Centroid-Based Scene Classification (CBSC): Using Deep Features and Clustering for RGB-D Indoor Scene Classification

Ali Ayub, Alan Wagner
The Pennsylvania State University
State College, PA 16802
{aja5755, alan.r.wagner}@psu.edu

Abstract
This paper contributes a novel method for RGB-D indoor scene classification. Recent approaches to this problem focus on developing increasingly complex pipelines that learn correlated features across the RGB and depth modalities. In contrast, this paper presents a simple method that first extracts features for the RGB and depth modalities using Places365-CNN and fine-tuned Places365-CNN on depth data, respectively and then clusters these features to generate a set of centroids representing each scene category from the training data. For classification a scene image is converted to CNN features and the distance of these features to the closest learned centroids is used to predict the image’s category. We evaluate our method on two standard RGB-D indoor scene classification benchmarks: SUNRGB-D and NYU Depth V2 and demonstrate that our proposed classification approach achieves superior performance over the state-of-the-art methods on both datasets.

Introduction
Classifying images taken from indoor scenes is an important area of research. The development of an accurate indoor scene classifier might improve indoor localization and decision-making for domestic robots, offer new applications for wearable computer users, and generally result in better vision-based situation awareness thus impacting a wide variety of applications.

The development of deep learning, the creation of numerous large-scale datasets, and the introduction of specialized computing hardware have all contributed to the rapid improvement in image classification performance. One reason for deep learning’s success has been the ability to learn multiple layers of generic image features that can then be used on other related computer vision problems. For instance, features from object trained image classifiers have been used to train indoor scene classifiers (Wang et al. 2016).

Yet, indoor scene classification is a challenging problem on its own. Although the presence of certain objects may provide evidence that an image is from a scene category, realistic indoor scenes are often cluttered with numerous objects that are unrelated to the scene’s category. Moreover, images taken from an indoor scene often lack category specific information or include information that could be from several different categories. Furthermore, indoor scene datasets tend to have an uneven distribution of images across scene categories. Finally, a single scene category (i.e. office in Figure 1) may include a variety of layouts that are composed of different objects and orientations. In general, each image from a scene category can only represent a specific part of the scene containing specific objects. Overall, indoor scenes tend to have high intra-class variation and low inter-class variation (Figure 1).

The development of low-cost RGB-D sensors, such as Microsoft’s Kinect, has generated additional interest in indoor scene classification. It has been argued that depth data might offer more robust geometric cues which would be particularly valuable for indoor scene classification (Zhu, Weibel, and Lu 2016). Recent research has successfully used this depth information to learn an additional set of discriminating features which, when used with the RGB features, improves indoor scene classification accuracy (Cheng et al. 2017). State-of-the-art approaches for indoor scene classification thus focus on developing methods that correlate features across the RGB and depth modalities and learn the relationships of features across modalities, while classification is performed using either a softmax classifier or SVM (Li et al. 2018; Yuan, Xiong, and Wang 2019). In contrast, this paper presents a remarkably simple and fast centroid-based scene classification (CBSC) approach to deal with
some of the problems of indoor scene classification mentioned above. This approach uses the Places365-CNN (Zhou et al. 2017) to extract a feature map for the RGB modality and fine-tuned Places365-CNN on depth data to get feature map for the depth modality. Then agglomerative clustering is applied on the feature maps of the training data for each scene category to get a set of centroids for that category. Finally, the distance of the feature maps of an unlabeled RGB-D image to the \( n \) closest centroids is used to predict its category. Our approach is tested on the two standard benchmark datasets for RGB-D indoor scene classification: the SUN RGB-D and the NYU Depth V2 datasets and it outperforms the state-of-the-art methods on both.

The main novelty of our approach is not in a particular technical component, a combination of clustering and deep features has been used in the past for different computer vision problems (Arandjelovic et al. 2016; Ng, Yang, and Davis 2015). Rather the unique combination of these techniques (deep features, clustering, weighted voting scheme for classification) to create a non-linear centroid-based classifier for the application of RGB-D scene classification is novel. The resulting classifier produces state-of-the-art performance on RGB-D scene classification benchmarks and provides further advantages such as improved model interpretability and faster learning time that are not shared by current deep learning techniques.

