A Comparative Analysis and Study of Multiview CNN Models for Joint Object Categorization and Pose Estimation

Mohamed Elhoseiny[*], **Tarek El-Gaaly**[*] M.ELHOSEINY@CS.RUTGERS.EDU, TGAALY@GMAIL.COM
Amr Bakry[*], **Ahmed Elgammal** AMRBAKRY@CS.RUTGERS.EDU, ELGAMMAL@CS.RUTGERS.EDU
Rutgers University, Computer Science Department, Piscataway, NJ, USA, 08854, [*] Equal contribution

Abstract

In the Object Recognition task, there exists a dichotomy between the categorization of objects and estimating object pose, where the former necessitates a view-invariant representation, while the latter requires a representation capable of capturing pose information over different categories of objects. With the rise of deep architectures, the prime focus has been on object category recognition. Deep learning methods have achieved wide success in this task. In contrast, object pose estimation using these approaches has received relatively less attention. In this work, we study how Convolutional Neural Networks (CNN) architectures can be adapted to the task of simultaneous object recognition and pose estimation. We investigate and analyze the layers of various CNN models and extensively compare between them with the goal of discovering how the layers of distributed representations within CNNs represent object pose information and how this contradicts with object category representations. We extensively experiment on two recent large and challenging multi-view datasets and we achieve better than the state-of-the-art.

1 Introduction

Impressive progress has been made over the last decade towards solving the problems of object categorization, localization and detection. It is desirable for a vision system to address two tasks under general object recognition: object categorization and object pose estimation (estimating the relative pose of an object with respect to a camera). Pose estimation is crucial in many applications. These two broad tasks are contradicting in nature. An optimal object categorization system should be able to recognize the category of an object, independent of its viewpoint. This means that the system should be able to learn viewpoint-invariant representations of object categories. In contrast, a pose estimation system requires a representation that preserves the geometric and visual features of objects in order to distinguish their pose. This gives rise to a fundamental question: *should categorization and pose estimation be solved simultaneously, and if so, can one aid the other?* Contrasting paradigms approach this question differently. Traditional instance-based 3D pose estimation approaches solve the instance-recognition and pose estimation problems simultaneously, given model bases of instances in 2D or 3D (Grimson & Lozano-Perez, 1985; Lamdan & Wolfson, 1988; Lowe, 1987; Shimshoni & Ponce, 1997). Most recent object pose estimation approaches solve the problem within the detection process, where category-specific object detectors that encode part geometry are trained (Savarese & Fei-Fei, 2007; Savarese & Li, 2008; Mei et al., 2011; Payet & Todorovic, 2011; Schels et al., 2012; Pepik et al., 2012). Since part-geometry is a function of the pose, these approaches are able to provide coarse estimate of the object pose with the detection. However the underlying assumption here is that the categorization is done a-priori, and the representation is view-variant. Other recent approaches try to solve the pose estimation simultaneously with categorization through learning dual representations: view-invariant category representation and view-variant category-invariant representation (Zhang et al., 2013a; Bakry & Elgammal, 2014).

With the rise of deep architectures, the main focus has been on category recognition. A wide success has been achieved on this task. In contrast, pose estimation has not received much attention. The impressive results of Convolutional Neural Networks (CNNs) in tasks of categorizations (Krizhevsky et al., 2012) and detection (Sermanet et al., 2013; Girshick et al., 2013) motivated many researchers to explore their applicability in different tasks. Several approaches recently showed successful results where they used networks that are pre-trained for a specific task (*e.g.* categorization) and then transfer the learnt deep representations for other tasks (Frome et al., 2013;

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

Donahue et al., 2013; Zeiler & Fergus, 2013). This process is known as transfer learning. As pointed out by (Yosinski et al., 2014), this process is useful when the target task has significantly smaller training data than what is needed to train the model. Typically the first n-layers are copied from the pre-trained network to initialize the corresponding layers for the target task. Within the CNN literature, typically the layers up until FC7 (i.e., last layer before the output layer) are used for that purpose (Frome et al., 2013).

Pose estimation is an example of a task that inherently suffers from lack of data. In fact the largest available dataset for multiview recognition and pose estimation has 51 object categories with a total of about 300 instances (Lai et al., 2011a). It is hard to imagine the availability of a dataset of thousands of objects where different views are sampled around each object in order to be able to train a learning machine such as a CNN with millions of parameters. Therefore, transfer learning is critical for this task. However the challenge lies in the contradicting objective that has been described in the first paragraph. Current CNN models are optimized for categorization, and therefore they are expected to achieve view invariant representation. Therefore it is not expected that feature representation at deeper layers are useful for pose estimation. However, feature representation in shallower layers tend to be more general and less category-specific and thus may hold enough information to discriminate between different poses. This is a key hypothesis that is explored in this paper and this work is the first exploration of the capability of CNNs on the task of object pose estimation.

The contributions of this paper are: (1) we show how CNNs can be adapted to the task of simultaneous categorization and pose estimation of objects, (2) we investigate how each of these tasks affect the other, *i.e.*how category-specific information can help estimate the pose of an object and how a balance between these contrasting tasks can be achieved, (3) we analyze different CNN models and extensively compare between them to find an efficient balance between accurate categorization and robust pose estimation, (4) we validate our work by extensive experiments on two recent large and challenging multi-view datasets. We achieve better than state-of-the-art performance on both datasets.

2 Related Work

Due to the surge of work in deep architectures over the last few years, there has amassed a large number of research studies. Despite this, using CNNs for regression and capturing pose information is still a relatively unexplored area. This motivates the goals of this paper.We focus on the most relevant work, in particular, studies that focus on understanding the functions of CNN layers and CNNs that solve for pose information. We also briefly touch upon previous approaches in object categorization and pose estimation.

