Enhancing the performance of indoor-outdoor image classifications using features extracted from depth-maps

George Ciubotariu
Babeș-Bolyai University

WeADL 2021 Workshop

The workshop is organized under the umbrella of WeaMyL, project funded by the EEA and Norway Grants under the number RO-NO-2019-0133. Contract: No 26/2020.

Working together for a green, competitive and inclusive Europe

Contents

Introduction

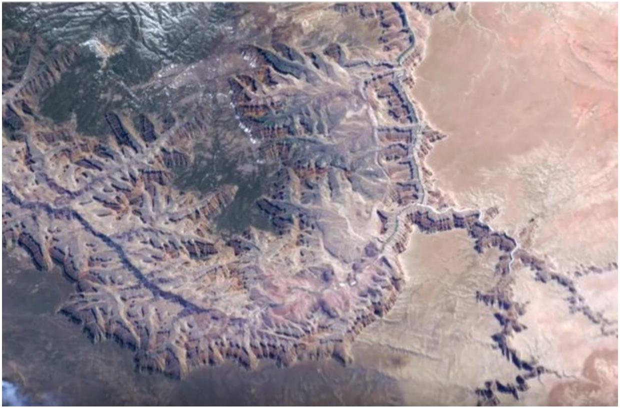

Figure: A picture taken from space

Introduction

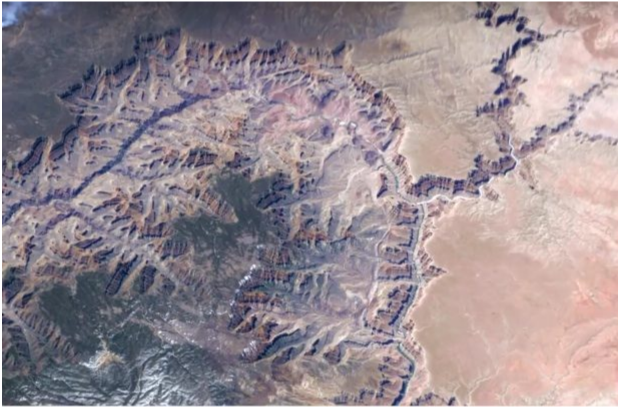

Figure: The same picture, but flipped upside down

Introduction

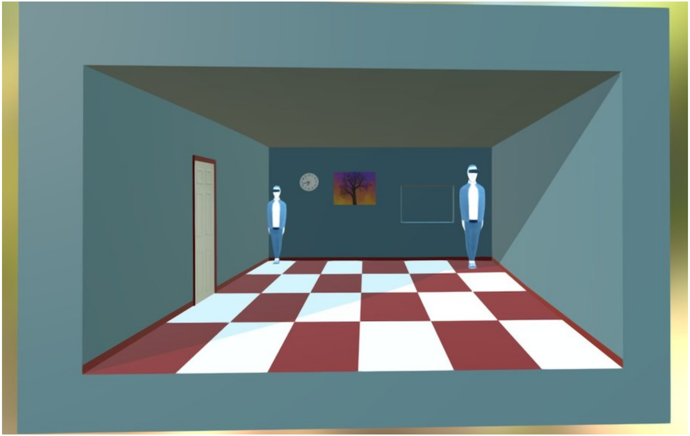

Figure: An illusion of depth

Research Questions and Original Contributions

- **RQ1**: *How relevant are depth maps in the context of indoor-outdoor image classification?*
  - Unsupervised learning based analysis on DIODE dataset for indoor-outdoor classification
    - t-SNE clustering support for further supervised investigations
- **RQ2**: *To what extent does aggregating visual features into more granular sub-images increase the performance of classifiers?*
  - Supervised learning based classification for supporting the unsupervised approach
    - Multilayer Perceptron (MLP) classifier tested to confirm hypothesis
- **RQ3**: *How correlated are the results of the unsupervised based analysis and the performance of supervised models applied for indoor-outdoor image classification?*
  - Comparative analysis on image features aggregation

Computer Vision (CV) and Deep Learning (DL)

Most recent work implement **Convolutional Neural Networks** (CNNs) in dense visual tasks such as *Semantic Segmentation* (SS) or *Depth Estimation* (DE).

