ScenarioNet: An Interpretable Data-Driven Model for Scene Understanding

Zachary A. Daniels, Dimitris N. Metaxas
zad7@cs.rutgers.edu, dnm@cs.rutgers.edu

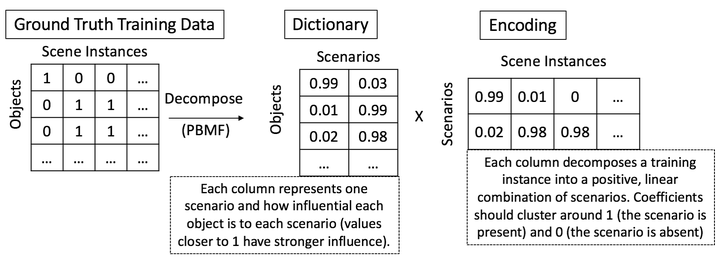
Center for Computational Biomedicine Imaging and Modeling
Department of Computer Science
Rutgers, The State University of New Jersey

Problem Overview

- Many applications require computational agents to reason about the high-level content of real world scene images.
- We introduce **scenarios** as a new way of representing scenes.
- Scenarios are sets of frequently co-occurring objects that satisfy four properties:
  1. Scenarios are composed of one or more objects.
  2. The same object can appear in multiple scenarios, and this should reflect the context in which the object appears, e.g., "screen" plays different roles in the {keyboard, screen, mouse} and {remote control, screen, cable box} scenarios.
  3. Scenes can be decomposed into combinations of scenarios. A bathroom scene instance might decompose into: {shower, bathtub, shampoo} + {mirror, sink, toothbrush, toothpaste} + {toilet, toilet paper}.
  4. Scenarios should be flexible and robust to missing objects. A scenario can be present in a scene without all of its constituent objects being present.
- We introduce **ScenarioNet**: a convolutional neural network architecture that integrates a novel **pseudo-Boolean matrix factorization** in order to learn to identify and recognize scenarios.
- We desire a single model capable of capturing high-level information about scene images at multiple levels including 1) scene classification, 2) multi-object recognition, 3) scenario identification and recognition, and 4) content-based scene image retrieval.
- ScenarioNet is designed to be more interpretable and more efficient than comparable standard CNN architectures.

Identifying Scenarios from Data Using Pseudo-Boolean Matrix Factorization (PBMF)

- We begin by asking "how can scenarios be learned from groundtruth data?". We propose a novel dictionary learning formulation to address this problem.
- Assume we have a ground-truth binary matrix A where each row represents an object and each column represents a training scene instance.
- Element A_{ij} is 1 if object i is present in scene instance j and 0, otherwise.
- To learn an initial dictionary of scenarios, we decompose A into two *approximately* binary matrices:
  1. W: a dictionary of scenarios where each column represents one scenario
  2. H: an encoding matrix that tells us how to decompose a scene as an additive linear combination of scenarios.

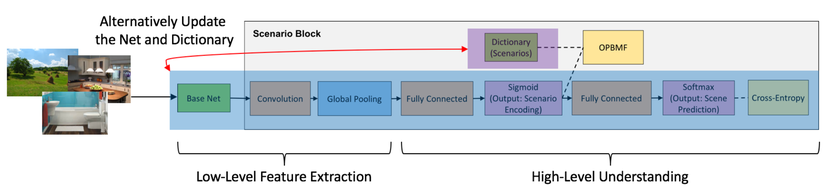


- We formulate PBMF as:

$$\begin{aligned} &\min_{W,H} \|\Omega \bullet (A - \min(WH, 1 + 0.01WH))\|_F^2 \\ &+ \alpha_1 \|W^T W - \mathrm{diag}(W^T W)\|_F^2 + \alpha_2 \|W\|_1 + \alpha_3 \|H\|_1 \\ &\text{s.t. } W \in [0,1], H \in [0,1], \\ &\Omega_{ij} = \max\left(A_{ij} * \left(1 + \log\left(\frac{N_{instances}}{N_{objects}}\right)\right), 1\right) \end{aligned} \quad (1)$$