Although fundamentally different to object pose estimation, some research has explored using CNNs to recognize human pose (Toshev & Szegedy, 2013; LI et al., 2014; Pfister et al., 2014). Recently (Gkioxari et al., 2014) proposed joint optimization on human pose and activity. In human pose estimation, there is no problem getting millions of image of people at different postures. Human activities are correlated with human poses, while in the object-pose domain the category is independent of pose. This makes joint learning of category and pose more challenging than joint learning of human activity and pose. Some very recent work has explored joint detection and pose estimation using CNNs (Tulsiani & Malik, 2014; Long et al., 2014; Tulsiani et al., 2015; Carreira et al., 2015).

Recent in-depth studies explore the intricacies of CNNs; including the effects of transfer-learning and fine-tuning, properties and dimensions of CNN layers and the study of invariances captured in CNN layers (*e.g.*(Yosinski et al., 2014; Zeiler & Fergus, 2013; Chatfield et al., 2014)). A data-centric analysis of existing CNN models for object detection has appeared in (Pepik et al., 2015a).

A comprehensive review of recent work in object recognition and pose estimation is detailed in (Savarese & Fei-fei, 2010). We highlight the most relevant research. Successful work have been done in estimating the object pose of a single object (Cyr & Kimia, 2004; Mei et al., 2011; Schels et al., 2012). This model, referred to as single-instance 3D model, has the limitation of being category-specific and does not scalable to a large number of categories and deal with high intra-class variation. Recently, detection and pose have been solved simultaneously (*e.g.*(Tulsiani & Malik, 2014; Pepik et al., 2015b; Xiang et al., 2014; Savarese & Fei-Fei, 2007; Savarese & Li, 2008; Mei et al., 2011; Payet & Todorovic, 2011; Schels et al., 2012; Pepik et al., 2012; Lai et al., 2011b)). Most of these methods belongs to the category of limited-pose (discrete-pose) approaches since it uses classification for pose estimation. Very few works formulate the pose estimation problem as regression over a continuous space. In (Lai et al., 2011b), an object pose tree is built for doing multi-level inference. This involves a classification strategy for pose which results in coarse estimates and does not utilize information present in the continuous distribution of descriptor spaces. Work presented in (Zhang et al., 2013b) and (Torki & Elgammal, 2011) explicitly model the continuous pose variations of objects but the scalability of these models is limited. A more recent work (Bakry & Elgammal, 2014) proposes a feedforward approach to solve the two problems jointly by balancing between continuous and discrete modeling of pose in order to increase performance and scalability. In these models, the nonlinearity in the representations are not modeled, which is mandatory for many applications.

3 Motivation

The first question we pose in this paper is *how good are pre-trained representations of different CNN layers, without fine-tuning, for the task of pose estimation?* To answer this we analyzed a state-of-the-art CNN trained on ImageNet (Krizhevsky et al., 2012) by testing it on dense multiview images from the RGBD dataset (Lai et al., 2011a) to see how well it represented object view-manifolds and hence able to estimate object poses. This CNN is composed of 8 layers: Conv1, Pool1, Conv2, Pool2, Conv3, Conv4, Conv5, Pool5, FC6, FC7, FC8. Pool indicates Max-Pooling layers, Conv indicates convolutional layers and FC indicates fully connected layers.

In order to quantitatively evaluate the representations of pose within the CNN, we trained both a pose regressor (using Kernel Ridge-Regression) and an SVM classifier for categorization (linear one-vs-all) on the features extracted from each of the layers. Fig. 1-Left is the result of the regressor and classifier. It clearly shows the conflict in representation of the pre-trained network. For pose estimation, the performance increases until around Pool5 and then decreases. This confirms that shallow layers that have sufficient abstractive representation offer better feature encoding for pose estimation. It appears that Pool5 provides a representation that captures the best compromise in performance, between categorization and pose discrimination.

In Fig 1-Left we report cross-evaluation of categorization using pose features and vice versa. FC8 (output) which is task specific, does not perform good pose estimation, while FC6/FC7 perform much better. It is interesting to observe the opposite is not true; when optimizing on pose, much of the category-specific information is still represented by the features of the CNN, as seen by the increase in performance of category recognition using the pose-optimized features.

We further explored using other regressors on multiple views of a single object instance (GPR (Rasmussen & Williams, 2005), WKNN (H et al., 1996), SVR (H et al., 1996), KTA (H et al., 1996)). We use a coffee mug instance that has enough visual and shape features to discriminate its poses. Fig 1-Right shows the MAE of the pose regression. The results confirm that the pose representation improves as we approach Pool5. This indicates that Pool5 has the best representation of the object's view-manifold. We also found that the performance of features based on Pool5 are the closest to correlate with the performance when using HOG features on the objects' multiview images (Fig 1-Right). This further proves that Pool5 has the abstraction capability to represent pose efficiently. It is important to point out here that, in addition to our analysis, in-depth manifold analysis was conducted to analyze the object-view manifolds and their representations within CNN layers(Bakry et al., 2015). This in-depth study corroborates the conclusions we make here.

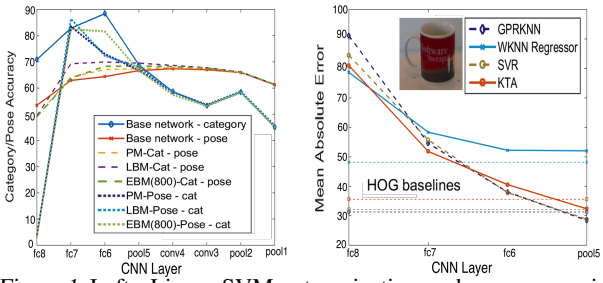

Figure 1. Left: Linear SVM categorization and pose regression performance based on feature encoding of different layers of a pre-trained CNN over all objects. The dotted lines are for cross-evaluation: for PM-Cat, LBM-Cat and EBM(800)-Cat represent the models' category representations evaluated on the task of pose estimation (to observe the effect of how category representations encode pose information). PM-Pose, LBM-Pose and EBM(800)-Pose are evaluated on the task of categorization to see how well pose representations in the CNN encodes categories. This is to show the complete pose-invariant representations of the layers when learning to categorize. Right: Pose regression on a single object - showing the Mean Absolute Error (in degrees) of various regression algorithms from FC8 to Pool5. The horizontal lines represent the regression performance on HOG feature descriptors computed on the images.

