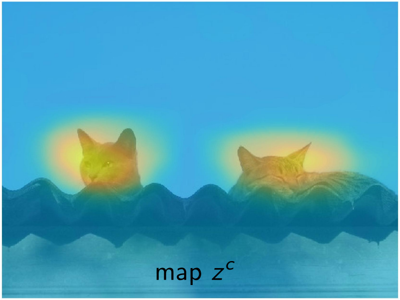
map z^c

Weakly supervised learning

- Make learning and recognition more challenging
- Adapt deep architecture
    - **Pooling function $\Rightarrow$ global label from local predictions**



How to pool?



spatial pooling $\longrightarrow$ • score y^c

Max [Oquab, CVPR15]

$$y^c = \max_{i,j} z^c_{ij}$$

Use 1 region

Average (GAP) [Zhou, CVPR16]

$$y^c = \frac{1}{N} \sum_{i,j} z^c_{ij}$$

Use all regions

Average pooling limitation

- Classifying with all regions
- Not efficient for small objects: lots of "noisy" regions

Max pooling

$$y^c = \max_{i,j} z_{ij}^c \quad (1)$$

- Classifying only with the max scoring region

- Loss of contextual information

Max pooling

$$y^c = \max_{i,j} z_{ij}^c \quad (1)$$

- Classifying only with the max scoring region

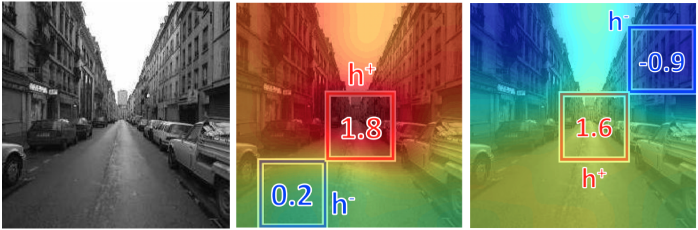
- Loss of contextual information

max+min pooling [Durand et al., 2015]

- **Contribution:** `max+min` **pooling function**

$$y^c = \max_{i,j} z_{ij}^c + \min_{i,j} z_{ij}^c \tag{2}$$

- $\mathbf{h}^+$: presence of the class → high $\mathbf{h}^+$
- $\mathbf{h}^-$: localized evidence of the absence of class: **negative evidence**

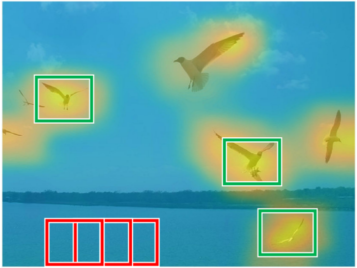


street image **x** — $s(\textbf{street}) = 2$ — $s(\textbf{highway}) = 0.7$

WELDON pooling [Durand et al., 2016]

- Extension of `max+min` pooling
- Using several regions, more robust region selection

k=1



k=3

$$y^c = s^{top}_{k^+}(z^c) + s^{low}_{k^-}(z^c) \tag{3}$$

$$s^{top}_{k^+}(z^c) = \frac{1}{k^+}\sum_{i=1}^{k^+} i\text{-th-max}(z^c) \quad s^{low}_{k^-}(z^c) = \frac{1}{k^-}\sum_{i=1}^{k^-} i\text{-th-min}(z^c)$$

WILDCAT pooling [Mordan et al., 2017a]

- `max+min` pooling:
  - Both types of region are important
  - Complementary information
  - Not the same importance
- Pooling function

