Natural image understanding

principles, challenges and research outlook

Siniša Šegvić, Ivan Krešo, Marin Oršić and Petra Bevandić
ZEMRIS UniZg-FER

Agenda

- Introduction: natural image understanding
- Convolutional models: operational principles and current performance
- Challenges towards achieving truly intelligent perception
- Prominent research directions and results
- Conclusions

Introduction: image classification

Story of natural image understanding starts with image classification

Consider the problem of discriminating bison from oxen:

[image-net.org]

Hard because we do not know where is the defining object

Especially hard when intra-class variance is large and inter-class variance is small

Introduction: a rule-based approach?

We could try to solve image classification by a rule-based system:

1. concentrate on image regions projected from four legged animals
2. oxen have longer horns or no horns at all
3. bison are dark brown, etc, etc

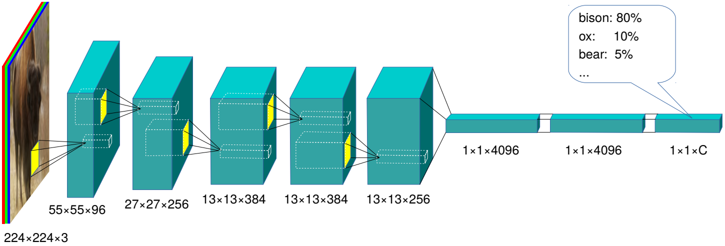
It's explainable and neat, but nobody succeeded to make that work!

[image-net.org]

Convolutional models: architecture

Instead of handcrafted features and rules, we prefer **training on data**

Consider a deep convolutional model trained in an end-to-end fashion:



- **input**: image; **output**: distribution over known classes
- **structure**: a succession of convolutions and poolings
  - gradual decrease of resolution and increase of the semantic depth
  - recent architectures: O(10^3) classes, O(10^2) layers, O(10^6) parameters, O(10^9) multiplications for a 224x224 image!
- **fitness criterion**: (average) log probability of the correct class

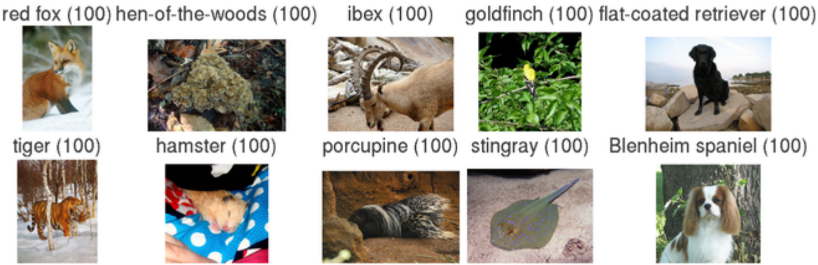
Convolutional models: ImageNet

A lot of data is required to learn O(10^6) parameters!

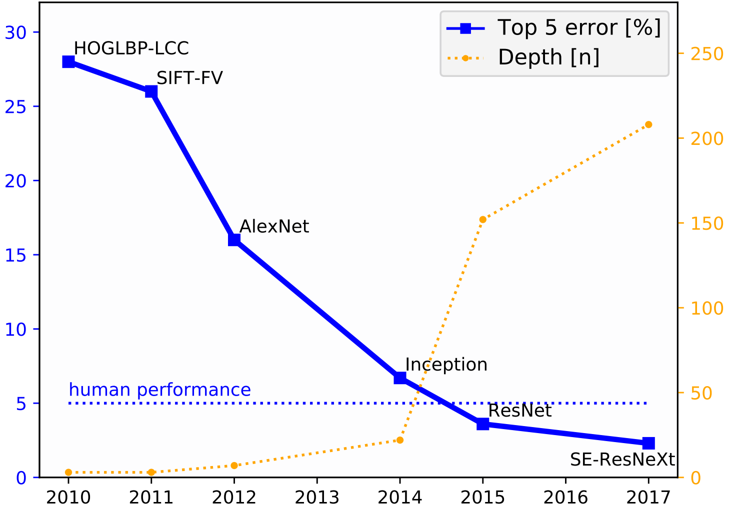
- hence, deep models for vision are usually trained on ImageNet

One of the most popular vision datasets [russakovsky15ijcv]

- we focus on the ILSVRC subset: 10^6 images, 10^3 classes
- annual challenges: classification, localization, detection in video
- fine-grained animals, objects, materials, sports, dishes...