The main contributions of this paper are: 1) A new RGB-D indoor scene classification method is proposed that uses a combination of CNN extracted features and agglomerative clustering to achieve state-of-the-art performance. To the best of our knowledge, a clustering-based approach has not been used for RGB-D indoor scene classification before. 2) An approach that produces an interpretable model quickly. Lack of model interpretability has been one criticism of traditional deep learning (Lipton 2017). 3) The time to learn the set of centroids for the proposed approach is much lower than training a deep network.

Related Work

Approaches for indoor scene classification have been influenced by different research directions. This section first reviews methods for scene classification, then RGB-D scene classification, and finally centroid-based classification.

Scene Classification. Techniques for scene classification have rapidly improved. Early work relied on handcrafted features and low-level spatial information (Szummer and Picard 1998). Recent approaches have tended to focus on methods using Convolutional Neural Networks (CNN) to extract features from indoor scenes (Zhou et al. 2014; Zhou et al. 2017) notes that the features extracted from an ImageNet (Russakovsky et al. 2015) trained CNN results in poor performance on indoor scene classification. They therefore created the Places dataset which includes over 10 million labeled images of different scenes. Another approach has been to pool local image features using Fisher Vectors (FV) or Vector of Locally Aggregated Descriptors (VLAD) (Wang et al. 2017). Yet the classification performance of pooled image features is impacted by noisy image patches (Doshi, Kira, and Wagner 2015), however, demonstrates an early system that uses features generated by a CNN pretrained on ImageNet encoded as Fisher Vectors to classify scenes from streaming, first-person video. Unfortunately, they do not evaluate their method on standard datasets or compare its performance to other methods.

RGB-D Scene Classification. The availability of low-cost RGB-D sensors has encouraged the development of methods seeking to improve indoor scene classification by using features from both the RGB and depth modalities. In an early approach, (Gupta et al. 2013) proposed to extract local features by quantizing segmentation outputs and detecting contours on depth images. More recent work focuses on using CNNs because of their performance on object classification tasks (Krizhievsky, Sutskever, and Hinton 2012). Current state-of-the-art approaches focus on developing better methods to represent and correlate the RGB and depth features. (Cheng et al. 2017) use modality-specific features learned separately from RGB and depth images and then fuse the results at the score level. Depth features have also been learned independently and different fusion strategies have been explored in an attempt to maximize performance (Song et al. 2019). Li et al. (Li et al. 2018) presents a classification pipeline that learns and uses a fusion network. Current state-of-the-art indoor classification accuracy on the SUN RGB-D dataset is achieved by (Yuan, Xiong, and Wang 2019) using a cross-modal graph convolutional network to capture and use the RGB and depth relationships. These approaches have improved RGB-D scene classification by mainly focusing on learning better scene representations for the RGB and depth modalities, we present an entirely different classification approach which clusters the deep Places365 CNN features of the RGB-D images.

Centroid-based Classification. Clustering is a fundamental approach to pattern recognition and machine learning. Clustering groups data into self-similar collections, the central member of the group being a centroid. Once a set of centroids is obtained, the classification of an unlabeled data point is accomplished by matching the data point to its closest centroid, often using a distance metric such as euclidean distance or cosine distance (Jain, Murty, and Flynn 1999). A variety of clustering algorithms can be used to generate a set of centroids. Methods for clustering can be categorized into hierarchical and partition-based approaches (Jain, Murty, and Flynn 1999). Hierarchical clustering techniques do not require a priori determination of the number of clusters, but does require a stopping criterion. Agglomerative clustering is a hierarchical approach that creates a cluster hierarchy from the bottom up, initially assigning each data instance to its own cluster (Gowda and Krishna 1978; Dgalayahu, Weimshall, and Werman 2001). Next, the closest two clusters in terms of a distance metric are merged to create a new cluster. This procedure continues until some stopping distance threshold is achieved. By defining a distance (or similarity) threshold one can control the similarity of data points within a single centroid. For indoor scene recognition, agglomerative clustering can be applied to each scene category separately. Different distance thresholds result in different number of centroids for different scene categories. The standard algorithm for agglomerative clustering
Figure 2: The framework for our proposed approach generates the centroids for each scene category using agglomerative clustering and uses them for classifying unlabeled RGB-D images. (FT means fine-tuned)

has a time complexity of $O(n^3)$ and requires $O(n^2)$ memory, which makes it extremely slow for even medium data sets. For this reason, we propose a slight variation of the standard algorithm which has time complexity of $O(n)$ and performs better than the standard algorithm on the RGB-D scene classification benchmark datasets.