4 Analyzed Models

We used a state-of-the-art CNN built by (Krizhevsky et al., 2012) as our baseline network in our experiments (winner of the LSVRC-2012 Imagenet Challenge (Russakovsky et al., 2014)). We refer to this model as *Model0: base network*. This model is pre-trained on Imagenet. The last fully connected layer (FC8) is fed to a 1000-way softmax which produces a distribution over the category labels. Dropout was employed during training and Rectified Linear Units (ReLU) were used for faster training. Stochastic gradient descent is used for back propagation. Model0 is not fine-tuned, and thus an analysis of it shows how the layers of a CNN trained on categorization of ImageNet lacks the ability to represent pose efficiently. Throughout this study we vary the architecture of the base network and the loss functions. All other models described are pre-trained on ImageNet and fine-tuned on each of the two large dataset we experimented on. The models could be downloaded via `https://goo.gl/5ao9CN`

We propose and investigate four different CNN models: Parallel Model (PM), Cross-Product Model (CPM), Late Branching Model (LBM) and Early Branching Model (EBM). PM is a parallel version of the base network; two parallel and independent versions of the base network, one for categorization and one for pose. CPM has an output layer with units for each category and pose combination to jointly train (depicted in Figure 2-a). LBM and EBM models are also depicted in Figure 2. LBM branches into two last layers, one for categorization and one for pose. LBM is similar to the model proposed in (Gkioxari et al., 2014) for a different problem (action detection and human pose

estimation). Finally, EBM performs early branching into two subnetworks, each specialized in categorization and pose estimation, respectively. The output layer FC8 is not merged but instead the LBM and EBM networks are optimized over two loss functions, one concerned with building view-invariance representations for categorization and the other with category-invariant representations for pose estimation. Because of the branching, this causes two units to be active, one in each branch, at the same time. All losses are optimized by the multinomial logistic regression objective, similar to (Krizhevsky et al., 2012) (*Softmax* loss). We denote softmax loss of label $l \in \{l^c, l^p\}$ and image x as $loss_i(x, l)$, where i indicates if this loss is over category or pose modes, l^p and l^c are the labels for pose and category, respectively. We now describe each model in detail.

Parallel Model (PM): This model consists of two base networks running in parallel, each solving categorization and pose estimation independently. There is no sharing of information between the two networks. The goal of this model is to see how well the traditional CNN is capable of representing object-view manifolds and hence estimating object pose, independent of category-specific information. The category and pose losses minimized in this model are: $loss_c(x, l^c)$ and $loss_p(x, l^p)$, one for each of the tasks of categorization and pose estimation, respectively.

Cross-Product Model (CPM): CPM explores a way to combine categorization and pose estimation by building a last layer capable of capturing both (see Fig2-a). We build a layer with the number of units equivalent to the number of combinations of category and pose, *i.e.*the cross-product of category and pose labels. The number of categories varies according to the dataset as we will see. The pose angles (in this case yaw or azimuth angle of an object) is discretized into angle bins across the viewing circle. This is the case with all our pose estimating models. The loss function for CPM is the softmax loss over the cross-product of category and pose labels: $loss(x, l^p \times l^c)$.

Late Branching Model (LBM): We introduce a change in the architecture by splitting/branching the network into two last layers, each designed to be specific to the two tasks: categorization and pose estimation. Thus, this network has a shared representation for both category and pose information up until layer FC7 (see Fig2-b).

The goal is to learn category and pose information simultaneously from the representations encoded in the previous layers of the CNN. The question behind this model is whether one last layer would be enough to recover the pose information from the previous layers, in other words untangle the object view-manifold and give accurate pose estimates. In other words, one can think of this as testing the ability of the deep distributed representations of a CNN in holding both category-specific pose-invariant information as well as pose-variant information. LBM is trained using a linear combination of losses over category and pose: $\lambda_1 \cdot loss_c(x, l^c) + \lambda_2 \cdot loss_p(x, l^p)$ where λ_1, λ_2 are weights found empirically (see sec 7 in the supplementary).

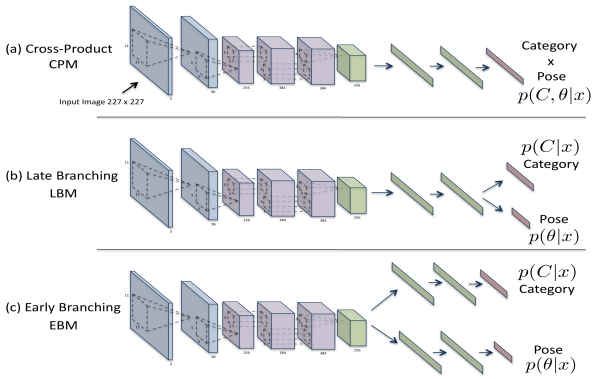

Figure 2. The studied models showing the joint loss layer in CPM, late branching in LBM and early branching in EBM. Blue layers correspond to layers with convolution, pooling and normalization. Violet colored layers correspond to layers with just convolution. Green layers correspond to fully-connected layers.

Early Branching Model (EBM): The question of moving the branching to an earlier layer in the network poses itself here: *Can the branching be moved earlier in the network to where the pose knowledge is still well preserved and in fact maximal across the layers?*

From our experiments (described later on) we observe that the objects' view-manifolds are maximally represented at Pool5. Thus, this network has a shared representation for both category and pose information up until layer Pool5. At Pool5 it branches out into two subnetworks, that are jointly optimized using a combined loss function (same as for LBM): $\lambda_1 \cdot loss_c(x, l^c) + \lambda_2 \cdot loss_p(x, l^p)$. Similar to LBM, it is important to note that this network optimizes over two losses. This model achieves the most efficient balance between categorization and pose estimation and achieve state-of-the-art results on two large challenging datasets, as we shall see in Section 7.