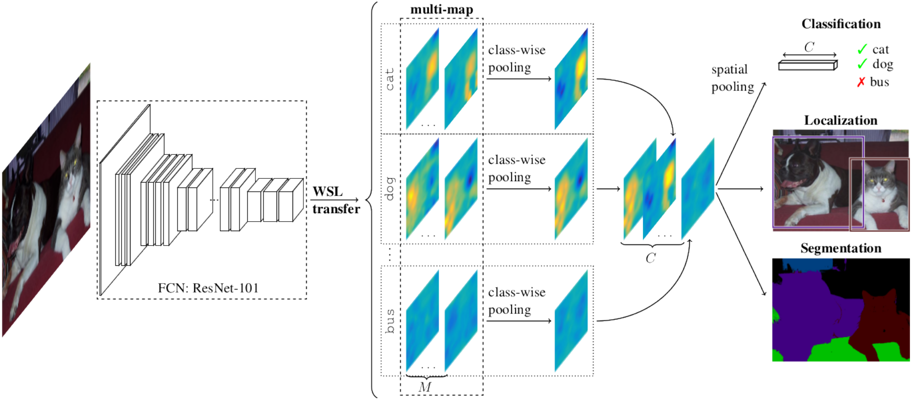
$$y^c = s_{k^+}^{top}(z^c) + \alpha \cdot s_{k^-}^{low}(z^c) \tag{4}$$

  - $\alpha \in [0,1]$: trade off parameter

Pooling	k^+	k^-	α
`max`	1	0	0
GAP	n	0	0
`max+min`	1	1	1
WELDON	k	k	1

Negative Evidence Models: Conclusion

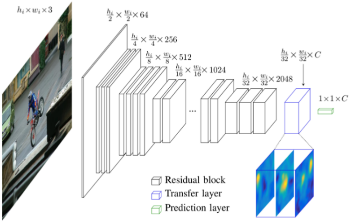
- Global archi applicable for weakly supervised localization & segmentation
  - Extended PAMI version [Durand et al., 2019]



- State-of-the-art for many image classification datasets
- **Structured output prediction:** AP ranking

Weakly Supervised Object Detection: Task 3.2

- ⊕ Use start/end drone detection presence in video stream (GPS-RTK)
  - Improving annotation granularity with GPS + optronic system orientation $\Rightarrow$ limiting drone RoI search
- Evaluation of WSL models : needs test annotations!
  - Using full supervision for a small data sub-set?
  - Localization accuracy with WSL: relatively coarse
    - OK or object proposals?



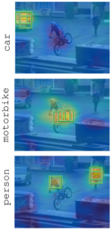
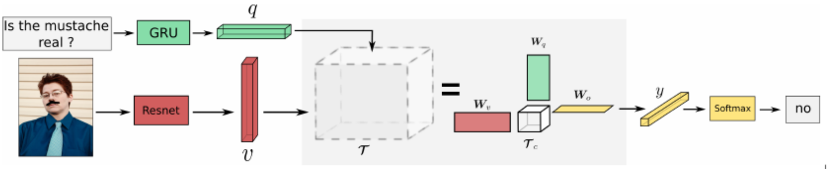


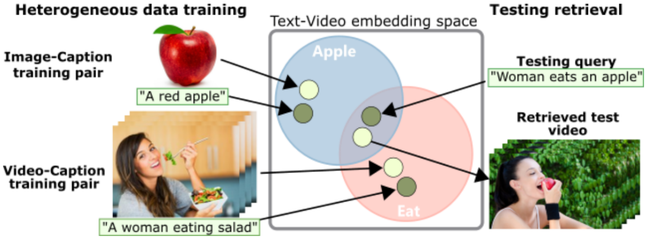
Image size	Size before pooling	k^+, k^-
224 × 224	7 × 7	5
280 × 280	9 × 9	10
320 × 320	10 × 10	20
374 × 374	12 × 12	30
448 × 448	14 × 14	50
560 × 560	18 × 18	75
747 × 747	24 × 24	100