Many different variations of combining clustering and deep neural networks have been developed in the past (Aljallout et al. 2018). (Yang, Parikh, and Batra 2016) develops an unsupervised method for joint CNN and agglomerative clustering representation learning in which CNN features are used as an input to the agglomerative clustering procedure. Their method demonstrates near state-of-the-art performance on several object detection datasets. With respect to supervised machine learning, clustering has mostly been used for text classification tasks in recent years (Liu et al. 2017). Although, k-nearest neighbors has been applied for object-centric image classification tasks (Meyer, Harwood, and Drummond 2018), to the best of our knowledge, agglomerative clustering with CNN features has not been applied to the problem of RGB-D indoor scene classification.

Methodology

Figure 2 graphically depicts our approach. The first step extracts features for the RGB modality using VGG-16 CNN (Simonyan and Zisserman 2014) pre-trained on the Places365 dataset (Places365-CNN) and for the depth modality using a fine-tuned Places365-CNN on depth data. Next agglomerative clustering is applied to the training dataset for each category of scene resulting in a set of centroids for each scene category. Finally, the centroids are used to classify unlabeled scene images.

Centroid-Based Scene Classification

Our approach (Figure 2) begins with the generation of separate features maps for the two modalities. VGG16 CNN pre-trained on the Places365 dataset is used to generate features map for the RGB modality and the same pre-trained VGG16-CNN is fine-tuned on depth data to get feature map for the depth (HHA encoded) modality. Using the notation of (Yuan, Xiong, and Wang 2019), we denote the input RGB data as $x_{rgb}$ and the input depth data as $x_d$ and the feature maps generated by the second fully-connected layer of VGG as $F_{rgb}$ and $F_d$.

The proposed approach varies the standard agglomerative clustering algorithm in several ways. First, the traditional agglomerative clustering method initializes each data point as a cluster and the closest two clusters are then merged together iteratively until a stopping condition is met. Our variant of agglomerative clustering only initializes the first data point (image) as a centroid and then iterates through the remaining images in a category, merging the centroids based upon the pre-defined distance threshold (described below). Second, traditional agglomerative clustering first calculates a distance matrix containing the distance of each data point as a cluster and the closest two clusters are then merged together iteratively until a stopping condition is met. Our variant of agglomerative clustering only initializes the first data point (image) as a centroid and then iterates through the remaining images in a category, merging the centroids based upon the pre-defined distance threshold (described below). Second, traditional agglomerative clustering first calculates a distance matrix containing the distance of each data point with every other data point in the dataset. Our variant does not require this initial distance matrix. Finally, with the traditional method, after the distance matrix is created, the actual data points are not used. To merge the two closest centroids the algorithm simply replaces the distances of the individual clusters from the other clusters with the distance of the merged cluster from the other clusters. Moreover, the distance of the merged cluster to other clusters is calculated by using a pre-defined linkage method, such as the single linkage, on the individual cluster distances. In contrast, our variant merges the two closest centroids using the weighted mean (described below). The distances of the merged clusters from the other clusters are different for our proposed variant as compared to the traditional agglomerative clustering when using a weighted mean linkage for both approaches.

In the learning phase, our variant of agglomerative clustering is applied to each feature map in a scene category. This step begins by creating two centroids, one for each modality, from the first image in each scene category. Next, for each image in all categories, feature maps $F_{rgb}$ and $F_d$ are generated and compared using the euclidean distance to all the centroid pairs in the category. A weighted average
of the euclidean distances for each modality is used to calculate the overall RGB-D distance, \( \text{dist}_{\text{RD}}(c_{rdj}^i, F_{rdj}) \), between the \( i \)-th centroid pair \( c_{rdj}^i \) (for both modalities) in the scene category \( W_j \) and the feature map pair \( F_{rdj} \):

\[
\text{dist}_{\text{RD}}(c_{rdj}^i, F_{rdj}) = \frac{1}{2} \left( w_{rgb} \times \text{dist}(c_{rgbj}^l, F_{rgb}) + w_D \times \text{dist}(c_{d}^l, F_{d}) \right)
\]

where,

\[
\text{dist}(F, F') = \sqrt{\sum_{i=1}^{z} (F_i - F'_i)^2}
\]

is the euclidean distance between two feature maps \( F \) and \( F' \) of dimension \( z \) and \( c_{rgbj}^l \) and \( c_{d}^l \) are the \( i \)-th RGB and depth centroids for the scene category \( W_j \). The weights \( w_{rgb} \) and \( w_D \) are hyper-parameters we term fusion weights with values ranging from 0 to 1, for the RGB and depth distances.