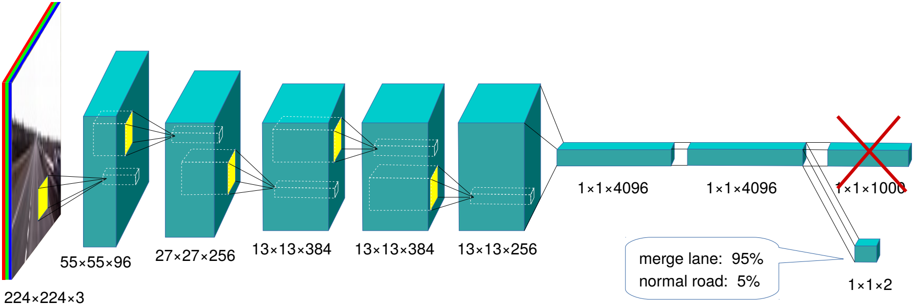
Convolutional models: ImageNet performance



Convolutional models: knowledge transfer

A deep classification model can be fine-tuned for another (easier) task:

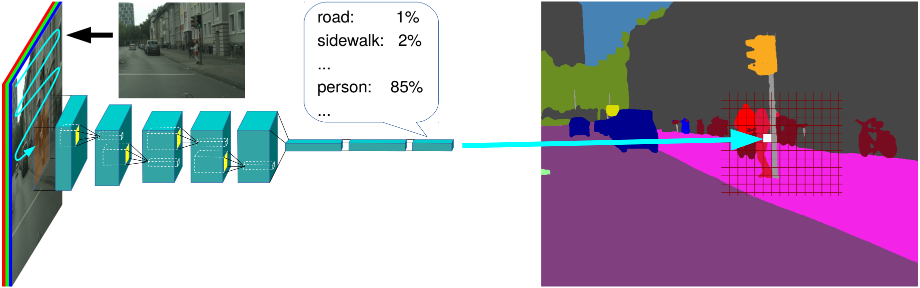
- cut-off the last few layers
- connect the remaining layers with a back-end for the new task
- train the resulting model on new images
- inherited layers are well-trained so we can train with less data (few thousands images)



Convolutional models: sliding window

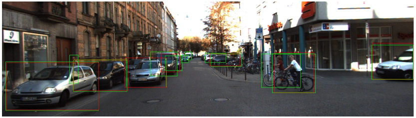
Dense prediction achieved by classifying image patches one by one:

- analyze each image patch in a **sliding window** fashion
    - a little bit more complicated than that in practice
- each patch corresponds to one pixel of the semantic map
- semantic groundtruth allows fine-tuning in an end-to end fashion
- each pixel becomes one component of the fitness criterion



Convolutional models: dense prediction tasks

- Object localization (and tracking):
  - detect objects...
  - ...and indicate their location



- Semantic segmentation:
  - label all image pixels...
  - ...with semantic classes

- Stereoscopic reconstruction:
  - label all image pixels...
  - ... with metric distances

Demonstrations are available at:

[1] http://www.zemris.fer.hr/~ssegvic/pubs/orsic18Linthescher.mp4

[2] http://www.zemris.fer.hr/~ssegvic/pubs/kreso17semseg_stuttgart_00.mkv

[3] http://www.zemris.fer.hr/~ssegvic/pubs/orsic17stereo_kitti19.mp4

Challenges: false positives due to context

Current models tend to produce false positive detections due to context

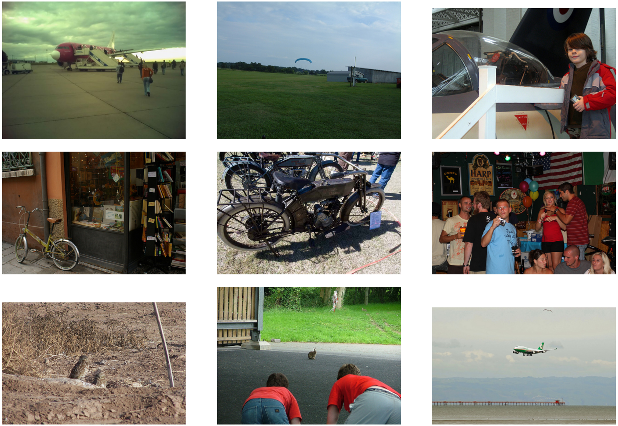
- good performance likely due to recognition of **easy context**

Results on Pascal VOC 2012 (confident TP, high FP, low FN):

airplane

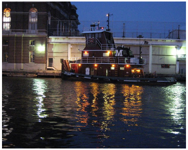
bicycle

bird



Challenges: false positives due to context (2)

boat



bottle


bus



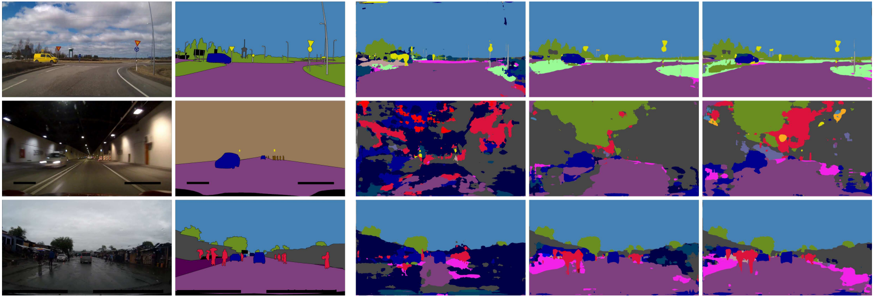
car



Challenges: cross-dataset generalization

Often our models generalize well only within dataset

For instance, images from the novel WildDash dataset (left) fool most models trained on popular datasets such as Cityscapes or Vistas (right)



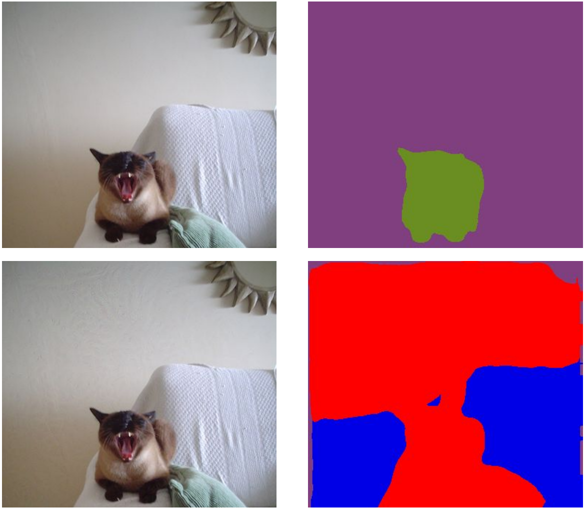
[zendel18eccv]

Conclusion: models tend to *overfit* to **dataset specifics**

- camera, weather, environment, climate, ...

Challenges: Adversarial examples

Imperceptible perturbations may invalidate prediction [szegedy14iclr]:



[kreso18ep]

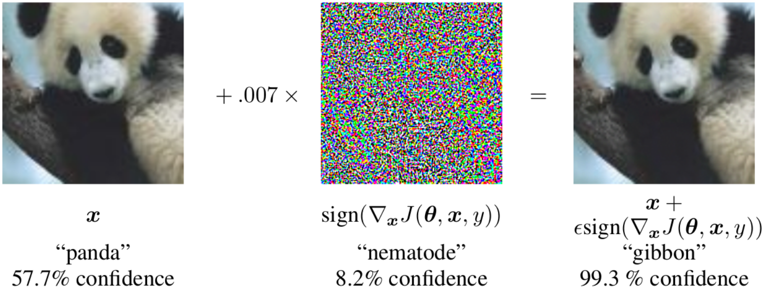
Challenges: Adversarial examples (2)

