Enhancing Deep Semantic Segmentation of RGB-D data with Entangled Forest

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Intelligent Autonomous Systems Laboratory (IAS-Lab), University of Padova, Italy

M. TERRERAN, E. BONETTO, S. GHIDONI

ABSTRACT

Semantic segmentation is a problem which is getting more and more attention in the computer vision community. Nowadays, deep learning methods represent the state of the art to solve this problem, and the trend is to use deeper networks to get higher performance. The drawback with such models is a higher computational cost, which makes it difficult to integrate them on mobile robot platforms. In this work we want to explore how to obtain lighter deep learning models without compromising performance. To do so we will consider the features used in the 3D Entangled Forests algorithm and we will study the best strategies to integrate these within FuseNet deep network. Such new features allow us to shrink the network size without loosing performance, obtaining hence a lighter model which achieves state-of-the-art performance on the semantic segmentation task and represents an interesting alternative for mobile robotics applications, where computational power and energy are limited.

CONTRIBUTIONS

While deep learning networks like **FuseNet** [1] require high computational power (i.e. high-end GPUs), the **3D Entangled Forest Classifier (3DEF)** [2] achieves state-of-the-art performance on indoor semantic segmentation tasks using a standard CPU and powerful hand-crafted features. In this work we investigate how to combine the two approaches:

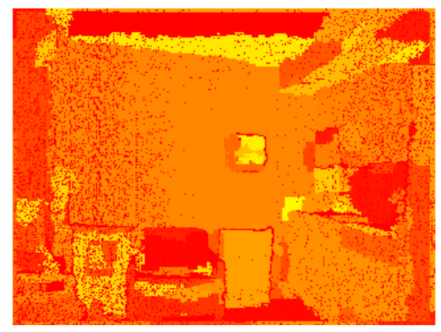
- Analysis of **different configurations to embed hand-crafted features** in a deep learning networks such as FuseNet;
- Study of the **potential of 3DEF hand-crafted features** in DL networks;
- **Possibility to shrink a DL network** without reducing performance by exploiting additional information.

3DEF FEATURES

3DEF classifier relies on a preliminary over-segmentation in clusters and the computation of two types of hand-crafted features:

- **Unary features**, describe local information (e.g. color, geometric moments);
- **Entangled features**, describe relations between clusters to consider also information in the space domain.

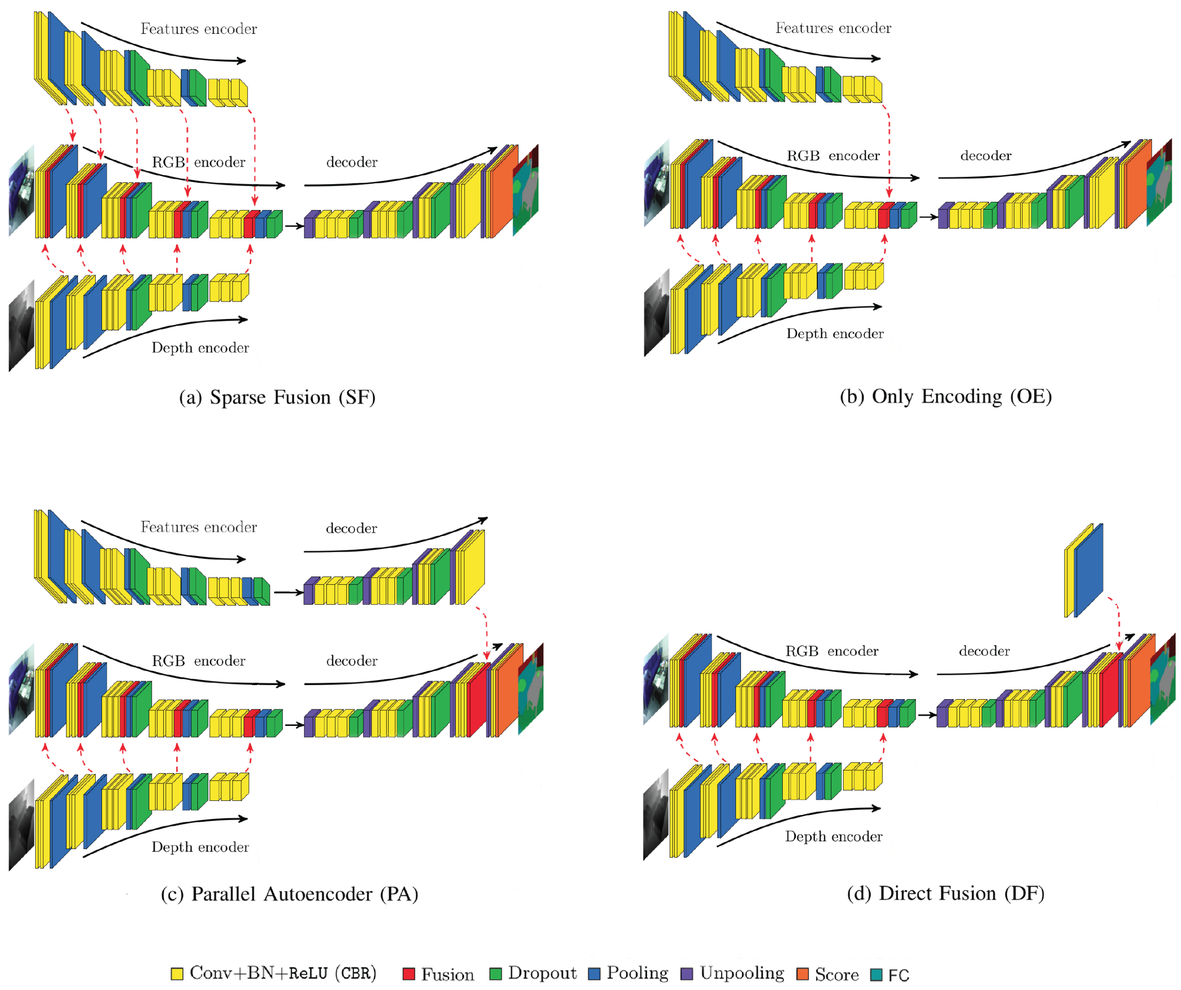
3DEF features represent cluster information, not pixels as in FuseNet architecture; to obtain a similar representation we redefined 3DEF features in a 2D space.