**Algorithm 1** Generate centroids for each scene category

**Inputs:** \( F_{\text{RD}} \): RGB and depth feature map pairs, of the training dataset with \( m \) categories and \( k \) samples

\( D \): Distance threshold

\( w_{rgb}, w_D \): RGB and depth fusion weights

**Output:** A collection containing a set of centroid pairs for each category of indoor scene, \( C_{\text{RD}} = S_{\text{RD}1}, ..., S_{\text{RDm}} \)

1. \( F_{\text{RD}j} \): the set of RGB and depth feature map pairs labeled as category \( W_j \), with \( k_j \) samples, where \( f_{rdj}^i \) represents the \( i \)-th sample in \( F_{\text{RD}j} \).

2. \( C_{\text{RD}j} \): set of RGB and depth centroid pairs for category \( W_j \), where \( c_{rdj}^i \) represents the \( i \)-th centroid pair in \( C_{\text{RD}j} \).

3. for \( j = 1; j \leq m \) do \( C_{\text{RD}j} \leftarrow \{ f_{rdj}^1 \} \)

4. for \( j = 1; j \leq m \) do

5. for \( i = 2; i \leq k_j \)

6. \( d_{\text{min}} \leftarrow \min_{1 \leq l \leq \text{size}(C_{\text{RD}j})} \text{dist}_{\text{RD}}(c_{rdj}^l, f_{rdj}^i) \)

7. \( x \leftarrow \arg \min_{1 \leq l \leq \text{size}(C_{\text{RD}j})} \text{dist}_{\text{RD}}(c_{rdj}^l, f_{rdj}^i) \)

8. Set \( x_{rdj}^\star \) to be the nearest centroid pair

9. Set \( w_j \) to be the number of images clustered

10. in the \( x \)-th centroid pair of category \( W_j \)

11. if \( d_{\text{min}} < D \) then

12. Use Eq. 2 to update centroid pair \( c_{rdj}^\star \)

13. else

14. \( C_{\text{RD}j}.\text{append}(f_{rdj}^i) \)

For each image in all categories, the distance \( \text{dist}_{\text{RD}}(c_{rdj}^i, F_{rdj}) \) (eq. (1)) is computed for all current centroid pairs in a category to the image feature map pair, \( F_{rdj} \). If the distance to the closest centroid pair is below a pre-defined threshold \( D \), the centroid for each modality is updated by calculating a weighted mean of the centroids and the feature maps of the new image:

\[
C_{\text{new}} = \frac{w_C \ast C_{\text{old}} + F}{w_C + 1}
\]

where, \( C_{\text{new}} \) is the updated centroid, \( C_{\text{old}} \) is the centroid before the update, \( w_C \) is the number of data points (images) already represented by the centroid and \( F \) is the feature map of the new image. The above equation is used to calculate the updated centroids for both the RGB and depth modalities.

If the overall distance between the new image and the nearest centroid is higher than the distance threshold \( D \), a new centroid pair (for both modalities) is created for that category and equated to the feature map pair for the image. This approach is repeated for each image in the training data for each category. The result of this process is a collection containing a set of centroid pairs for each category of indoor scene, \( C_{\text{RD}} = S_{\text{RD}1}, S_{\text{RD}2}, ..., S_{\text{RDm}} \) where \( m \) is the number of scene categories and \( S_{\text{RDi}} \) represents the set of centroids of a scene category \( W_i \), where \( 1 \leq i \leq m \). This procedure is formally described as Algorithm 1. By choosing the optimal distance threshold \( D \), we can find a set of centroid pairs for each category such that each centroid pair represents a different layout in the scene.

**Classification of Unlabeled Images**

The output from Algorithm 1, a collection \( C_{\text{RD}} \) containing a set of centroid pairs for each category of indoor scenes, is used to classify unlabeled images. To classify an unlabeled image, we first use the Places365-CNN to calculate the unlabeled image’s feature maps, \( F_{\text{rgb}} \) and \( F_{\text{d}} \). The result is the selection of \( n \) closest centroid pairs to the unlabeled image. The contribution of each of the \( n \) closest centroid pairs to the determination of a scene category \( W_i \) is a conditional summation:

\[
\text{Pred}(W_i) = \sum_{j=1}^{n} \frac{1}{\text{dist}_{\text{RDj}}[y_j = W_i]}
\]