Adversarial perturbation:

$$\delta = \arg\min_{\delta} p(Y = y_i | \mathbf{x}_i + \delta, \Theta)$$

Adversarial example: $\mathbf{x}_i + \delta$

Existence of adversarial examples suggests that current vision systems are *free-riding* on **easy features** while ignoring the gist of the scene



[goodfellow15iclr]

Solutions: approaches

Presented challenges suggest that state-of-the-art systems:

- are unable to comprehend limits of their expertise
- under-achieve by relying on **easy** but **non-robust** features
  - humans also jump to conclusions: e.g. a person without a wedding ring is considered single
  - such inferrence is not appropriate for mission critical systems

Prominent ways for getting closer to truly intelligent artificial vision:

- improve training data [rob18cvpr]
- detect out-of-distribution (OOD) input [bevandic18arxiv]
- improve the training process [tsipras18arxiv]

Solutions: better data (1)

Idea: improve cross-dataset generalization by training on multiple datasets

Robust vision challenge at CVPR'18, semantic segmentation contest:

- train the common segmentation model on four datasets
  - Cityscapes, WildDash, KITTI, ScanNet
- submit the same model to four benchmarks
- the model has to predict 19 "driving" and 20 "indoor" classes
- not known: will additional classes cause performance drop?
- our submission ranked #2 and received the **runner-up prize**

Solutions: better data (2)

www.robustvision.net/leaderboard.php?benchmark=semantic

Semantic Segmentation Leaderboard

	Method	KITTI (Detailed subrankings)	ScanNet (Detailed subrankings)	Cityscapes (Detailed subrankings)	WildDash (Detailed subrankings)
1	MapillaryAI_ROB	1	1	1	1
	In-Place Activated BatchNorm for Memory-Optimized Training of DNNs [Project page] - Submitted by Peter Kontschieder (Mapillary Research)				
2	LDN2_ROB	3	2	2	3
	Ladder-style DenseNets for Semantic Segmentation of Large Natural Images [Project page] - Submitted by Ivan Krešo (University of Zagreb, Faculty of Electrical Engineering and Computing)				
3	IBN-PSP-SA_ROB	2	3	3	4
	Submitted by Anonymous				
4	AHiSS_ROB	5	8	5	2
	Training of Convolutional Networks on Multiple Heterogeneous Datasets for Street Scene Semantic Segmentation [Project page] - Submitted by Panagiotis Meletis (Eindhoven University of Technology)				
5	VENUS_ROB	4	4	4	9
	VENUS-Net for RobustVision - Submitted by Anonymous				
6	AdapNetv2_ROB	5	5	6	7
	Submitted by Anonymous				
7	VlocNet++_ROB	7	5	10	5
	Submitted by Anonymous				
8	APMoE_seg_ROB	8	7	7	10
	Pixel-wise Attentional Gating for Parsimonious Pixel Labeling [Project page] - Submitted by Shu Kong (University of California at Irvine)				
8	BatMAN_ROB	8	10	9	6
	Submitted by Robert Peharz (University of Cambridge)				
10	GoogLeNetV1_ROB	10	12	7	8
	Baseline - Submitted by Jonas Uhrig (ROB Team)				

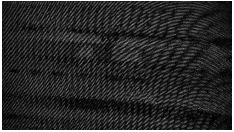
[rob18cvpr]

Solutions: better data (3)

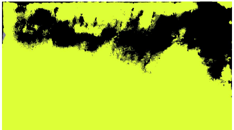
Conclusions from our ROB experiments [kreso18arxiv]:

- "foreign" predictions rare, no performance drop wrt standard setup:


- foreign predictions occur in negative WildDash images:



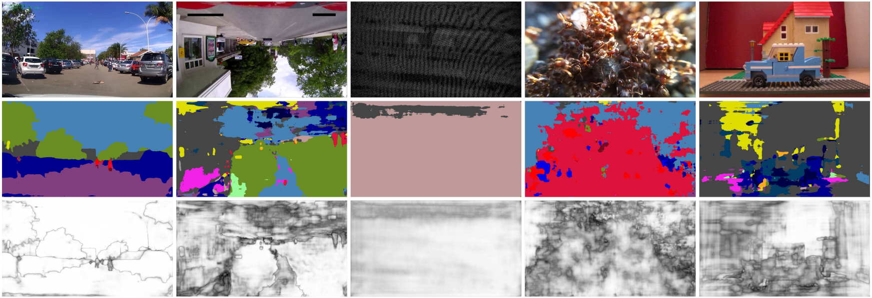




SOLUTIONS: DETECT OOD INPUT

Idea: intercept false positive predictions by detecting OOD input

- unfortunately, most public datasets do not include OOD samples
- WildDash is currently the only public semantic segmentation dataset with "negative" (OOD) images

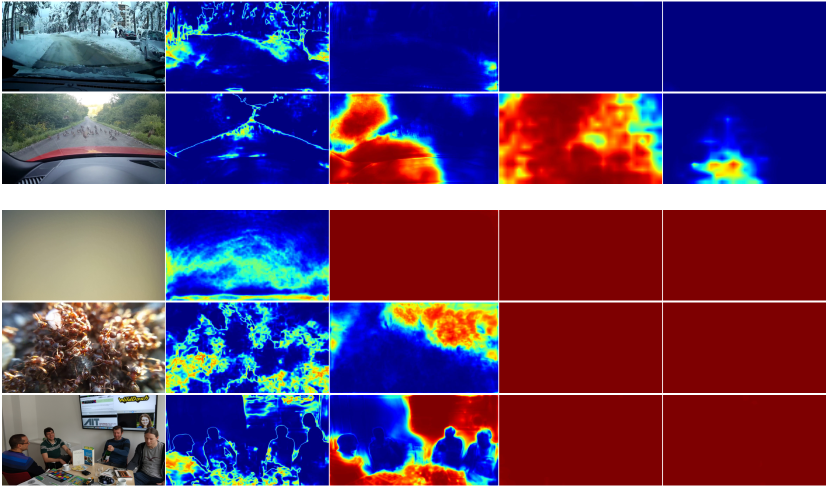


[zendel18eccv]

The creator of WildDash will give a talk here in Zagreb on October 18 (ACROSS workshop, UniZg-FER, register at: `http://goo.gl/1UK1vz`)

Solutions: detect OOD input (2)

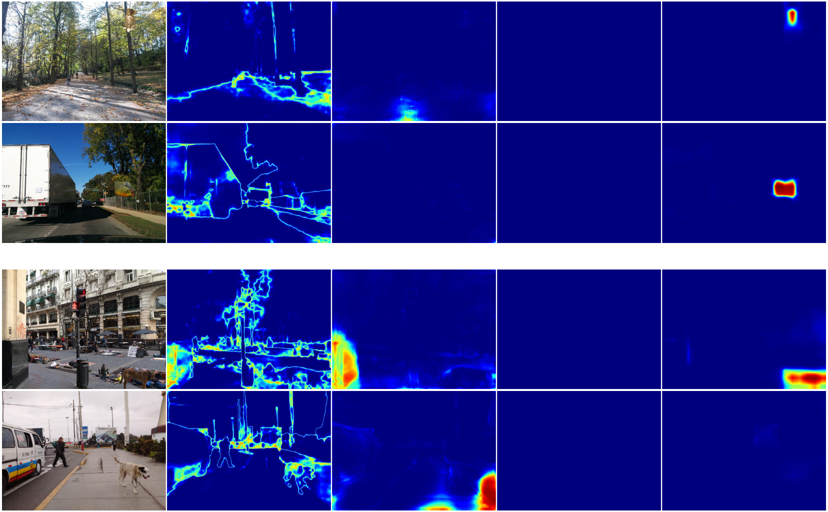
Our approach: detect outliers by training a model which distinguishes in-distribution data from some broad dataset (e.g. ImageNet)



[bevandic18arxiv]

Solutions: detect OOD input (3)

