High-level Cues for Predicting Motivations

Arun Mallya Svetlana Lazebnik
University of Illinois at Urbana Champaign
amallya2,slazebni@illinois.edu

The success and effectiveness of deep networks on the popular tasks of image classification [10, 11], object detection [9, 3], *etc*. has set the foundation for the study of tasks based on high-level image concepts such Visual Question Answering (VQA) [1] and image captioning [7]. Despite the broad scope of required knowledge and successes in scene, action, and attribute prediction, most methods use generic features from the VGG-16 network [11] trained on the ImageNet classification task [10] for solving such tasks. A few recent works [12, 14, 15] have tried to use combine and use a variety of cues to improve performance on the task of VQA. These works claim that using specialized cues helps improves performance on tasks that require higher-level concepts. In this work, we further support this case by using action and scene cues to achieve state-of-the-art performance on predicting 'motivations' of humans in images. We propose a simple Canonical Correlation Analysis (CCA) [4] model based on scene and action cue features that achieves significantly better performance compared to prior work on the Motivations dataset [13].

The Motivations dataset consists of images each containing one selected person and 3 sentences describing the action being performed, the scene in the image, and the believed motivation of the person, respectively. The dataset aims to enable the creation of methods that can understand *why* a person is performing an action in the given setting. Along with the dataset, Vondrick *et al*. [13] proposed a structured-SVM–based method to retrieve and rank the 3 types of sentences over the test set. Their method uses 3 types of features - 1) VGG-16 *fc7* feature from the image, 2) VGG-16 *fc7* feature from the person bounding box, and 3) Language Model (LM) probabilities of all possible subsets of {Action, Scene, Motivation}. Their VGG-16 network was trained on the Places scene classification dataset [16] and the language model was trained on 6TB of web data.

In this work, we use the multimodal normalized CCA [4] which has been shown to be very effective on retrieval tasks [5, 4], instead of a structured-SVM. We represent images with action and scene cues, and sentences with skip-thought vectors [6] of length 4800. More specifically, we extract one action cue and one scene label cue. We use the action labels predicted by the Fusion network [8] based on VGG-16, trained on the HICO [2] action recognition dataset which has 600 labels. For scenes, we use labels predicted by a VGG-16 network trained on the Places dataset which has 365 labels. Thus, our image is represented by a feature vector of length 600+365 = 965 v/s 8192 in prior work. Further, each element of this feature vector has an associated human-interpretable label. Fig 1 shows samples from the Motivations dataset, along with the predicted cue labels. We train 3 separate CCA models, one for each of the Action, Scene, and Motivation sentences. We use action features for predicting Action and Motivation sentences, and scene features for predicting Scene sentences. We use an embedding dimension of 300 for all CCA models. We also try a baseline CCA model that uses *fc7* features from a VGG-16 network trained on ImageNet. The baseline and cue CCA use a regularization factor of 1e-2 and 1e-3 respectively. We split the train dataset into a train and val set of size 6133, and 1532 respectively. The test set contains 2526 images. All models were trained on the train set only and parameters were tuned on the val set. We try two text settings: 1) Replacing all sentences with cluster centers due to sentence similarity, as done in prior work [13], and 2) Using sentences as-is.

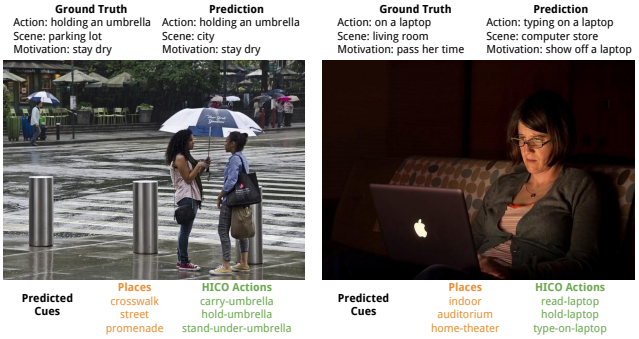

Figure 1: A sample image and associated sentences from the Motivations dataset. Below the image are top 3 labels predicted by our cue networks.