NETWORK CONFIGURATIONS

Starting from FuseNet architecture, we propose **4 different configurations** to embed 3DEF hand-crafted features, investigating both middle-fusion and late-fusion strategies.
For each configuration we also consider a **reduced version**, obtained by removing the intermediate layers, to further analyze the role of 3DEF features.

Network	Original		Reduced	
	Param	Epoch time	Param	Epoch time
Sparse Fusion	58.908	125	6.968	83
Only Encoding	58.908	125	6.968	83
Parallel AE	73.591	143	8.666	94
Direct Fusion	44.224	104	5.269	70
FuseNet	44.173	90	5.218	60



(a) Sparse Fusion (SF) (b) Only Encoding (OE)

(c) Parallel Autoencoder (PA) (d) Direct Fusion (DF)

EXPERIMENTS

Evaluation of all configurations on NYU Depth v2 indoor dataset [3], using both 13 and 40 classes mapping. With 3DEF features we achieve slightly better performance than FuseNet and improve loss function convergence.

Configuration		13-classes			40-classes		
		Global acc.	Mean acc.	IoU	Global acc.	Mean acc.	IoU
SF	UN	76.44	67.52	54.93	68.63	47.08	35.72
	ENT	75.19	65.28	52.83	66.69	42.82	32.95
	ENT+UN	76.86	67.63	55.29	68.60	47.95	36.28
OE	UN	76.29	66.81	54.36	68.76	47.24	35.82
	ENT	76.91	67.61	55.21	68.16	44.65	34.05
	ENT+UN	76.56	67.73	55.03	68.76	46.67	35.54
PA	UN	76.50	67.02	54.74	**69.02**	46.27	36.10
	ENT	76.16	67.31	54.06	67.35	42.88	33.24
	ENT+UN	76.77	67.75	55.27	68.77	47.06	35.89
DF	UN	76.61	67.59	54.94	68.72	46.86	35.86
	ENT	76.17	67.26	54.14	68.13	44.71	34.71
	ENT+UN	**76.92**	**68.21**	**55.49**	69.00	**47.64**	**36.62**
FuseNet [1]		76.40	66.93	54.74	68.76	46.42	35.48

Considering the reduced configurations, large improvements are obtained with middle-fusion strategy and 3DEF Entangled features; therefore, such features can be used to obtain DL models with state-of-the-art performance and a reduced computational load.

Configuration		13-classes			40-classes		
		Global acc.	Mean acc.	IoU	Global acc.	Mean acc.	IoU
SF	UN	67.92	56.95	43.73	58.73	32.11	23.02
	ENT	**71.20**	**60.35**	**47.79**	**62.28**	**36.84**	**26.51**
	ENT+UN	67.83	56.66	43.59	59.04	33.83	24.00
OE	UN	67.77	56.52	43.46	59.03	32.77	23.34
	ENT	70.11	58.42	46.12	60.78	34.44	24.79
	ENT+UN	68.03	57.06	43.80	59.05	33.13	23.71
PA	UN	67.82	56.31	43.38	58.95	32.68	23.30
	ENT	68.38	55.89	43.90	59.68	33.59	23.82
	ENT+UN	68.10	57.11	43.87	59.12	33.14	23.53
DF	UN	67.82	56.36	43.47	59.02	32.71	23.40
	ENT	68.02	54.98	43.10	59.52	32.91	23.44
	ENT+UN	67.76	56.62	43.43	59.31	32.69	23.50
FuseNet reduced		68.01	56.50	43.70	59.02	33.00	23.42

[1] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture", in ACCV (1), pp. 213-228, 2016.
[2] D. Wolf, J. Prankl, and M. Vincze, "Enhancing semantic segmentation for robotics: the power of 3-d entangled forest", IEEE Robotics and Automation Letters, vol. 1, no. 1, pp.49-56, 2016.
[3] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in ECCV, 2012.

Part of this work was supported by MIUR (Italian Minister for Education) under the initiative "Departments of Excellence" (Law 232/2016).