DeepStat: learning statistics from images using deep learning

First attempts

Joep Burger in collaboration with Tim de Jong, Lyana Curier, Marc Ponsen, Dewi Peerlings, Han van Leeuwen, Krzysztof Cybulski and Jan van den Brakel

UNECE project on the use of ML in OS – WP2 Imagery – April 9, 2020

Background

– Sustainable Development Goals (UN 2015)

– MAKSWELL
  - EU-funded project
  - Harmonize indicators on sustainable development and well-being
  - WP3 – measurement of regional poverty
    - Tax register
    - Earth observation

Research questions

- Can we learn poverty from images?
- What is the effect on prediction quality of
  - sample size?
  - sampling design (including non-probability sample)?
  - grid size (image height × width)?
  - remote sensing indices (image depth)?
- What features are learned?

Statistics Netherlands has the labels needed for this supervised ML task

Data

- Input images

Data source	Resolution	Bands	Color depth	Available since
Aerial images	0.25 m	3	8-bit	2016
Satellite images (Landsat 8)	30 m	11	16-bit	2013

- Grid (coordinate system EPSG:28992)
  - 1 ha (100 m × 100 m) – 400 × 400 aerial pixels or 4 × 4 Landsat pixels
  - 25 ha (500 m × 500 m) – 2000 × 2000 aerial pixels or 17 × 17 Landsat pixels
- Output labels
  - income-related poverty indices
  - open grid statistics

https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/kaart-van-100-meter-bij-100-meter-met-statistieken

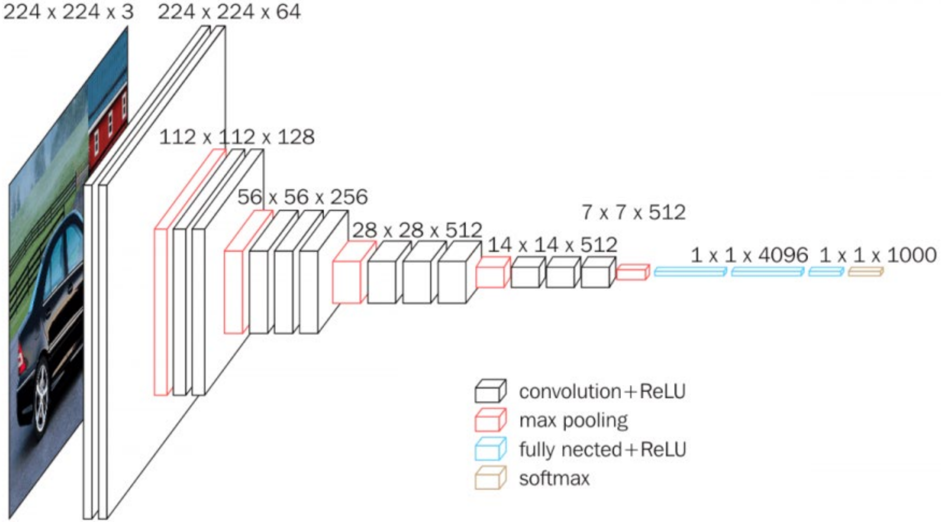
Data preparation

- NL 40k km^2
  - 4 mln 1-ha squares
  - 160k 25-ha squares
- 300k 1-ha squares after simple random sample and linking grid statistics (unknown/unreliable/undisclosed → bias)
- Data augmentation (rotation, shift, zoom, shear, flip)
- 4-class label (quartiles)
  - Number of inhabitants
  - Number of households
  - Number of dwellings

Examples

Convolutional Neural Network (CNN)

- VGG16
- ResNet
- Adjust final layer
- Transfer learning



Preliminary results

			Accuracy	
Output label	# classes	train/test	VGG16	ResNet
Number of households quantile	4	20k/10k	0.68	
Number of inhabitants quantile	4	20k/10k	0.65	
Number of dwellings quantile	4	20k/10k		0.69 (0.58*)
	4	30k/10k		0.73
	4	50k/20k		0.74
	2	20k/10k		0.87 (0.74*)

*min-max normalized

Conclusions

– Statistical information can be learned from images
– CNNs require specialized IT hardware and skills
  - Input is big
  - Algorithms evolve quickly
  - Many (hyper)parameters to estimate/tune
  - Output is privacy-sensitive

Next steps

1. Move to secure environment
   - Link income-related poverty indices
2. Use ordinal loss function
3. Optimize architecture and (hyper)parameters
4. Quantify effect on prediction quality of
   - sample size
   - sampling design
   - grid size
   - remote sensing indices
5. Visualize learned features
6. Compare or combine aerial with satellite images and traditional ML

Acknowledgments

UNECE

- Abel Coronado
- Jimena Juarez
- Claude Julien
- …

CBS

- Tim de Jong
- Lyana Curier
- Marc Ponsen
- Dewi Peerlings
- Han van Leeuwen
- Krzysztof Cybulski
- Jan van den Brakel