Object Agnostic Activity Recognition

Harshal Priyadarshi †
University of Texas at Austin
Austin, TX 78712
harshal@cs.utexas.edu

Tushar Nagarajan †
University of Texas at Austin
Austin, TX 78712
tushar.nagarajan@utexas.edu

Abstract

Current activity recognition datasets are heavily object-centric. For an action 'phoning', a naive classifier trained to identify 'phone' objects is enough to reliably classify the action. In the process, cues about human pose, relative position of the agent and the object, and any other activity-specific features may be lost in favor of stronger object identity features. As a consequence, instances from other activities such as 'photographing' (where phone objects occur) are easily misclassified. In this paper we investigate this object-activity correlation, and present a Siamese network (called object-activity net), and an Adversarial Network to learn object agnostic features for activity recognition. We obtain fair results using the object-activity net, but could not confidently attribute the results to the object agnostic activity features. Still, the direction and approach is promising towards truly object-agnostic activity recognition to counter dataset bias.

1. Introduction

Activities and objects are tightly interrelated - having a good understanding of the objects present in an image, and the roles that they play, is essential for accurate activity recognition. For humans, common object-activity associations are driven by observations like "dogs usually fetch" and "cats usually claw", which in turn provide strong priors about the activity in unseen scenes. However, in activity recognition datasets, these associations manifest themselves as a large skew in the dataset with respect to certain objects and activities, which make learning more difficult. Traditional classification models operating in this setting learn to exploit these object-activity correlations to improve activity recognition, as the loss function does not incentivize against it. This can be detrimental when the same object is performing a different activity, resulting in the model wrongly predicting the activity that the object is most commonly involved in.

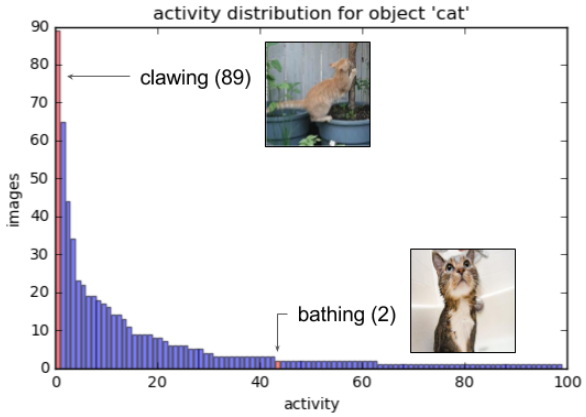

Figure 1. Distribution of activities across instances containing the object *cat*. Note the severe skew towards activities like clawing versus activities like *bathing*.

Figure 1 highlights the skew in the dataset that gives rise to these activity-object correlations. The figure shows the distribution of activities where the object *cat* occurs - *clawing* is the dominant activity. Activities with very low representation in the dataset, like *bathing*, are very difficult to classify if correlations between *cat* and *clawing* are learned as can be seen in Figure 2. Given a new image with a *cat* in it, the network has a propensity to predict the dominant activity based on object presence, rather than activity specific cues.

For the activity *bathing*, a similar skew exists, with the dominant object being a *bathtub*. Other activity specific cues such as the presence of soap bubbles, running water etc. are potentially ignored in favor of stronger object identification cues.

In this work, we investigate the idea of decorrelating the task of object and activity recognition, so that important activity-specific cues are not lost, and may be identified. We hypothesize that this decorrelation can be achieved by forcing the network to learn representations for images during activity recognition that are as *different* as possible from the representation learned for the same image during object recognition. We propose two models to achieve this,

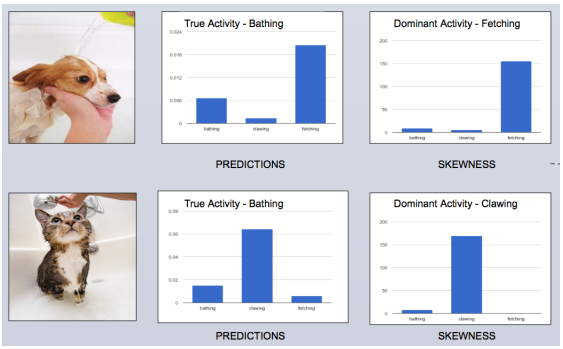

Figure 2. Demonstration of dominant activity's effect on prediction of other activities performed by the same object.

and evaluate our performance on a large activity recognition dataset, as well as a specially curated subset of the data which emphasizes the dataset skew.