where \( \text{Pred}(W_i) \) is the prediction weight of the scene category \( W_i \), \( y_j \) is category label of \( j \)-th closest centroid pair and \( \text{dist}_{\text{RDj}} \) is the distance (calculated using Eq. (1)) between the \( j \)-th closest centroid pair and the feature maps (both modalities) of the test image. The prediction weights for all the categories are initialized to zero. Then, for the \( n \) closest centroid pairs the prediction weights are updated for the categories that each of the \( n \) centroid pairs belong to. The prediction weight for each class is further multiplied with the inverse of the total number of images in the training set of the class to deal with class imbalance. The test image is classified based on the category with the highest prediction weight. From equation (3), the prediction weight of a category is directly proportional to the number of centroid pairs, among the \( n \) centroid pairs, that belong to the category, and inversely proportional to the distance of those centroid pairs from the test image’s feature maps. Hence, the closest centroid pair contributes the most to the prediction of the category of the test image. The hyper-parameter, \( n \geq 1 \), was chosen empirically. Intuitively, because a scene category typically has multiple layouts, a test image can have different patches of pixels that match to different scene layouts. Hence, more than one closest centroid pairs (repre-
senting different layouts of a scene category) are considered when predicting the unlabeled image’s category.

**Experiments**

The proposed approach was evaluated on two standard RGB-D scene classification datasets: SUN RGB-D and NYU Depth V2. The datasets are first briefly described and then the performance of our approach is compared to the state-of-the-art methods for indoor scene classification.

**Datasets**

The SUN RGB-D dataset is the largest publicly available dataset for RGB-D indoor scene classification. It includes 10,335 RGB and depth image pairs captured from a variety of different camera and depth sensors. To be consistent with our predecessors’ experimental setup, we use 19 categories, all of which have more than 80 images (Song, Lichtenberg, and Xiao 2015). We also keep the standard splits, with 4,845 images for training and 4,659 images for testing.

The NYU Depth V2 dataset includes 1,449 RGB and depth image pairs consisting of 27 different categories. Because many categories have few training examples, (Silberman et al. 2012) reorganized the 27 categories into 10 categories with 9 usual scene types and one ”others” category. To again be consistent with our predecessor’s experimental setup, we follow the same category settings and the data split settings as in (Silberman et al. 2012) with 795 training images and 654 testing images. We also present the results from tests using all 27 categories but with the same split size.

**Implementation Details**

To generate the feature maps for the images, VGG-16 was implemented with the Keras deep learning framework (Chollet and others 2015) and using the pre-trained Places365 weights. A TITAN RTX GPU was used for feature extraction and fine-tuning and a Ryzen Threadripper 1920x CPU was used to create the centroids and classify test images. The input images were resized to 256 × 256 and randomly cropped to 224 × 224 as the input to the network. For the depth modality of the SUN RGB-D dataset, we fine-tune the Places365 CNN for 2500 epochs and for the depth modality for NYU Depth V2 dataset, we fine-tune the VGG16 network already fine-tuned for SUN RGB-D for another 1000 epochs. For both of the datasets we use a fixed learning rate of 0.0001, cross-entropy loss with mini-batches of size 56 and optimized with stochastic gradient descent. The hyper-parameters (distance threshold $D$, fusion weight $w_D$ and number of closest centroids for classification $n$) were tuned using cross-validation. For the SUN RGB-D dataset, the values of the hyper-parameters $D$, $w_D$ and $n$ are set as 80, 0.70 and 13, respectively. For NYU Depth V2 dataset, $D$, $w_D$ and $n$ are set as 165, 0.73 and 5, respectively, for the best results. For both the datasets, $w_R$ is set to 1.0.

**Results on the SUN RGB-D Dataset**

Table 1 compares our approach to nine state-of-the-art methods on the SUN RGB-D test set in terms of classification accuracy(%):

| Methods | RGB | Depth | Fusion |
| -- | -- | -- | -- |
| (Song, Lichtenberg, and Xiao 2015) | – | – | 39.0 |
| (Liao et al. 2016) | 36.1 | – | 41.3 |
| (Zhu, Weibel, and Lu 2016) | 40.4 | 36.5 | 41.5 |
| (Wang et al. 2016) | 40.4 | 36.5 | 41.5 |
| (Song, Jiang, and Herranz 2017) | – | 40.1 | 52.3 |
| (Du et al. 2018) | 42.6 | 43.3 | 53.3 |
| (Li et al. 2018) | 46.3 | 39.2 | 34.6 |
| (Song et al. 2019) | 44.6 | 42.7 | 33.8 |
| (Yuani, Xiong, and Wang 2019) | 45.7 | – | 35.1 |
| CBSC ($D = 85$, $w_D = 0.70$, $n = 13$, for Fusion) | **48.82** | 36.22 | **57.84** |