Our results are summarized in Table 1. Our baseline CCA trained on VGG-16 *fc7* features outperforms the method of Vondrick *et al*. [13] in spite of not using the language model features. Further, our method based on image cues instead of generic *fc7* features outperforms our baseline, in both text settings. We observe that scenes are adequately represented

Method	Motivation				Action				Scene			
	MR	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	R@1	R@5	R@10
Sentences replaced with cluster centers – #Clusters (Motivation, Action, Scene) : (256, 100,100)												
Image *fc7* [13]	39	-	-	-	17	-	-	-	4	-	-	-
Image *fc7* + Person *fc7* [13]	42	-	-	-	18	-	-	-	4	-	-	-
Image *fc7* + Person *fc7* + LM [13]	28	-	-	-	14	-	-	-	**3**	-	-	-
CCA – ImageNet VGG *fc7*	19	9.0	26.4	38.8	11	14.5	36.3	49.0	**3**	**35.2**	**61.5**	72.2
CCA – Action & Scene Cues	**14**	**12.0**	**34.1**	**44.9**	**8**	**18.6**	**43.3**	**56.3**	**3**	33.4	61.0	**73.8**
Sentences used as-is – #Sentences : 2526												
CCA – ImageNet VGG *fc7*	171.5	0.9	4.1	7.7	187	1.3	5.0	9.5	116	**1.1**	4.6	7.9
CCA – Action & Scene Cues	**130**	**1.4**	**6.5**	**12.0**	**117**	**1.8**	**7.8**	**13.2**	**113**	1.0	**4.9**	**8.8**

Table 1: Retrieval results obtained on the test set of the Motivations dataset. MR stands for Median Rank of correct sentence, R@x stands for Recall@x. Lower is better for MR, higher is better for R@x. Using high-level cues obtains better performance than the methods of Vondrick *et al*. [13] as well as *fc7* features from the VGG-16 trained for ImageNet classification. Further, the cue-based representation is compact ($965 < 4096 < 8192$) and interpretable.