Through our experiments on object-activity correlation, we show that there is indeed a correlation that develops between the features learned for the two tasks as training proceeds. Furthermore, we show that there exists a difficult subset of the dataset that contains instances of decorrelated object-activity pairs, upon which all models perform substantially worse. We emphasize that this subset is important, and that performance on this subset is a stronger indication of how well an activity recognition system performs than average performance over the whole dataset.

All our models and dataset curation code is made available online. [1]

2. Related Work

Activity recognition in static images has been studied extensively, especially in the context of human activity [5]. Work in this area has focused on features involving the human body, body parts, objects, human-object interaction and scene understanding to correctly predict activity.

More recently, the focus on human-object interaction has been taken a step further in work by Yatskar et al. [12] where an activity is decomposed into a situation with objects, agents, and the role that each of them play. The participating objects and agents are identified jointly in a conditional random field (CRF) based model.

2.1. Activity-Object Correlations

Gkioxari et al. [4] explore the observation that activity occurs with strong contextual cues (where the context here is elements of the scene, or secondary objects present in the image) that can be exploited to improve activity recognition. By adding the secondary region score to the activity prediction score, better performance is achieved.

[1] https://github.com/harpribot/AgnosticActivity

Strong object-activity correlations have been explored in [10], where activity was predicted from a bag of objects representation, generated for each frame of a video (in the egocentric domain). Rabinovich et al. [11] develop a model to use object context cues as a post-processing step over any classification system to improve results.

Models that do not subscribe to learning high-level object representations fall into categories that model human pose, layout, and agent-object interaction [1, 6, 2]. We aim to move the focus away from the object identities, but do not want to explicitly model pose or layout.

2.2. Learning Orthogonal Features

Khosla et al.[8] explore the idea of removing the dataset specific bias that exists for a specific task, and leave only the visual world vectors, which are common across datasets used in vision.

To combat unintended correlation of features with attributes, Jayaraman et al.[7] learn attribute models that are uncorrelated in the feature space. We intend to explore a similar approach but we wish to remove correlation of the object with activity, rather than any activity-activity correlation that might exist between similar activities.

Ganin et al. [3] propose a method to learn representations that are predictive of a particular task, but are invariant to domain shifts in training data. This forces the representations learned to capture non domain-specific features. In our work, we adapt the ideas from [3] and treat one domain as object recognition, and the other as activity recognition to learn *non object-specific, activity features*.

3. Dataset

3.1. ImSitu

We use the ImSitu dataset developed by Yatskar et al. [12] for our experiments. The dataset consists of 504 verbs (spread across 126,102 images), which we use as activity labels for our recognition task. The dataset also provides fine-grained situation annotations regarding which objects, agents and scenes are present, and the roles that each object/agent takes in the situation. We use the object labels in Section 4.3 for one of our models, but we do not exploit the remaining annotations.

This dataset is unique in that it is much larger than traditional activity recognition datasets in terms of the number of images and the number of classes. In addition, this dataset does not focus only on human activities.

3.2. ImSitu Selections

As observed in Section 1, activities in this dataset have very skewed object distributions, with some having the dominant object in upto 94% of all the images for that activity. Some objects also have similarly *dominant* activities.

We construct a subset of the ImSitu dataset that contains those images that do not possess strong object-activity correlations. These images are the *difficult* instances, that systems which overly rely on such correlations, will perform poorly on.

The dataset is constructed by looking at the tail of the object distribution for each activity. Images with objects that occur very infrequently in the context of an activity are good candidates for this subset. In addition, these objects must also be fairly highly occurring in another activity - The ratio of its occurrence in its dominant activity to the ratio of the next most occurring object in that activity must be above a certain threshold for it to be considered.

In this manner, a subset of 509 images are selected from ImSitu which represent images with little object-activity correlation.

4. Technical Approach

4.1. Baseline Model

The baseline model is modeled using the AlexNet architecture [9] till the $fc7$ layer, followed by a 504-way SoftMax layer (where each element of the output vector represents the probability of the 504 possible activities). The baseline model can be seen in Figure 3 (Top).

For this model, the weights are initialized with ImageNet pre-trained weights for the AlexNet model. The choice of this model for the baseline is justified, as this does not try to explicitly model object and activity prediction differently, and thus allows object prediction to affect activity prediction. The model is subsequently fine-tuned to convergence on the activity recognition dataset.