- [ZWZ+20] **Split-Attention Network** (ResNeSt)
  - efficient network that outperformed other similar models in what regards both computational costs and performance
  - the model introduced a new split-attention block for dense task prediction.
- [LRSK19, RBK21] **Dense Prediction Transformers** (DPT)
  - model that leverages visual transformers instead of convolutions.
  - its results outperform ResNeSt models that have previously been considered state-of-the-art.

Vision Transformers for Dense Prediction (DPT)

Model	Image resolution	# extracted features after encoder	# extracted features after decoder
Depth Estimation / **Semantic Segmentation**	384×384	49152	12582912

Table: DPT architectures details

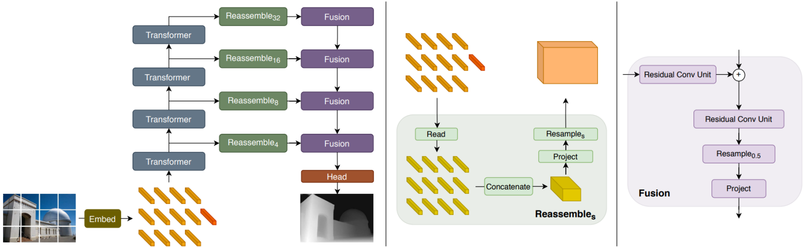

Figure: DPT architecture

DIODE (Dense Indoor and Outdoor DEpth)

- Data has been collected with a **FARO Focus S350**
- It consists of 27858 1024$\times$768 **RGB-D** images
- Photos have been taken both at daytime and night, over several seasons (summer, fall, winter)

Apart from RGB-D images, DIODE dataset also provides us with normal maps that could further enhance the learning of depth and vice-versa

DIODE (Dense Indoor and Outdoor DEpth)