Outline

Deep Multi-modal Fusion

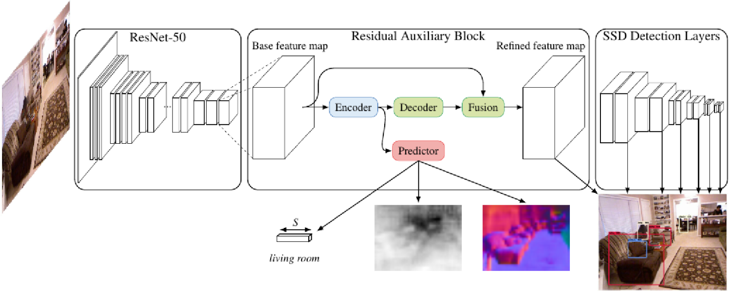
- Fusion at intermediate representation levels *vs* early / late fusion
- Used for VQA and VRD [Ben-younes et al., 2017, Ben-younes et al., 2019]
  - Missing / Incomplete data [Miech et al., 2018]





Multi-task Learning [Mordan et al., 2018b]

- Multi-task: Primary task (focus) $\neq$ Auxiliary tasks (help)
  - Related to privileged information (LUPI) [Vapnik and Vashist, 2009]



- Can be leveraged for combining detection with complementary info (spectral target signature, device orientation)
  - Available data at test time?

Thank you for your attention !

Questions ?

References I

[Azizpour et al., 2016] Azizpour, H., Razavian, A. S., Sullivan, J., Maki, A., and Carlsson, S. (2016).
Factors of transferability for a generic convnet representation.
IEEE Trans. Pattern Anal. Mach. Intell., 38(9):1790–1802.

[Bearman et al., 2016] Bearman, Russakovsky, Ferrari, and Fei-Fei (2016).
What's the Point: Semantic Segmentation with Point Supervision.
ECCV.

[Ben-younes et al., 2017] Ben-younes, H., Cadène, R., Cord, M., and Thome, N. (2017).
MUTAN: multimodal tucker fusion for visual question answering.
In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2631–2639.

[Ben-younes et al., 2019] Ben-younes, H., Cadene, R., Thome, N., and Cord, M. (2019).
Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection.
In *AAAI*.

[Durand et al., 2015] Durand, T., Thome, N., and Cord, M. (2015).
MANTRA: Minimum Maximum Latent Structural SVM for Image Classification and Ranking.
In *International Conference on Computer Vision (ICCV)*.

[Durand et al., 2016] Durand, T., Thome, N., and Cord, M. (2016).
WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks.
In *Computer Vision and Pattern Recognition (CVPR)*.

[Durand et al., 2019] Durand, T., Thome, N., and Cord, M. (2019).
Exploiting negative evidence for deep latent structured models.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2):337–351.

[Girshick et al., 2014] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014).
Rich feature hierarchies for accurate object detection and semantic segmentation.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

References II

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).
Imagenet classification with deep convolutional neural networks.
In *Advances in neural information processing systems*, pages 1097–1105.

[Miech et al., 2018] Miech, A., Laptev, I., and Sivic, J. (2018).
Learning a Text-Video Embedding from Imcomplete and Heterogeneous Data.
In *arXiv*.

[Mordan et al., 2017a] Mordan, T., Durand, T., Thome, N., and Cord, M. (2017a).
WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Localization and Segmentation.
In *Computer Vision and Pattern Recognition (CVPR)*.

[Mordan et al., 2017b] Mordan, T., Thome, N., Hénaff, G., and Cord, M. (2017b).
Deformable part-based fully convolutional network for object detection.
In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*.

[Mordan et al., 2018a] Mordan, T., Thome, N., Henaff, G., and Cord, M. (2018a).
End-to-end learning of latent deformable part-based representations for object detection.
International Journal of Computer Vision.

[Mordan et al., 2018b] Mordan, T., Thome, N., Hénaff, G., and Cord, M. (2018b).
Revisiting multi-task learning with ROCK: a deep residual auxiliary block for visual detection.
In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 1317–1329.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014).
Two-stream convolutional networks for action recognition in videos.
In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 568–576. Curran Associates, Inc.

[Uijlings et al., 2013] Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., and Smeulders, A. W. M. (2013).
Selective search for object recognition.
International Journal of Computer Vision, 104(2):154–171.

References III

[Vapnik and Vashist, 2009] Vapnik, V. and Vashist, A. (2009).
A new learning paradigm: Learning using privileged information.
Neural Networks.