Object Agnostic Activity Recognition

Tushar Nagarajan, Harshal Priyadarshi

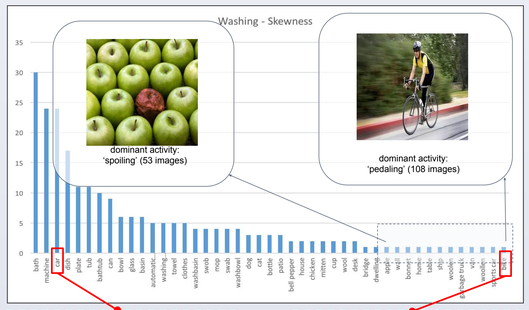
University of Texas at Austin

Problem Statement

Activity datasets are biased towards certain objects, for instance clinging is usually performed by koala bears, and eating by humans. Thus, simple ConvNets models might learn to exploit these object-activity associations to better predict activity, 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 object is most commonly involved in. This is indeed the case in the our findings. We show in the motivation section that the dataset is very skewed, and thus can confuse the convnet model in such scenarios.

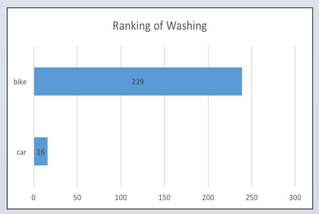
Motivation

Intra-Activity Skewness

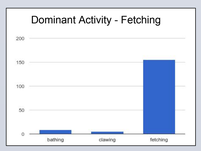


Too much skewness makes it harder to predict when the same activity is performed by a new, or relatively unseen (in the context of that activity) object