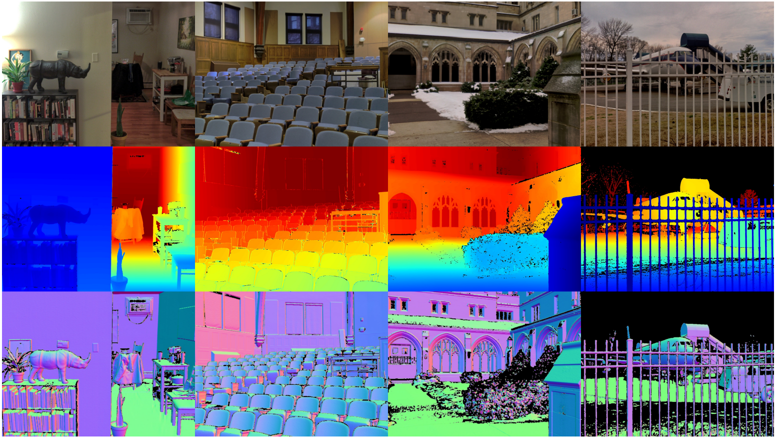

Figure: Sample images from DIODE dataset

DIODE Structure

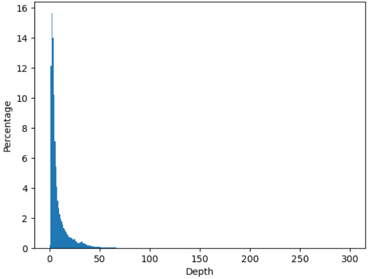

Figure: Histogram of depth values frequency (%) for the whole train set

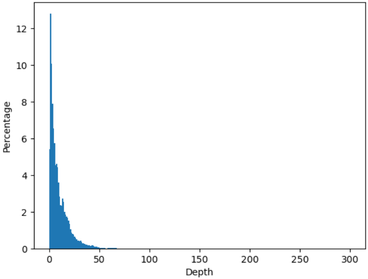

Figure: Histogram of depth values frequency (%) for the whole validation set

DIODE Structure

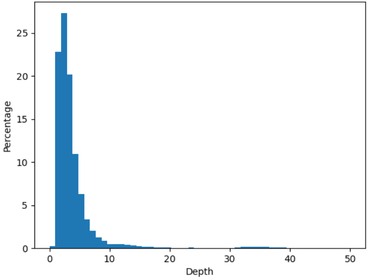

Figure: Histogram of depth values frequency (%) for indoor train set

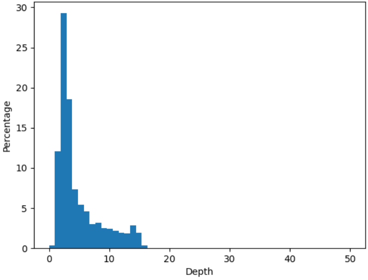

Figure: Histogram of depth values frequency (%) for indoor validation set

DIODE Structure

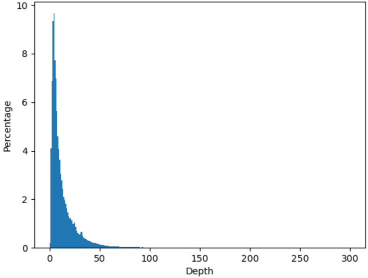

Figure: Histogram of depth values frequency (%) for outdoor train set

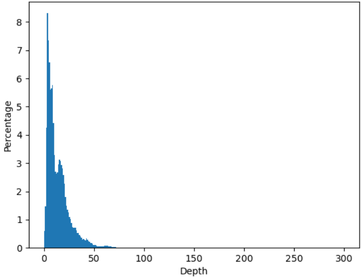

Figure: Histogram of depth values frequency (%) for outdoor validation set

Unsupervised Learning Approach for Analysing the Data

- *3D t-SNE* unsupervised clustering
  - used for *non-linear* dimensionality reduction
  - able to uncover more useful patterns in data
  - uses *Student t-distribution* to better disperse the clusters
- *data normalization* with the **inverse hyperbolic sine (asinh)**
  - increased sensitivity to particularly small and large values
- parameters used
  - **perplexity** of 20
  - **learning rate** of 3.0
    - for a slower converging but finer learning curve
  - 1000 **iterations**

Relevance

Unsupervised learning-based analysis provide useful insight about data organization and features' importance.

Automatic Feature Extraction

1. **aggregating RGB from sub-images**
   - $3 \cdot k$ dimensional vector ($k = 1, 4, 16$)
   - average RGB values for each sub-image
2. **aggregating RGBD from sub-images**
   - $4 \cdot k$ dimensional vector ($k = 1, 4, 16$)
   - average RGBD values for each sub-image
3. **features from DPT encoder/decoder**
   - trained for SS
   - trained for DE