5 Datasets

5.1 RGBD Dataset

One of the largest and challenging multi-view datasets available is the RGB-D dataset (Lai et al., 2011a). It consists of 300 tabletop object instances over 51 different categories. Images are captured of objects rotating on a turntable, resulting in dense views of each object. The camera is positioned at three different heights with elevation angles: 30°, 45° and 60°.

In previous approaches the middle height (45°) is left out for testing (Lai et al., 2011b; Zhang et al., 2013a; Bakry & Elgammal, 2014; El-Gaaly et al., 2012). This means that object instances at test time have been seen before from other heights during training. For this dataset it was impor-

tant to experiment with an additional training-testing split of the data to give more meaningful results. We wanted to ensure that objects at test time had never seen before. We also wanted to make sure that the instances we are dealing with have non-degenerate view manifolds. We observed that many of the objects in the dataset are ill-posed, in the sense that the poses of the object are not distinct. This happens when the objects have no discriminating texture or shape to be able to identify its poses (*e.g.* a texture-less ball or orange). This causes object-view manifold degeneracy. For this reason, we select 34 out of the 51 categories as objects that possess variation across the viewpoints, and thus are not ill-posed with respect to pose estimation. We split the data into training, validation and testing. In this dataset, most categories have few instances; therefore we left out two random instances per category, one for validation and one for testing. In the case where a category has less than 5 instances, we form the validation set for that category by randomly sampling one object instance from the training set. We also left out all the middle height for testing. Thus, the testing set is composed of unseen instances and unseen heights and this allows us to more accurately evaluate the capability of CNNs in discriminating categories and poses of tabletop objects. We call this split, Split 1. In order to compare with state-of-the-art we also used the split used by previous approaches (we call this Split 2).

5.2 Pascal3D+ Dataset

We experiment on the recently released challenging dataset of multi-view images, called Pascal3D+ (Xiang et al., 2014). Pascal3D+ consists of images *in the wild*, in other words, images of object categories exhibiting high variability, captured under uncontrolled settings, in cluttered scenes and under many different poses. Pascal3D+ contains 12 categories of rigid objects selected from the PASCAL VOC 2012 dataset (Everingham et al., 2010). These objects are annotated with pose information (azimuth, elevation and distance to camera). Pascal3D+ also adds pose annotated images of these 12 categories from the ImageNet dataset (Russakovsky et al., 2014). The *bottle* category is omitted in state-of-the-art results. To be consistent, we do the same. This leaves 11 categories to experiment with. There are about 11,500 and 7,000 training images in ImageNet and Pascal3D+ subsets, respectively. We take a small portion of these images for validation and use the rest for training. For testing, there are about 11,200 and 6,900 testing images for ImageNet and Pascal3D+, respectively. On average there are about 3,000 object instances per category in Pascal3D+ captured *in the wild*, making it a challenging dataset for estimating object pose.

6 CNN Layer Analysis

Similar to the analysis performed in Section 3, we do the same on all our described models. This gives insight into the ability of these models to represent pose and the intrinsic differences between them. We perform kernel Ridge-regression and SVM classification on each layer of the CNN models. The results of this analysis on the two mutiview datasets are shown in Fig. 3 and 4.

From Fig. 3 and 4, we can see that the base network monotonically decreases in pose accuracy after layer Pool5. Pool5 seems to again hold substantial pose information, before it is lost in the following layers. This is the premise behind the design of our EBM model. EBM is able to efficiently untangle the object-view manifold and achieve good pose estimation on the branch specific to pose estimation. From Fig. 3 and 4, LBM is able to achieve a good boost in pose performance at its last layer. Fig. 3-right shows that layers Conv4 and Pool5 EBM have slightly worse accuracy than LBM and PM on the RGBD dataset. This indicates that the optimization is putting emphasis on the category information just before branching to achieve better pose estimation at deeper layers.

CPM does quite worse than the other models on both datasets, in both categorization and pose estimation. This can be seen in Fig. 3-Left and 3-Right and to some extent in Fig. 4. The reason for this lies in the fact that CPM shares information to jointly optimize over category and pose. The drop is more evident in the task of categorization, indicating again that category information aids in estimating the pose, but not the other way round. The drop

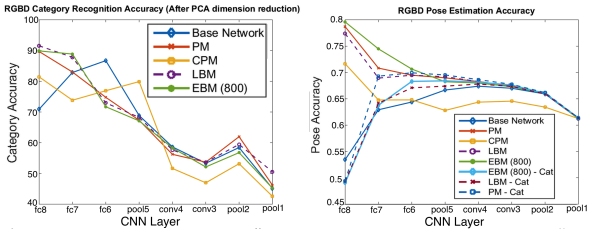

Figure 3. Analysis of layers trained on the RGBD dataset. Left: the performance of linear SVM category classification over the layers of different model. Right: the performance of pose regression over the layers of different models (including the category parts of some of the models - this shows the lack of pose information encoded within the object category representations)

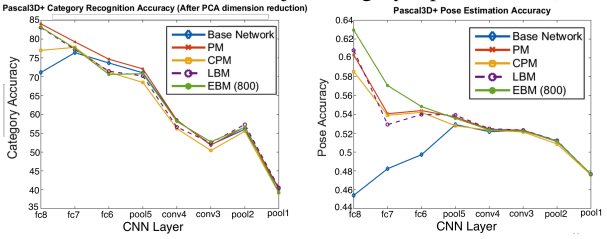

Figure 4. Analysis of layers trained on the Pascal3D+ dataset. Left: the performance of linear SVM category classification over the layers of different model. Right: the performance of pose regression over the layers of different models