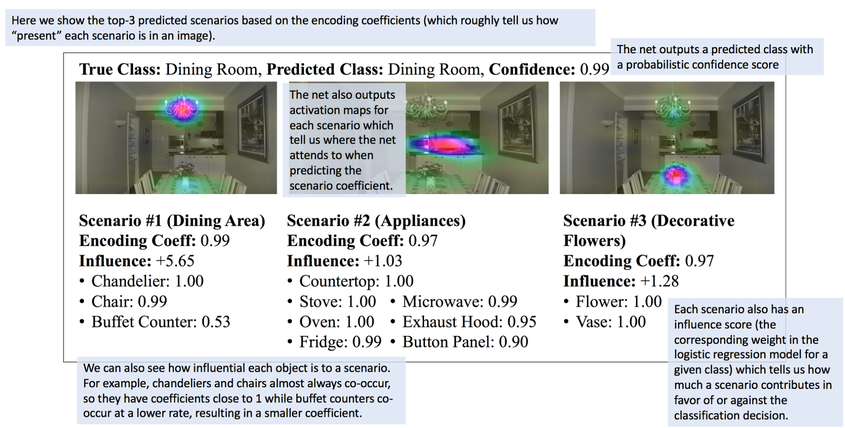
- $\bullet$ denotes element-wise matrix multiplication. The αs represent tradeoff parameters. The first term approximates reconstruction error under Boolean multiplication. The second term promotes orthogonality of the basis vectors leading to more diverse scenarios. The third and fourth terms promote sparsity in the dictionary and encoding matrices. Ω is a weight matrix that decreases the importance of common objects and increases the importance of rare objects during the factorization

The ScenarioNet Architecture



- In practice, we need to recover the encoding matrix H from visual data, and we should adapt the scenario dictionary W to account for noisy estimation of H.
- We propose the ScenarioNet which:
  1. Maps visual data to the scenario encoding space
  2. Finetunes the scenario dictionary based on noisy visual scenario recognition
  3. Performs scene classification in an interpretable manner
- The principle contribution is the **scenario block** which replaces the final fully connected layers of a standard CNN and consists of:
  1. Global pooling layers that identify which parts of an image ScenarioNet attends to when recognizing whether a scenario is present in a given image
  2. Layers that use a PBMF-based loss function to finetune a dictionary of scenarios and predict the presence of each scenario for a given image
  3. Layers equivalent to multinomial logistic regression that use scenarios as low-dimensional features for scene classification.
- The net is trained using alternating minimization:
  1. Finetune the scenario dictionary to account for the noisy prediction of the scenario encoding.
  2. Finetune the neural network to adapt to changes in the scenario dictionary.

Generating Evidence: Interpreting the Output of ScenarioNet



Efficiency

- ScenarioNet has substantially fewer parameters than equivalent base architectures.
  - The final convolutional layers of VGG-16 consist of a 4096-by-4096 matrix followed by a 4096-by-#classes matrix for a total of $4096(4096 + \#classes)$ parameters.
  - ScenarioNet uses a 512-by-#scenarios matrix followed by a #scenarios-by-#classes matrix for a total of $\#scenarios(512 + \#classes)$ parameters.
  - Since #scenarios $<<$ 4096 (we use between $k = 25$ and $k = 70$ scenarios in our experiments), this results in over a 100x reduction in the number of parameters in the final layers, reduces the memory footprint of the *total* net by a factor of ~10, and the net is ~15% faster during testing.

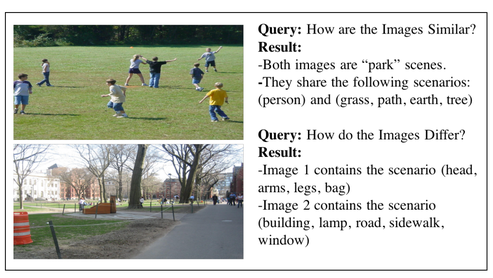
Multi-Object Recognition

- Task: Predict all of the objects present in an image.
- We can recover an *approximate* hypothesis about which objects are present in a scene by recovering the object-scene data matrix (A) using the learned scenario dictionary (W) and predicted encoding matrix (H): $A \approx WH$. This recovery gives us a list of the *possible* objects in a scene, but this recovery is not noiseless since the factorization is noisy and the predicted encoding matrix is imperfectly predicted from visual data.
- Result: The noisy recovery is almost as good as training a neural net to directly perform multi-object recognition, and in one experiment, even outperforms trained object detectors, suggesting that PBMF is an effective decomposition mechanism!
- Please see paper for detailed results.

Scene Classification

- Task: Predict the category of a scene image (e.g. kitchen, bathroom, park, etc.)
- Result: Despite being lower-parameter, lower-dimensional, and designed with interpretability in mind, ScenarioNet is competitive with standard CNNs and other methods for scene classification in terms of accuracy.
- Please see paper for detailed results.

Content-Based Querying and Comparison



- Task: Quickly compare two images based on high-level semantic content.
- Task: Retrieve images based on scene categories, scenarios, and individual objects.
- Please see paper for detailed results.

This work is partly supported by the AFOSR under the DDDAS program. This material is based upon work supported by the NSF Graduate Research Fellowship under Grant No. DGE-1433187.

XAI, 13 July 2013, Stockholm, Sweden