Reason -
Unwanted object correlation with activity prediction task

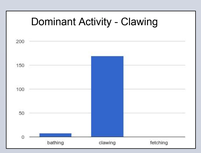


Inter-Activity Skewness



PREDICTIONS

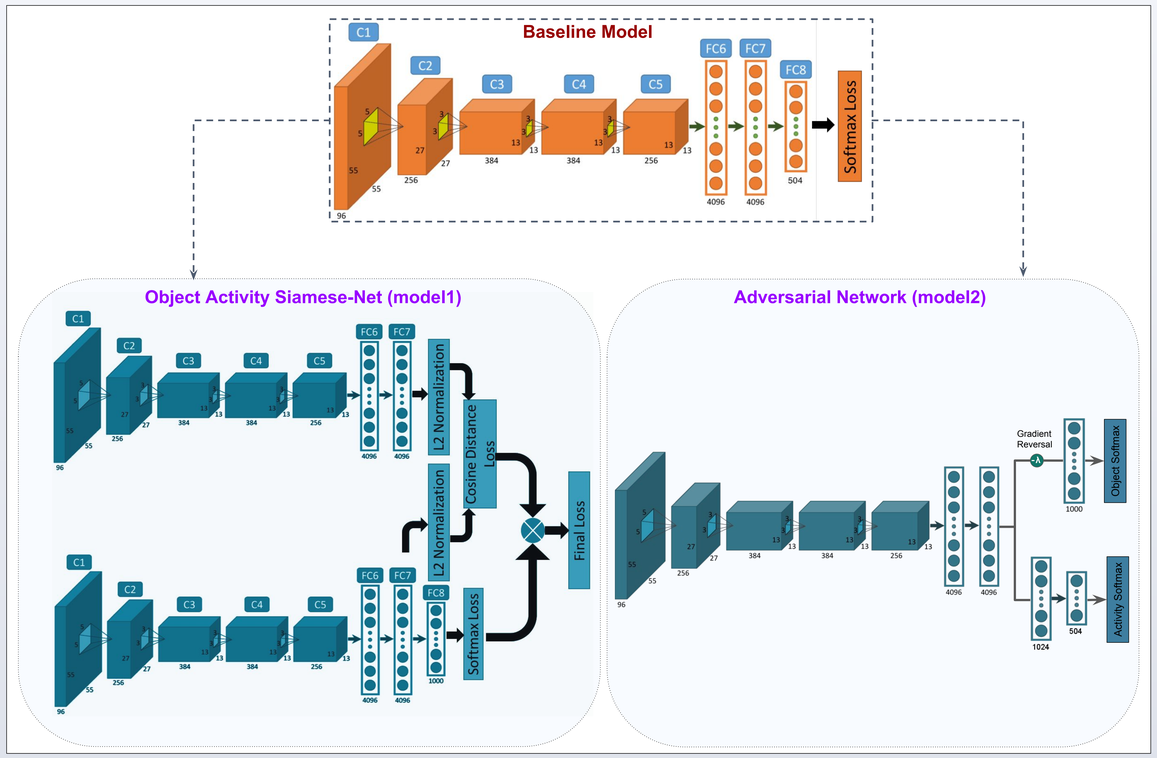
SKEWNESS



PREDICTIONS

SKEWNESS

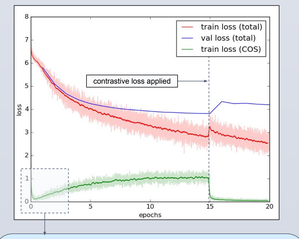
Model Architecture



Observations

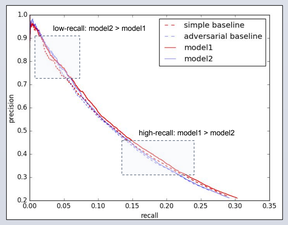
Of the models created and tested, model1 performed the best, despite its simple architecture, and is used to illustrate the observations below.

Training Curve - Contrastive Loss



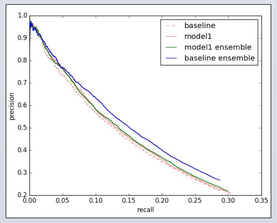
During the activity training stage, the contrastive loss **decreases** for several epochs. However, as activity recognition becomes better, the contrastive loss **increases** - this supports the hypothesis that activity features are correlated to object features

Model Comparison - imSitu



model1 outperforms the baselines over the entire recall range. While the final precision@1 of all models are comparable, discriminating object and activity features leads to fewer examples being misclassified as the 'dominant' activity based on the objects present.

Ensemble Methods

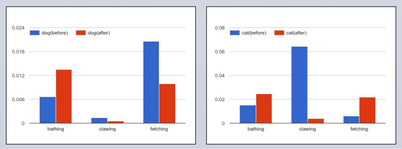


To test whether orthogonal information is learned, each model is ensembled with the baseline model.

A small gain is seen for **baseline + model1**, which hints towards the fact the model possess unique information.

To test if the increase is just a result of ensembling, two independently trained baseline models are ensembled as well. Curiously, this model performs the best - it is unclear what to attribute the performance jump to.

Effect of Feature Decorrelation



The sample of the baseline model predictions on the images of the cat and the dog, before and after feature decorrelation is enforced. While the overall confidence is still low, the **bias towards the dominant activity is heavily reduced.**

Methodology

We approach the task of de-coupling object and activity representations using two complementary models, and evaluate their performance on a curated subset of the ImSitu dataset [1] that exposes the undesirable object - activity correlations. Both models aim to produce representations that are orthogonal to those learned in object recognition tasks. The models are built upon the **Alexnet architecture** (baseline model) with the following modifications:

- **Object Activity Siamese-Net**: A contrastive loss is applied between the FC7 representations from the baseline branch (ImageNet pre-trained), and from the activity branch. Coupled with the activity classification loss, this will learn to **classify actions** as well as learn **dissimilar representations**. The object branch weights are frozen. Model is trained on Caffe [3].

- **Adversarial Network**: Following [2], we branch the FC7 layer into the activity and object branch, and treat the object branch as the adversarial domain. By multiplying all object-branch gradients by a negative scalar, **domain co-adaptation** is prevented.

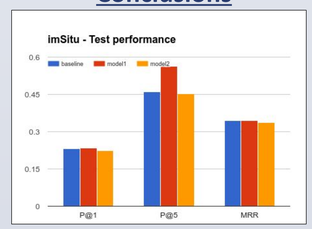
Dataset

A curated subset of imSitu is created from the tail of each activity-object distribution. A total of ~500 images are selected

	p@1	p@3	p@50	MRR
baseline	0.118	**0.235**	0.719	**0.217**
model1	**0.122**	0.223	0.713	0.214
model2	0.101	0.213	**0.729**	0.2

The subset is difficult - the models achieve an average **p@1 ~11%**, nearly **half** of what is seen on the whole dataset.

Conclusions



- We highlight the **strong object-activity correlation** in activity recognition datasets.
- Evaluation on the imSitu test set show **large improvements due to model1.**
- A qualitative analysis of skewed instances reveal that the **feature decorrelation helps avoid skewed predictions.**

From our quantitative experiments on ensemble methods, it is unclear whether to attribute these improvements to orthogonal features, or better initializations.

References

[1] ImSitu Dataset: Yatskar, Mark, Luke Zettlemoyer, and Ali Farhadi. "Situation Recognition: Visual Semantic Role Labeling for Image Understanding."
[2] Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." Journal of Machine Learning Research 17.59 (2016): 1-35.
[3] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." Proceedings of the 22nd ACM international conference on Multimedia. ACM 2014