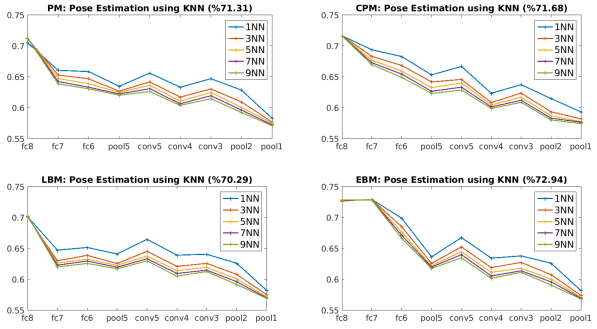

Figure 5. Comparison between the CNN models pose estimation at each layer of the CNN using k-NN with varying $k = \{1, 3, 5, 7, 9\}$ (on the Pascal3D dataset)

is more on the RGBD dataset because there are a lot more categories than Pascal3D+ and thus a lot more inter-class confusion. This is analogous to using category labels to separate between objects of different categories which may help bring similar posed objects of the same category together in the *latent* space encoded in the layers. On the other hand there is no clear untangling of the object-view manifold, where the pose information is stored, and thus this lack of pose information negatively impacts the categorization of objects.

k-NN Layer Analysis We conduct $k - NearestNeighbor$ pose estimation over the Pascal3D dataset on all the layers of the 4 models with varying neighborhood sizes (shown in Fig. 5). Comparing the two models (LBM and EBM), we gain slight improvement in categorization and a large improvement in pose estimation performance when using EBM. From Fig. 5 (and Fig. 1 in the supplementary materials), we conclude that as we go deeper into the network - up to layer Conv5 - we gain more category separation and object-view manifold preservation. This shows how the early branching better resolves the contradiction between the pose estimation and categorization tasks while sharing the low level filter representations that are helpful for both tasks. After Conv5, there are two common layers in EBM. In these two layers, linear separability between categories increases (seen in Fig. 1 in the supplementary materials), but the object-view manifolds collapse (as seen in Fig. 5). This hurts the pose estimation. At the same time, this supports the aforementioned claim that enforcing better categorization (fine-tuning) hurts pose estimation. In our best performing model (EBM), in Figure 5, remarkable improvement to the pose object-view manifold is attained. For pose, the drop in KNN-classifier as the K increases vanishes when going deeper in network; see FC6 and FC7 layers EBM in Fig. 5. KNN figure for categorization on Pascal3D dataset could be seen in the supplementary materials (Fig. 1). In a similar behavior EBM behaves better than CPM and PM. An interesting behavior that CPM works clearly better on Pascal3D dataset compared to RGBD; see Fig. 2 and Fig. 3 in the supplementary materials (Sec. 2) for KNN analysis on RGBD dataset. This is due that RGBD dataset has both dense poses and also larger number of categories (5 times Pascal3D). This increases the information/uncertainty to model that are beyond the capacity of CPM for RGBD dataset and generally as the number of categories and poses increase.

Local Pose Measurement Analysis: In the supplementary materials (Sec2), we further performed pose analysis for our models using four local pose analysis measurements proposed in (Bakry et al., 2015) to analyze each layer of the models. The purpose of this analysis is to show how the learning representations for each model is untangle to the circle manifold where the pose inhabits. The main conclusion that it shows the advantage of how EBM untangle the pose manifold locally compared to other CNN models.

7 Experiments

Here we describe the experimental setup and present the quantitative results of our experiments as well as comparisons with state-of-the-art.

7.1 Training and Testing

All classification losses are optimized by the multinomial logistic regression objective. Similar to (Krizhevsky et al., 2012), we optimized it by maximizing the average of the log-probability of the correct label under the prediction distribution across training cases. The pose softmax output (FC8) layer produces the pose probability distribution given the image. For each of the category and pose losses, the gradient with respect to the CNN parameters is computed which is then fed into CNN training for back propagation. More details about the training could be found in the supplementary materials (Sec 5).

The results presented in the paper were based on the prediction of $argmax_{pose} p(pose|x)$, where $pose$ is one of the 16 pose bins and x is the given image. In addition, we conduct an experiment where we predict the pose by computing the expected pose in the distribution of $p(pose|x)$; see Eq. 1.

$$E(pose|x) = \sum_i p(pose_i|x) \times \phi(pose_i) \qquad (1)$$

where $\phi(pose_i)$ is the center angle of the corresponding bin $pose_i$ (the pose of the i^{th} bin). The detailed definitions of the performance metrics used in our experiments are described in the supplementary (Sec 4).

7.2 Results

Table 1 shows the category recognition and pose estimation performance for the different models on the two training-testing splits of the RGBD dataset. Table 2 shows our best performing model EBM compared to state-of-the-art approaches. Using the pose prediction rule of eq. 1, the pose accuracy of EBM(800) increased from 78.83% to 79.30% for the $argmax$ prediction on split 2 of RGBD. Looking at the closest previous approach in Table 2, (Bakry & Elgammal, 2014) achieves 96.01% classification accuracy. This

is achieved using both visual and depth channels. We only used RGB (without depth) in our approach. (Bakry & Elgammal, 2014) achieves a lower 94.84% with RGB only, which shows the advantage of our CNN models for classification. We achieve 2.3% increase in category recognition and about 2% increase in pose estimation (79.30%) using EBM(800), when compared with state-of-the-art. These measurements are likely to increase further when using EBM(4096), as we see slight improvement of EBM(4096) over EBM(800) in Table 2. It is also possible that running k-NN on top of the layer features could improve performance further. We achieved 97.14% categorization using EBM. We also achieved 99.0% classification accuracy using Nearest Neighbor classification on the Pool5 layer of EBM, showing that we learned better convolutional filters.

Table 3 shows the performance of our models on Pascal3D+. We compare the accuracy achieved by our models with state-of-the-art results by (Zhang et al., 2015; Xiang et al., 2014; Pepik et al., 2015b; Tulsiani & Malik, 2014). It must be noted here that we are solving slightly different problems to some of these approaches. In (Xiang et al., 2014), the authors solve detection and pose estimation, assuming correct detection. On the other hand (Zhang et al., 2015) solve just pose estimation, assuming that the object categories are known. (Pepik et al., 2015b; Tulsiani & Malik, 2014) solve joint detection and pose estimation. In our case we are jointly solving both category recognition and pose estimation, which can be considered a harder problem than that of (Zhang et al., 2015) and (Xiang et al., 2014). Our pose estimation performance is better than all these previous approaches. For the sake of this comparison, we computed the pose performance using the metrics applied in (Zhang et al., 2015). These metrics are pose accuracy for images with pose errors $< 22.5^\circ$ and $< 45^\circ$.