4.2. Object-Activity Net Model

Our proposed model that tries to decorrelate activity and object features. The model architecture can be seen in Figure 3 (Bottom Left). The model is a Siamese-like Network with no weight sharing across the parallel branches. The top network represents a network trained for object prediction (object branch), while the bottom network is trained for activity recognition (activity branch). The object branch is initialized in the same way as the baseline model in Section 4.1, and then no further training is performed on this branch (the weights are frozen). The activity branch is also initialized with the same pre-trained weights, and then fine-tuned for activity recognition. Both branches of the Siamese network receive the same input during training.

The objective function being optimized is slightly modified. We calculate the cosine distance between L2-normalized $fc7$ layers of both the object and activity networks in the model. We then subtract this from the classification loss on the activity branch. We minimize this aggregate loss function. This allows us to learn activity recognition (by minimizing classification loss) while learning distinct features (by maximizing the cosine distance between the object and activity features). We hypothesize that maximizing the cosine distance between features serves as a proxy for breaking the correlations that exist between the two tasks.

4.3. Adversarial Network Model

The adversarial network consists of a common trunk (formed from AlexNet layers till $fc7$) which splits into two branches - the object branch and the activity branch. The activity branch has two additional linear layers (with ReLU activations) followed by a 504 way SoftMax for activity recognition. The object branch contains the fc8 layer, followed by a sigmoid Activation. A sigmoid activation is used here has there can be multiple objects present in the image.

A gradient reversal layer, as described in [3] is inserted after the $fc7$ layer in the object branch. This layer multiplies all incoming gradients by $-\lambda$, thus preventing the $fc7$ representations from learning features that would strongly help in the object recognition task. We use $\lambda = 0.5$ for our experiments.

The trunk and the object branch are initialized in the same way as the baseline model. Unlike the model in Section 4.2, none of the weights are frozen. Standard Softmax loss is used for the activity branch, while binary cross entropy is used for the object branch (to support multiple object classification). The loss function is a weighted average of both losses, with weights chosen to match the scales of the losses. Our model architecture can be seen in Figure 3.

5. Experiments

5.1. Evaluation Metrics

Activity recognition is a multi-class, single label classification problem, thus graded evaluation metrics are not suitable here. We borrow evaluation metrics from Information Retrieval literature to better evaluate our results.

5.1.1 Precision@k

$$Precision@k = \sum_{s=1}^{N} (L(p_i^s = p_{true}^s \forall i \in [1...k])) \quad (1)$$

where p_i^s is predicted activity at rank i for sample image s, and p_{true}^s is the true activity for the image. N is total number of sample images. This metric checks if the true label lies in top k predictions, and can be thought of as an accuracy measure for a given tolerance.

For our experiments we choose to use $Precision@1$ and $Precision@5$. These are common metrics in use in the visual recognition community for classification purposes.

