Controlled Visual Hallucination via Thalamus-Driven Decoupling Network for Domain Adaptation of Black-Box Predictors

Yuwu Lu, Chunzhi Liu

South China Normal University

Motivation

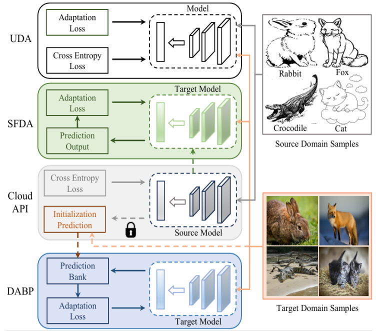
Domain Adaptation of Black-box Predictors (DABP) transfers knowledge from a labeled source domain to an unlabeled target domain, without requiring access to either source data or source model. Common practices of DABP leverage reliable samples to suppress negative information about unreliable samples. However, there are still some problems: 1) Excessive attention to reliable sample aggregation leads to premature overfitting; 2) Valuable information in unreliable samples is often overlooked.

Inspired by Thalamus-driven Decoupling Network (TDN), we propose a novel spatial learning method, named Controlled Visual Hallucination via Thalamus-driven Decoupling Network, to address the existing DABP problems.

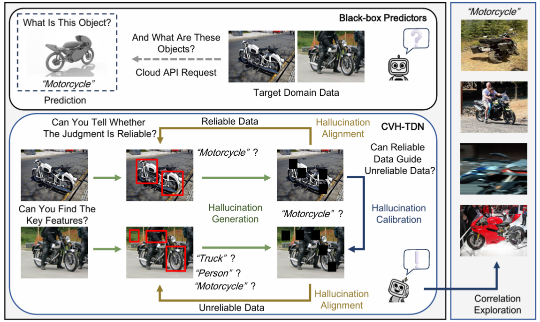
Contributions

- We observe the weaknesses of existing DABP methods and address them by proposing a novel method, called CVH-TDN, that significantly enhances the reasoning ability of model and discrimination capacity of classes.
- Based on the relationship between hallucination and cognition, CVH-TDN contains three parts: Hallucination Generation, Hallucination Alignment, and Hallucination Calibration, aiming to explore the spatial relationships between samples and hallucinations..
- We perform extensive experiments to verify the effectiveness of CVH-TDN, and the results show that it achieves SOTA performance on four benchmarks.

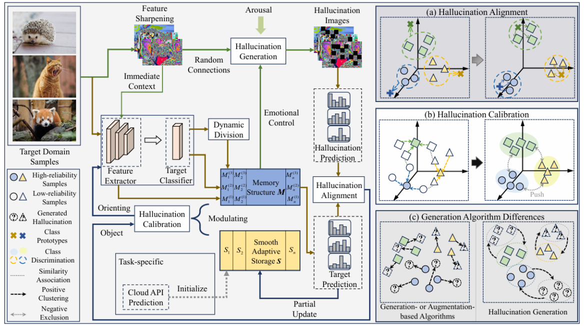
Method



The dotted lines in the figure indicate the operations performed by the cloud API with the source model under different settings. SFDA requires the entire source model to be obtained from the cloud API before training. DABP outperforms SFDA in data privacy protection and portability, simply requiring uploading target data to the cloud API and then downloading predictions before training.



Conceptual figure. The black-box predictors resemble agents with prior knowledge but lack the ability to perform targeted discrimination. HG controls mask formation by modeling the location where hallucinations are pathologically generated, driven by the key cognitive impairments observed in TDN. HA improves feature discrimination by simulating how humans deal with cognitive impairments. HC draws on neurotherapeutic principles to guide unreliable data through reasoning using reliable feature representations.



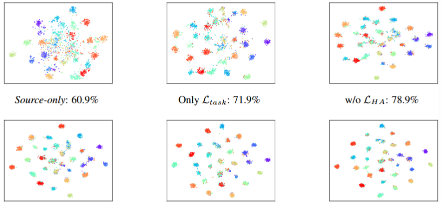
Feature extractor controls masking direction by evaluating knowledge from sharpened images, and the difference between the hallucination generation and other methods is shown in (c). In the hallucination alignment, we reduce the difference between samples and corresponding hallucinations by bidirectional alignment, as shown in (a). As shown in (b), we adopt hierarchical learning with the dynamic division for different types of samples based on spatial information in the hallucination calibration.

Quantitative Ablation Results

Loss $\mathcal{L}_{HA}$	$\mathcal{L}_{HC}$	HG	Office A→D	A→W	D→A	D→W	W→A	W→D	Mean	VisDA Mean
Source only			79.9	76.6	56.4	92.8	60.9	98.5	77.5	48.9
✓			89.3	86.4	73.6	96.5	74.4	99.0	86.5	75.7
	✓		89.8	86.8	74.5	98.5	75.2	99.6	87.4	78.1
✓	✓		94.4	90.4	74.7	98.9	80.1	99.8	89.7	85.1
✓		✓	93.8	92.1	74.5	98.6	77.1	**99.9**	89.3	83.5
	✓	✓	92.9	90.2	73.9	98.6	78.9	99.4	89.0	82.5
✓	✓	✓	**96.4**	**92.8**	75.6	**98.9**	**81.0**	99.6	**90.7**	**90.6**
$\mathcal{L}_{HA}$	$\mathcal{L}_{HC}^{Hsc}$	$\mathcal{L}_{HC}^{Cfe}$								
✓	✓		94.5	91.9	75.8	98.7	79.7	99.6	90.0	88.7
✓		✓	94.7	92.3	**76.1**	98.7	80.1	99.6	90.3	87.1

Qualitative Ablation Results

Scatter Plot :



Source-only: 60.9% — Only $\mathcal{L}_{task}$: 71.9% — w/o $\mathcal{L}_{HA}$: 78.9%

w/o $\mathcal{L}_{HC}$: 77.1% — w/o HG: 80.1% — *CVH-TDN*: 81.0%

Heat map :