Table 3 shows both our categorization and pose estimation results on Pascal3D compared against previous approaches. The table indicates 13.69% improvement of our method over (Zhang et al., 2015) (the best performing previous approach) in $pose < 22.5^\circ$ metric and 4% improvement in $pose < 45^\circ$, which are significant results. It is important to note that comparing to (Xiang et al., 2014; Tulsiani & Malik, 2014; Pepik et al., 2015b) (marked with (L)) is slightly unfair because these works solve for detection and pose simultaneously, while we do not solve detection.

We also show our performance when including ImageNet images in the training set and also the test set - see Table 3 (rows 7-8). The results show the benefit of ImageNet training images which boosts pose performance to 76.9% (from 57.89%) and 88.26% (from 63.0%) for $pose < 22.5^\circ$ and $pose < 45^\circ$, respectively. In Table 3, on the *in the wild* images of Pascal3D+, our EBM model achieves an impressive increase of $\sim$8% and $\sim$5% over the state-of-the-art models using the two pose accuracy metrics, respectively.

Table 1. A summary of all the results of the CNN models. Split 2 is the traditional RGBD dataset split. Split 1 is the one we describe that better evaluates our experiments. Split 2 is the state-of-the-art training-testing split. C indicates category performance and P indicates pose accuracy (where it is measured using 3 different metrics consistent with state-of-the-art. P ($< 22.5^\circ$) indicates that the pose accuracy is measured for objects where the pose error was less than 22.5°

Model	Split	C%	P% ($< 22.5^\circ$)	P% ($< 45^\circ$)	P% (AAAI)
PM	1	89.63	69.58	81.09	81.21
CPM	1	80.68	63.46	75.45	77.35
LBM	1	**91.48**	68.25	79.31	79.94
EBM (4096)	1	89.94	**71.49**	82.19	**82.00**
EBM (800)	1	89.84	71.29	**82.29**	81.91
EBM (400)	1	89.77	70.80	81.73	81.65
EBM (200)	1	90.11	67.70	79.43	79.77
EBM (100)	1	90.34	69.15	80.09	80.36
EBM (800)	2	**97.14**	**66.13**	**77.02**	**78.83**
EBM (4096)	2	97.07	65.82	76.51	78.66
SVM/Kernel Regression					
Model 0 (best category - FC6)	1	86.71	-	-	64.39
Model 0 (best pose - Conv4)	1	58.64	-	-	67.39
HOG	1	80.26	-	-	27.95

Table 2. RGBD Dataset: Comparison with state-of-the-art approaches on category recognition and pose estimation (Ours use only RGB channel).

Approach	Category %	Pose (AAAI) %
(Lai et al., 2011b)	94.30 (RGB + Depth)	53.50
(Bakry & Elgammal, 2014)	94.84 (RGB only)/96.01 (RGB+ Depth)	76.01
(Zhang et al., 2013a)	92.00 (RGB only)/93.10 (RGB + Depth)	61.57
(Bakry et al., 2016)	85.00	77.31
Ours (EBM(800))	**97.14**	**79.30**

Table 3. Pascal3D dataset (Xiang et al., 2014): Comparison with state-of-the-art approaches on category recognition and pose estimation. The AAAI pose metric is the performance metric used in(Zhang et al., 2015; Lai et al., 2011b; Zhang et al., 2013a).

Train: Pascal12, Test: Pascal12				
Approach	C	P% (< 22.5)	P% (< 45)	P% AAAI
(Xiang et al., 2014) (L)	-	15.60	18.70	-
(Pepik et al., 2012) (L)	-	17.30	21.50	-
(Pepik et al., 2015b) (L)	-	18.60	27.60	-
(Tulsiani & Malik, 2014) (L)	-	36.00	44.50	-
(Zhang et al., 2015)	-	44.20	59.00	-
EBM (4096)	83.0	**51.80**	**64.27**	**73.53**
EBM (800)	**83.10**	51.37	64.20	73.26
LBM	82.69	48.38	60.11	70.88
CPM	76.35	49.39	61.90	71.80
PM	**84.0**	47.34	61.30	71.60
Train: Pascal12 + ImageNet, Test: Pascal12				
EBM (800)	83.79	51.89	60.74	75.39
Train: Pascal12 + ImageNet, Test: Pascal12 + ImageNet				
EBM (800)	92.83	67.26	75.11	83.27

7.3 Computational Analysis and Convergence

We performed computational analysis on the convergence of the models and show that EBM converges substantially faster than all the other models. In Fig. 6 we show the convergence rates of the proposed models. EBM here is the larger EBM (4096) network. Despite having many more parameters than most of the other models (about $\sim$112 million parameters compared to 60 million in the base network), EBM converges substantially faster than all the other models. This shows the ability of this particular net-

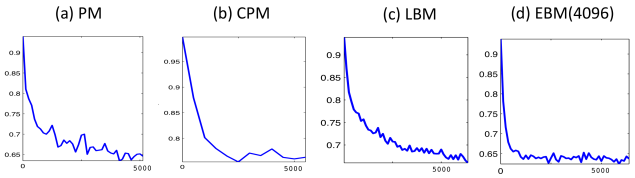

Figure 6. Comparison of convergence between the models. On the left is the category error and on the right is the pose error, on the validation set, respectively for each model (a) to (d). The error is computed per batch during each iteration. CPM shows the error for the joint category and pose. It can be seen that EBM (4096) converges much faster than the others which is another benefit of early branching. This is despite have a lot more parameters than the other models. This indicates that each of the subnetworks of EBM are able to specialize in both categorization and pose estimation faster. Each iteration is computed on one batch of 100 training samples.

work to specialize faster in the two tasks. The shared first five layers are able to build up the object-view manifolds, preserve them and enhance them in the pose subnetwork of the model, while the other subnetwork specializes in pose-invariant category recognition.