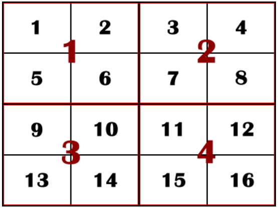

Figure: Structure of image splits

Deep Learning Tasks

- Indoor-Outdoor Classification
- Semantic Segmentation
- Depth Estimation

Features Extracted from DL models

- DPT trained for Semantic Segmentation

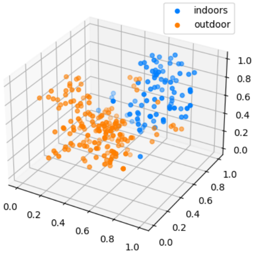

Figure: t-SNE of DPT encoder extracted features for SS

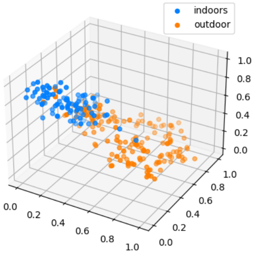

Figure: t-SNE of DPT decoder extracted features for SS

Features Extracted from DL models

- DPT trained for Depth Estimation

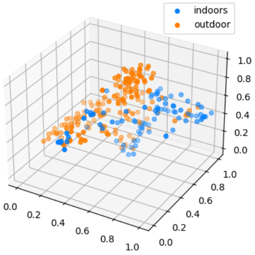

Figure: t-SNE of DTP encoder extracted features for DE

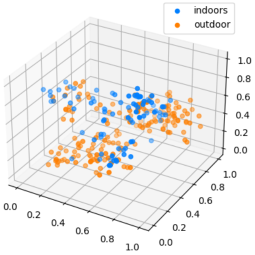

Figure: t-SNE of DTP decoder extracted features for DE

Features extracted aggregating RGB and RGBD values

- no splits

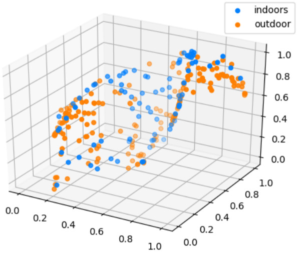

Figure: t-SNE for RGB without splits

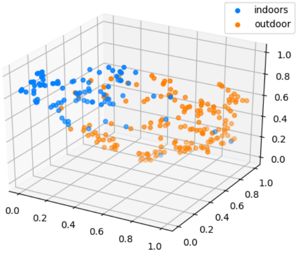

Figure: t-SNE for RGB-D without splits

Features extracted aggregating RGB and RGBD values

- 4 splits

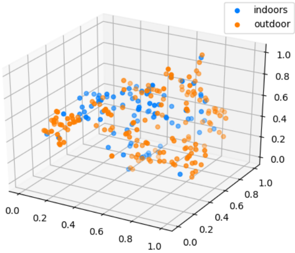

Figure: t-SNE for RGB with 4 splits

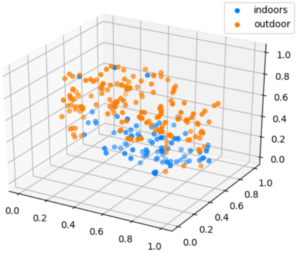

Figure: t-SNE for RGB-D with 4 splits

Features extracted aggregating RGB and RGBD values

- 16 splits

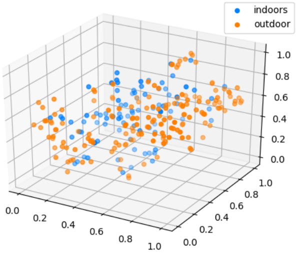

Figure: t-SNE for RGB with 16 splits

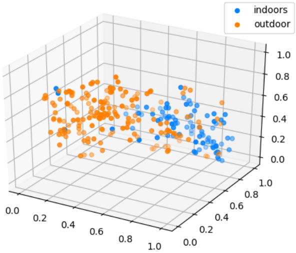

Figure: t-SNE for RGB-D with 16 splits

Supervised Learning Results

Features	# Splits	Accuracy	AUC	Specificity	Recall
MLP RGB	1	0.692±0.077	0.525±0.056	0.980±0.028	0.070±0.121
	4	0.688±0.064	0.517±0.022	**0.989**±0.014	0.046±0.049
	16	0.669±0.049	0.545±0.048	0.912±0.068	0.163±0.136
MLP RGBD	1	**0.880**±0.039	0.858±0.041	0.898±0.058	0.817±0.081
	4	0.876±0.043	**0.862**±0.044	0.894±0.046	**0.829**±0.063
	16	0.838±0.044	0.826±0.053	0.848±0.060	0.804±0.099
DPT encoder DE	1	0.823±0.131	0.831±0.076	0.812±0.185	0.850±0.069
DPT encoder SS	1	**0.953**±0.027	**0.944**±0.030	**0.974**±0.031	**0.915**±0.053

Table: Results of indoor-outdoor supervised classification on DIODE dataset

Best two performances (AUC)

1. DPT encoder SS.
2. RGBD with 4 splits.

Ongoing Experiments and Future Enhancements

- Identifying features that can be used in both SS and DE
- Identifying other problems that can be solved with adapted DL models
- Architecture Transfer from SS towards DE

Thank you!

Questions?

Bibliography I

Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun.
Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer.
CoRR, abs/1907.01341, 2019.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun.
Vision transformers for dense prediction.
CoRR, abs/2103.13413, 2021.

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander J. Smola.
Resnest: Split-attention networks.
CoRR, abs/2004.08955:1–12, 2020.