Our approach succeeds to detect animals by using the ImageNet as the positive class (outliers) and Vistas as the negative class (inliers):



Solutions: better training

Classic training: maximize log-likelihood of the data

$$\hat{\theta} = \arg\min_{\theta} \mathbb{E}_{\mathcal{D}} - \log P(y_i | \mathbf{x}_i, \theta)$$

Adversarial training: maximize log-likelihood of worst-case data

$$\hat{\theta} = \arg\min_{\theta} \mathbb{E}_{\mathcal{D}} \max_{\delta} - \log P(y_i | \mathbf{x}_i + \delta, \theta)$$

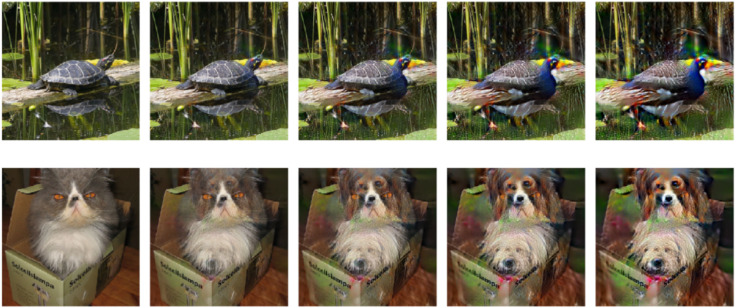
Adversarial training has been introduced years ago as a regularization technique [szegedy14iclr]

Recent results show additional benefits may be achieved by more aggressive search for the worst-case perturbation [madry18iclr]

SOLUTIONS: BETTER TRAINING (2)

Result 1: robustness to adversarial examples [madry18iclr]

Result 2: interpretable gradients [tsipras18arxiv]



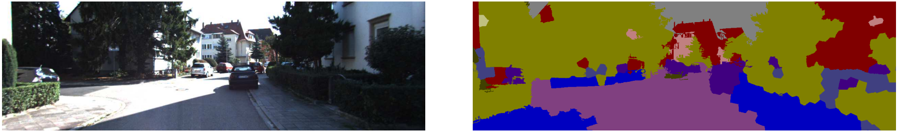
[tsipras18arxiv]

Downside 1: robust features currently do not lead to better performance

Downside 2: much more computationally intensive than classic training

Conclusion: dramatic improvement

Natural image understanding has experienced unprecedented progress in last few years:



2015 [ros15wacv]

2017 [kreso17cvrsuad]

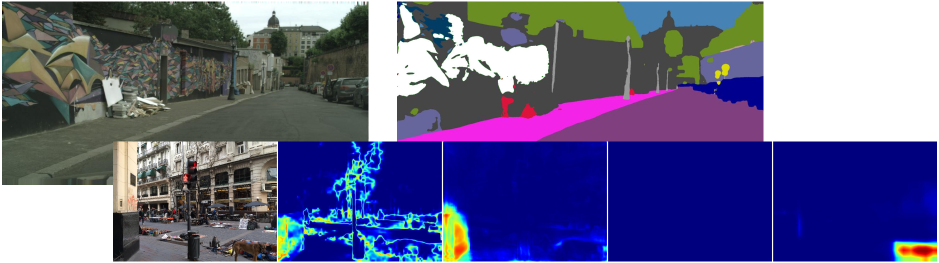
Conclusion: future work

Current failures are often due to models exploiting easy ways to perform their job

- maybe this is a sign of intelligence :-)

Current research strives to entice models to stop avoiding hard work:

- more involved datasets (e.g. WildDash, ROB 2018)
- open-set recognition, out-of-distribution awareness
- more involved optimization (e.g. adversarial training)



Conclusion: discussion

Thank you for your attention!

Questions?

Presented results have been achieved at projects Datacross (ERDF) [1], SafeTram (Končar i ERDF) [2] i MultiClod (HRZZ).

[1] https://across-datascience.zci.hr/datacross

[2] https://www.koncar-institut.hr/en/content-center/projects/safetram

[3] http://multiclod.zemris.fer.hr

The Titan X used in experiments was donated by NVIDIA Corporation.