8 Discussion

Analysis of the layers of all the CNN models is shown in Fig. 3 to 5 here and (Figs. 1 to 7 in the supplementary). We further provide quantitative results over two challenging datasets and summarize them in Tables 1, 2 and 3.

We compare our models with multiple baselines in Table 1: Linear SVM and Kernel Ridge-Regression on HOG descriptors (Dalal & Triggs, 2005) as well as on features extracted from the best performing layers of the base network on each task. These baseline results were expected to be quite lower than our models' performance due to the lack of fine-tuning in the base model and due to the sharing of the network layers between the two tasks of category recognition and pose estimation. Without fine-tuning the base network does not represent the object-view manifold well enough to estimate the pose efficiently. After fine-tuning on each of the respective datasets, we were able to achieve good category performance using the PM model. The downside of this model is its inability to perform robust pose estimation on the more challenging natural sparser-views of Pascal3D+. This is evident in the results shown in Table 3, where PM achieves less pose accuracy.

In Tables 1 and 3, we see that CPM does worse than the other models in both datasets. This is more evident in the task of categorization, *e.g.*, a drop of $\sim$7% and 2%-3% in category and pose accuracy on Pascal3D+, respectively, and similarly $\sim$10% and $\sim$6% on the RGBD dataset. This motivates the need for branching in the networks and branching at the particular layer that better represents both category and pose. Interestingly, we found that CPM performs relatively better on Pascal3D+. We argue that the reason is that object poses in natural images are dominated by a smaller range of viewpoints and hence most of the pose-bins have vanishing probability (easier to learn). In addition, Pascal3D+ has a smaller number of categories.

LBM performs relatively well on RGBD, but not on Pascal3D+. This can be attributed to the fact that RGBD has many more categories and is composed of images of objects under controlled settings and not *in-the-wild* like in Pascal3D+. The images in the RGBD dataset are captured at dense views as the object rotates on a turn-table. This is why the pose information is more prevalent in the last layers. This is evident from the steep monotonically increasing curve of LBM in Fig. 3-Left. This is not the case in Pascal3D+ where the increase is more steady and in fact there is a decrease after layer FC6 (see Fig. 4-Right).

The reason why EBM performs better than PM even though its weights are randomly initialized is that PM's FC6 and FC7 layers in the pose-specific branch are initialized with category-specific weights from pre-training. This adversely affects pose estimation since it is a contradictory task that requires *view-variant* representations and not view-invariant representations like that required in categorization. Therefore initializing FC6/FC7 by another network trained for a different task is not likely to help. We show that learning the convolutional filters jointly with categories help make them discriminative for both tasks and thus achieves a good accuracy on both tasks (see Fig. 3 to 5 and Tables 1, 2 and 3).

Comparing between EBM and LBM, we see that early branching is able to achieve a good balance between categorization and pose estimation by sharing the representations up to where we found the layer representation still capture pose information. We see this in Tables 1, 3 and Fig. 1 to 5, where better pose accuracy and slightly better categorization accuracy is achieved by EBM. We also see that in Fig. 5 that the object view-manifold collapses in the last two layers (one layer before LBM) and thus achieves better pose discrimination than LBM. The slight effect of decreasing the size of the layers in the pose subnetwork of EBM can be observed from the results in Table 1 and 3.

9 Conclusion

This paper is an exploration of using CNNs for joint object categorization and pose estimation. We present our analysis and comparison of CNN models with the goal of efficiently performing both both tasks simultaneously. Despite the dichotomy in categorization and pose estimation, we show how CNNs can be adapted to simultaneously solve both tasks. We make key observations about the intrinsics of CNNs in their ability to represent pose. We quantitatively analyze the models on two large challenging datasets with extensive experiments and achieve better than state-of-the-art accuracy on both datasets.

Acknowledgment: This work is funded by NSF-USA award # 1409683.

References

Bakry, Amr and Elgammal, Ahmed. Untangling object-view manifold for multiview recognition and pose estimation. *ECCV*, 2014.

Bakry, Amr, Elhoseiny, Mohamed, El-Gaaly, Tarek, and Elgammal, Ahmed. Digging deep into the layers of cnns: In search of how cnns achieve view invariance. *arXiv preprint arXiv:1508.01983*, 2015.

Bakry, Amr, El-Gaaly, Tarek, Elhoseiny, Mohamed, and Elgammal, Ahmed. Joint object recognition and pose estimation using a nonlinear view-invariant latent generative model. *To appear in IEEE Winter Conference on Applications of Computer Vision (WACV) 2016*, 2016.

Carreira, Joao, Agrawal, Pulkit, Fragkiadaki, Katerina, and Malik, Jitendra. Human pose estimation with iterative error feedback. *arXiv preprint arXiv:1507.06550*, 2015.

Chatfield, Ken, Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. Return of the devil in the details: Delving deep into convolutional nets. *CoRR*, abs/1405.3531, 2014. URL http://arxiv.org/abs/1405.3531.

Cyr, CM and Kimia, BB. A similarity-based aspect-graph approach to 3D object recognition. *International Journal of Computer Vision*, 2004. URL http://link.springer.com/article/10.1023/B%3AVISI.0000013088.59081.4c.

Dalal, Navneet and Triggs, Bill. Histograms of oriented gradients for human detection. *CVPR*, 2005.

Donahue, Jeff, Jia, Yangqing, Vinyals, Oriol, Hoffman, Judy, Zhang, Ning, Tzeng, Eric, and Darrell, Trevor. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

El-Gaaly, Tarek, Torki, Marwan, Elgammal, Ahmed, and Singh, Maneesh. Rgbd object pose recognition using local-global multi-kernel regression. *ICPR*, 2012.