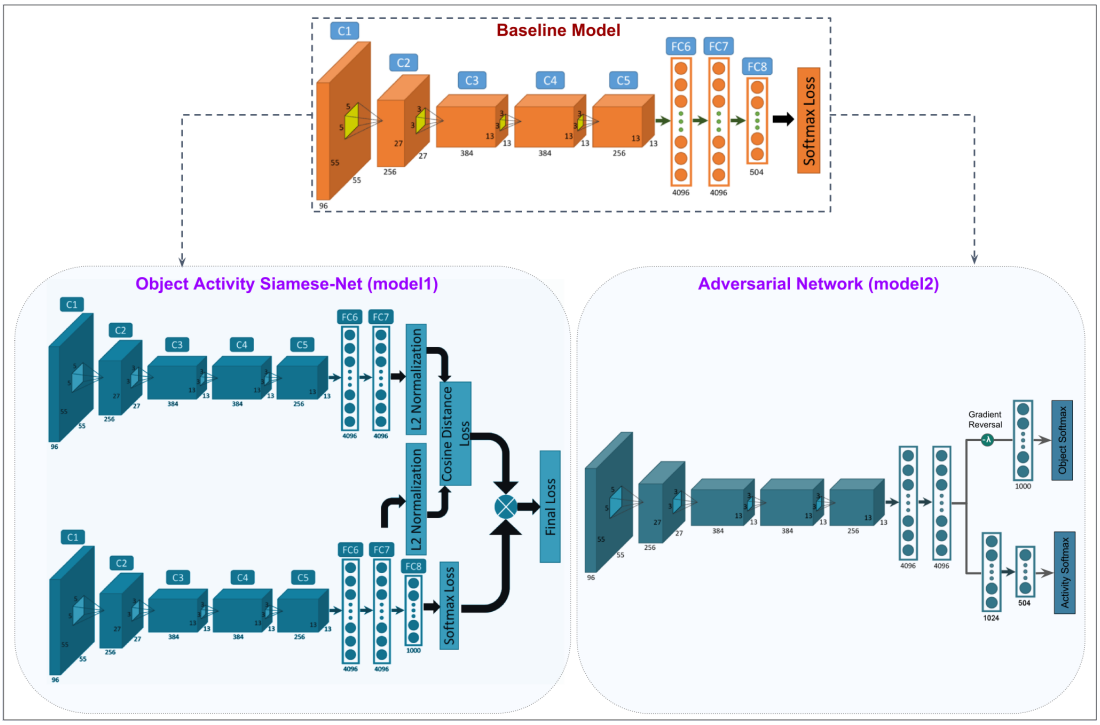

Figure 3. Top: Baseline Activity-Net model. Bottom Left: Object Activity Siamese Model, with frozen pre-trained weights for object side from ImageNet. Bottom Right: Adversarial Network where object prediction acts as an adversary to the activity prediction, by reversing the gradient that propagates back from the object branch, influencing the common representation layer.

5.1.2 Mean Reciprocal Rank (MRR)

Mean Reciprocal Rank is a stronger measure than Precision@k. It is a fine grained metric, that also gives higher value for rankings in which the true activity label is higher in the ranked list.

$$MRR = \sum_{s=1}^{N} \frac{1}{r^s_{true}} \tag{2}$$

where r^s_{true} is the predicted rank of the true activity by the model for the sample image s, when there are N sample images.

5.1.3 Precision-Recall Curve

We use PR curves to capture how well our models preform across the entire recall range. This is important as it captures the models confidence in predicting activities, which simple precision metrics do not. The curve is generated by sorting all predictions by their confidence score, and calculating precision and recall as this list is iterated over.

5.2. Dataset Skew Analysis

A look at the most skewed objects and activities present in the dataset reveal the extent to which the predictions could be affected. Figure 4 highlights the skew for the most extreme cases. The activity *piloting* has 20x more instances with *planes* in it than *choppers*. Similarly, the object *tent* dominantly occurs in instances of the activity *camping*. If an oracle system performs perfect object recognition, and always assigns the activity associated with the dominant object, an accuracy of 15.2% can be obtained. This is a very strong baseline compared to chance (2%), highlighting just how relevant dominant object associations are.

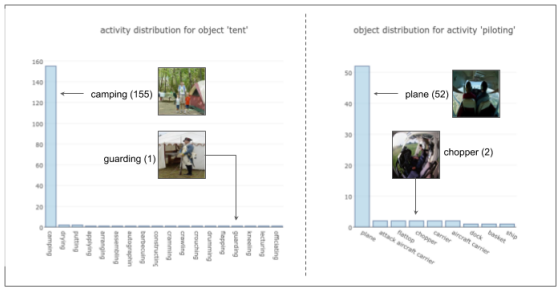

Figure 4. The object and activity distributions for the most skewed activity and object classes. There is a huge difference between the most dominant and second most dominant class in both cases.

This skew invariably affects the model performance. Our baseline model that is trained to predict activity assigns a higher score to the dominant activity *clawing* in 40% of all non-clawing images with cats in them.

Evidence of the effect of these correlations can also be

seen in the performance of all models on the curated ImSitu subset (Section 3.2). Precision values over the entire dataset across all models hovered around 20%, while the values on the subset (and relaxed subset versions with up to 2000 images) is almost half. This drop in performance can be attributed to the skew that we have discussed in this section.

5.3. Cosine Loss - Training Curve

To further support our hypothesis, we investigate the training curve of a neural network trained for activity recognition. Figure 5 highlights the evolution of the activity net's features with respect to the pre-trained object net features in the Object-Activity Siamese network of Section 4.2.

The green curve is the cosine-loss between the two networks activations. Notice that even though there is an initial drop in the cosine loss (the model is learning orthogonal features), as activity recognition becomes better, the cosine loss increases. That is, the features learned are becoming more similar to the features of the object net without any intervention. This loss continues to increase until the contrastive loss is added to the objective function being minimized, after which the loss goes down to zero, while training loss continues to reduce. This demonstrates that orthogonal features *can* be learned without compromising the ability of the network to predict activities.