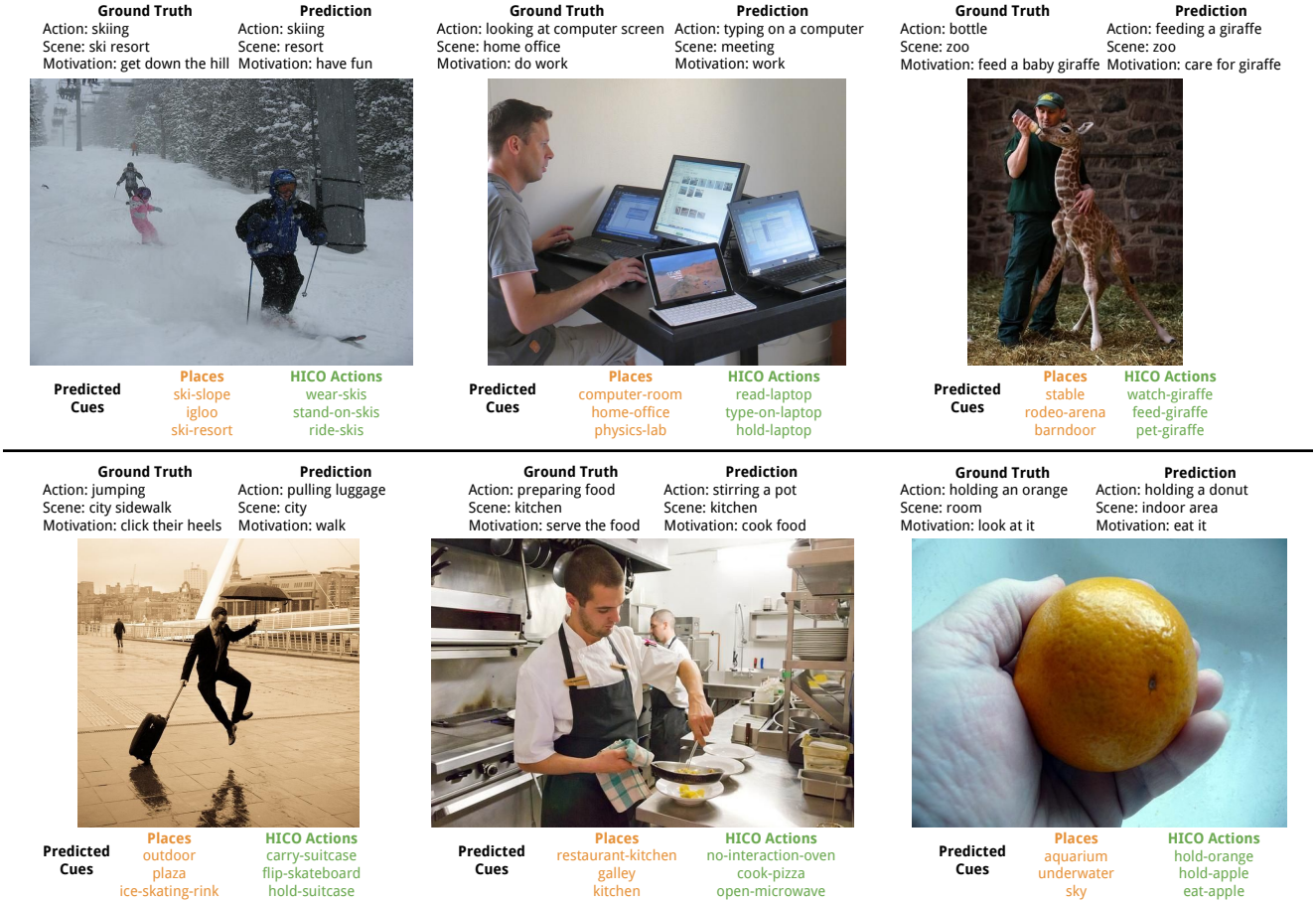

Figure 2: Sample predictions on the test set of the Motivations dataset. The ground truth annotation and predicted sentences are shown above each image. Below each image, the top 3 labels predicted by the scene and action networks are shown.

by the ImageNet VGG-16 *fc7* features, while action cues significantly improve motivation prediction. Using cue labels also helps us understand the workings of the model. Recall is rather low as sentences are similar to each other, motivating the use of sentence clustering in [13]. Interestingly, prior work of [13] achieved comparable performance on Scene sentences as they used features from a network trained on the Places dataset. Fig. 2 shows top retrieved sentences for some images of the test set. The top row shows some good predictions, where both cues are correctly predicted for the image. The bottom row shows cases where one or both of the cues are incorrectly predicted. Our action predictions are very accurate in most of the cases, while the places predictions can often be misleading, indicating scope for improvement.

From our results, we can conclude that high-level cues such as actions and scenes provide a compact image representation useful for solving complex tasks, while allowing humans to diagnose and understand model errors.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[2] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015.

[3] R. Girshick. Fast r-cnn. In *ICCV*, 2015.

[4] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 2014.

[5] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*. 2014.

[6] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *NIPS*, 2015.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*. 2014.

[8] A. Mallya and S. Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *ECCV*, 2016.

[9] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[12] T. Tommasi, A. Mallya, B. Plummer, S. Lazebnik, A. C. Berg, and T. L. Berg. Solving visual madlibs with multiple cues. In *BMVC*, 2016.

[13] C. Vondrick, D. Oktay, H. Pirsiavash, and A. Torralba. Predicting motivations of actions by leveraging text. In *CVPR*, 2016.

[14] P. Wang, Q. Wu, C. Shen, and A. v. d. Hengel. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. *arXiv preprint arXiv:1612.05386*, 2016.

[15] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, 2016.

[16] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.