Everingham, Mark, Gool, Luc Van, Williams, C. K. I., Winn, J., and Zisserman, Andrew. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 2010.

Frome, Andrea, Corrado, Gregory S., Shlens, Jonathon, Bengio, Samy, Dean, Jeffrey, Ranzato, Marc'Aurelio, and Mikolov, Tomas. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

Girshick, Ross B., Donahue, Jeff, Darrell, Trevor, and Malik, Jitendra. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013. URL http://arxiv.org/abs/1311.2524.

Gkioxari, Georgia, Hariharan, Bharath, Girshick, Ross B., and Malik, Jitendra. R-cnns for pose estimation and action detection. *CoRR*, abs/1406.5212, 2014. URL http://arxiv.org/abs/1406.5212.

Grimson, WJEL and Lozano-Perez, T. Recognition and localization of overlapping parts from sparse data in two and three dimensions. In *Robotics and Automation. Proceedings. 1985 IEEE International Conference on*, volume 2, pp. 61–66. IEEE, 1985.

H, Drucker, CJC, Burges, L, Kaufman, A, Smola, and V, Vapnik. Support vector regression machines. *Advances in Neural Information Processing Systems*, 1996.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoff. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pp. 1106–1114, 2012.

Lai, K., Bo, L., Ren, X., and Fox, D. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pp. 1817–1824. IEEE, 2011a.

Lai, Kevin, Bo, Liefeng, Ren, Xiaofeng, and Fox, Dieter. A scalable tree-based approach for joint object and pose recognition. In *AAAI*, 2011b.

Lamdan, Y. and Wolfson, H. Geometric hashing: A general and efficient model-based recognition scheme. 1988.

LI, Sijin, Liu, Zhi-Qiang, and Chan, Antoni B. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014.

Long, Jonathan L, Zhang, Ning, and Darrell, Trevor. Do convnets learn correspondence? In *NIPS*, 2014.

Lowe, David G. Three-dimensional object recognition from single two-dimensional images. *Artificial intelligence*, 31(3):355–395, 1987.

Mei, Liang, Liu, Jingen, Hero, Alfred, and Savarese, Silvio. Robust object pose estimation via statistical manifold modeling. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 967–974. IEEE, 2011.

Payet, Nadia and Todorovic, Sinisa. From contours to 3d object detection and pose estimation. In *ICCV*, 2011.

Pepik, Bojan, Stark, Michael, Gehler, Peter, and Schiele, Bernt. Teaching 3d geometry to deformable part models. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3362–3369. IEEE, 2012.

Pepik, Bojan, Benenson, Rodrigo, Ritschel, Tobias, and Schiele, Bernt. What is holding back convnets for detection? *CoRR*, abs/1508.02844, 2015a. URL http://arxiv.org/abs/1508.02844.

Pepik, Bojan, Stark, Michael, Gehler, Peter V., Ritschel, Tobias, and Schiele, Bernt. 3d object class detection in the wild. *CoRR*, abs/1503.05038, 2015b. URL http://arxiv.org/abs/1503.05038.

Pfister, Tomas, Simonyan, Karen, Charles, James, and Zisserman, Andrew. Deep convolutional neural networks for efficient pose estimation in gesture videos. In *Asian Conference on Computer Vision (ACCV)*, 2014.

Rasmussen, Carl Edward and Williams, Christopher K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael S., Berg, Alexander C., and Fei-Fei, Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. URL http://arxiv.org/abs/1409.0575.

Savarese, S. and Fei-Fei, Li. 3d generic object categorization, localization and pose estimation. *ICCV*, 2007.

Savarese, S. and Li, F.F. View synthesis for recognizing unseen poses of object classes. pp. III: 602–615, 2008.

Savarese, Silvio and Fei-fei, Li. Multi-view Object Categorization and Pose Estimation. In *Computer Vision, SCI 285*. Springer-Verlag Berlin Heidelberg, 2010.

Schels, Johannes, Liebelt, Joerg, and Lienhart, Rainer. Learning an object class representation on a continuous viewsphere. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3170–3177. IEEE, 2012.

Sermanet, Pierre, Eigen, David, Zhang, Xiang, Mathieu, Michaël, Fergus, Rob, and LeCun, Yann. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013. URL http://arxiv.org/abs/1312.6229.

Shimshoni, Ilan and Ponce, Jean. Finite-resolution aspect graphs of polyhedral objects. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(4):315–327, 1997.

Torki, M. and Elgammal, A. Regression from local features for viewpoint and pose estimation. 2011.

Toshev, Alexander and Szegedy, Christian. Deeppose: Human pose estimation via deep neural networks. *CoRR*, abs/1312.4659, 2013. URL http://arxiv.org/abs/1312.4659.

Tulsiani, Shubham and Malik, Jitendra. Viewpoints and keypoints. *CoRR*, abs/1411.6067, 2014. URL http://arxiv.org/abs/1411.6067.

Tulsiani, Shubham, Carreira, Joao, and Malik, Jitendra. Pose induction for novel object categories. In *ICCV*, 2015.

Xiang, Yu, Mottaghi, Roozbeh, and Savarese, Silvio. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? *ArXiv e-prints*, November 2014.

Zeiler, Matthew D. and Fergus, Rob. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. URL http://arxiv.org/abs/1311.2901.

Zhang, Haopeng, El-Gaaly, Tarek, and Elgammal, Ahmed. Joint object and pose recognition using homeomorphic manifold analysis. *AAAI*, 2013a.

Zhang, Haopeng, El-Gaaly, Tarek, Elgammal, Ahmed, and Jiang, Zhiguo. Joint object and pose recognition using homeomorphic manifold analysis. In *AAAI*, 2013b.

Zhang, Haopeng, El-Gaaly, Tarek, Elgammal, Ahmed, and Jiang, Zhiguo. Factorization of view-object manifolds for joint object recognition and pose estimation. *arXiv preprint arXiv:1503.06813*, 2015.