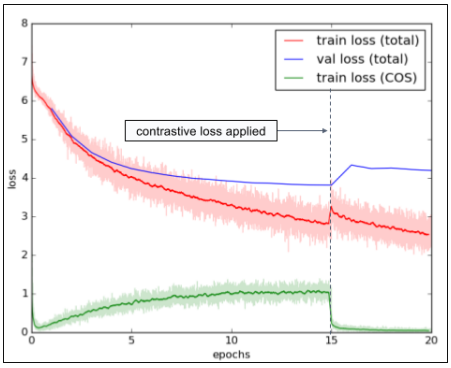

Figure 5. Profile of cosine loss during training of the Object-Activity Siamese Network. As training proceeds, the features learned by the activity net become more correlated with the features of the object net.

5.4. Model Ensemble

By decorrelating important object-activity associations, the performance of our model is bound to drop over the entire dataset. However, in the process, we hypothesize that the features learned contribute some new, orthogonal information. We ensemble the baseline models to our proposed models to verify if they have indeed learned orthogonal features, and if these orthogonal models lead to statistically significant improvements.

We performed a simple average ensemble of the probability distribution obtained by the baseline model and our model. The average ensemble gives a new probability distribution according to Equation 3.

$$p_{avg\ ensemble} = \frac{1}{2}\left[p_{baseline} + p_{proposed-model}\right] \quad (3)$$

The results obtained can be seen in the following tables

Table 1. Ensemble of Baseline and Siamese Net

	Imsitu-Validation	Imsitu-Test	Imsitu-Subset
P@1	0.231	0.235	0.121
P@5	0.462	0.466	0.292
MRR	0.344	0.348	0.222

Table 2. Ensemble of Baseline and Adversary Net

	Imsitu-Validation	Imsitu-Test	Imsitu-Subset
P@1	0.228	0.226	0.107
P@5	0.452	0.453	0.292
MRR	0.339	0.337	0.208

Table 3. Ensemble of Torch and Caffe Baseline Models

	Imsitu-Validation	Imsitu-Test	Imsitu-Subset
P@1	0.245	0.25	0.137
P@5	0.48	0.485	0.327
MRR	0.36	0.363	0.236

Table 4. Ensemble of Siamese Net and Adversary Net

	Imsitu-Validation	Imsitu-Test	Imsitu-Subset
P@1	0.239	0.241	0.126
P@5	0.473	0.474	0.327
MRR	0.354	0.355	0.23

6. Results

6.1. Model Comparison

For the ImSitu dataset, we can see in Figure 6, the performance of our models in comparison to the baseline model for both the validation and test dataset respectively. We see that both of our models consistently perform better than the baseline model, on all three evaluation metrics. We also see that the Object-Activity Net performs consistently better than the Adversarial Network. One possible reason for this behavior might be that, for the Adversary Network both the Object and Activity prediction tasks are trained together, and thus there is a high possibility that the object prediction half can get worse and thus might lose the object activity correlations that could be potentially exploited in this heavily skewed dataset.

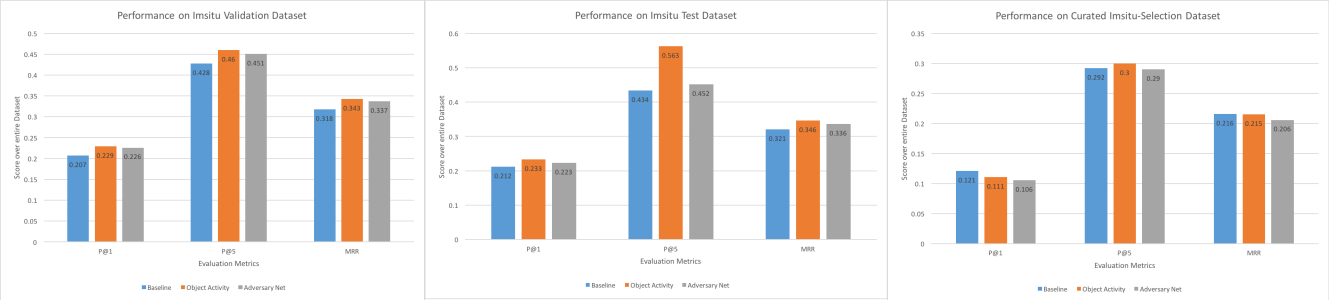

Figure 6. Evaluation of models over Left: ImSitu Validation data, Mid: ImSitu Test data, Right: ImSitu Curated Selection

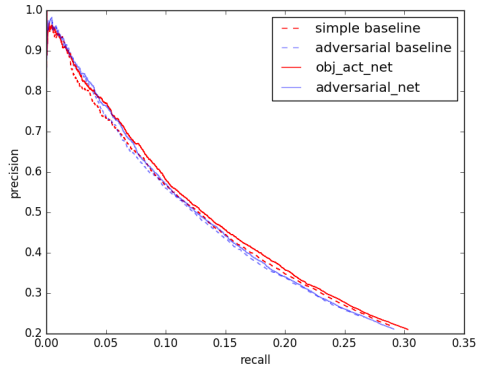

Figure 7. PR curve on ImSitu test set. In low recall ranges, all the models perform competitively, but at higher recall ranges, the object-activity net outperforms all other models by a decent margin.

However on the curated imsitu dataset, we don't see the hypothesized trend. As can be seen in Figure 6 (R), though the performance of both our models are slightly better than the baseline for $Precision@5$, but the performance is consistently slightly lower on $Precision@1$ and MRR metric. This result does not show that the models proposed are better in de-correlating object-activity association and at the same time improving performance on the images where this problem is present the most, in a statistically significant manner.

While precision metrics measure how accurate the model is, it is important to measure the models confidence across all the range of recall. From Figure 7, the object-activity net model outperforms the baseline models over high recall ranges, and matches the performance of the baseline models at low recall. A higher value across high recall regions demonstrates that even though feature decorrelation did not translate to accuracy, it reduced the volume of very confident, but incorrect predictions (where the confidence could be attributed to the dominant activity predictions arising from the dataset skew).

6.2. Ensemble Performance

We also hypothesized that our models learn features that are somewhat orthogonal to the original baseline model, as our model is better at identifying features that are learned when the object-activity associations are not considered, while the baseline model learns features when such interactions are present. Thus an ensemble of our model with the baseline model would lead to a performance improvement over and above each of the parent models.

When comparing to baseline results demonstrated in Figure 6, the ensemble performance of the Object Activity Net with the baseline model (Table 1), is slightly higher. However this slight edge is lost with the ensemble of baseline and Adversary Networks (shown in Table 2). In order to verify that the rise in performance of the former ensemble was because of the decorrelation of object-activity features, we ensemble two baseline models trained on different systems (Torch and Caffe), whose results are in Table 3. We see that this ensemble beats our best performing ensemble of baseline and object-activity net. With this control, it is uncertain whether the incremental advantage of the ensemble can be attributed to the feature decorrelation, or to advantage gained by ensembling models in general. We cannot confidently if truly orthogonal features have been learned by our models.

Further, the ensemble of our two models (Table 4) performed worse than the ensemble of our baseline models (Table 3). The performance of our models does not appear to be statistically significant over the ImSitu selection dataset. However, this might also be because of the heuristics applied in the automatic ImSitu data curation procedure. A more robust way of hand-picking the images would be the best, but would require substantial manual effort to curate such a collection, and thus was not performed given the time constraints of the project.

7. Conclusion

There is a large skew in activity recognition datasets caused by object-activity correlations. We analyze this skew, and show that these correlations severely impact

model performance by selecting a subset of data that represents the *hard* image instances where these correlations are absent. The models we present in this work aim at decorrelating the features learned by object and activity prediction tasks in an effort to circumvent this bias.

While our models showed improvements over the baselines over the entire recall range, the source of these improvements could not be confidently attributed to the orthogonal features learned. Further, the models performed poorly on the curated dataset which represents the difficult instances for object-activity correlation dependent models.

The idea of developing truly object-agnostic activity recognition systems is an important direction of research towards image understanding. It may have been an oversimplification to assume that maximizing distance between image representations is the same as learning orthogonal features, however, it is a promising start.

References

[1] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 9–16. IEEE, 2010.

[2] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *International journal of computer vision*, 95(1):1–12, 2011.

[3] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[4] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1080–1088, 2015.

[5] G. Guo and A. Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014.

[6] N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu. Recognizing actions from still images. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[7] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1629–1636, 2014.

[8] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] K. Matsuo, K. Yamada, S. Ueno, and S. Naito. An attention-based activity recognition for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 551–556, 2014.

[11] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[12] M. Yatskar, L. Zettlemoyer, and A. Farhadi. Situation recognition: Visual semantic role labeling for image understanding.