Table 1: Comparison with state-of-the-art methods on the SUN RGB-D test set in terms of classification accuracy(%)
improve scene classification. (Zhu, Weibel, and Lu 2016) attempt to exploit within-class and between-class correlations. A component aware feature fusion framework is developed by (Wang et al. 2016) to represent the scene. (Song, Jiang, and Herranz 2017) propose to use high level task specific features combined with low-level modality properties. (Li et al. 2018) focuses on improved multimodal feature learning by using discriminative and correlated features with structured loss and fusion learning. (Du et al. 2018) uses a two-step training strategy to build a GAN-based RGB-to-depth modality transition model in order to learn better depth features. Several different depth representations and fusion strategies are investigated by (Song et al. 2019). Most recently, cross-modal graph networks are utilized to extract semantic features for both modalities and then concatenated with the global features for improved accuracy (Yuan, Xiong, and Wang 2019).

Our centroid-based scene classification (CBSC) approach outperforms the other state-of-the-art methods on RGB only (48.82%) and RGB-D (57.84%) evaluations (Table 1). Average accuracy over all scene categories is reported. Figure 4 depicts the confusion matrix. For the RGB and RGB-D features, our approach outperforms the next best result by 2.52% and 2.74%, respectively. Our accuracy on the depth modality is the lowest reported because we are simply using the fine-tuned Places365-CNN for the depth modality. We have intentionally made no attempt to optimize depth features, choosing instead to focus on the value added by clustering optimized RGB features. Use of a network optimized for extracting depth features would undoubtedly further improve classification accuracy, although it is not clear by how much.

The time required to fine-tune the Places365 CNN for depth modality is about 18 hours. After that, the time required to generate the feature maps for the training images is 68.09s and the time required to generate the centroids is only 1.03s (using a single thread on a Ryzen Threadripper CPU). This is significantly faster than training a separate deep network.

Table 2 presents the results from an ablation study for our method. Baseline performance was generated by using the fine-tuned Places365-CNN where the streams for each modality are concatenated at the last fully connected layer. Our approach improves performance on the RGB modality over the baseline by a significant margin (6.55%). For the depth modality, we use the same features as the baseline (fine-tuned Places365-CNN features). The only difference between our approach and the baseline is that we use our clustering based classification while the baseline classification is performed using a softmax classifier. Hence, the improvement in classification accuracy is minimal for the depth modality (only 0.64%) and the value of our approach lies mostly in the RGB modality for this dataset. These results shows that even using the same features our classification approach is better than a softmax classifier. Finally, our approach improves performance by 8.00% over baseline VGG performance on the RGB-D data. We also compared the performance of our method to the same approach but using the traditional agglomerative clustering algorithm (Gowda and Krishna 1978; Gdalyahu, Weinshall, and Werman 2001) to generate the centroids. The best results for the traditional agglomerative clustering algorithm (Agglomerative RGB-D(HHA) in Table 2) were achieved when using the weighted mean linkage. Even using traditional agglomerative clustering algorithm, the results outperform the state-of-the-art methods but the results are inferior to our proposed variant (CBSC RGB-D in Table 2) which shows the value added by using our proposed variant of agglomerative clustering.

Hyper-parameter Analysis: The hyper-parameters $D$,
Table 3: Accuracy comparison with state-of-the-art methods on the NYU Depth V2 test set.

| Methods | RGB  | Depth | Fusion |
|---------|------|-------|--------|
| Wang et al. 2016 | 53.3 | 51.5 | 63.9 |
| Song, Jiang, and Herranz 2017 | - | - | 66.7 |
| Li et al. 2018 | 61.1 | 54.8 | 65.4 |
| Du et al. 2018 | 53.7 | 59.0 | 67.5 |
| Yuan, Xiong, and Wang 2019 | 55.4 | - | 67.4 |
| Song et al. 2019 | 53.4 | 56.4 | 67.5 |
| CBSC $(n = 5, D = 165, w_D = 0.73$ for fusion) | **66.44** | 49.50 | **69.7** |

Results on NYU Depth V2 Dataset

The confusion matrix can be viewed in Figure 5. All of the methods we compare to were introduced in the previous subsection. Our method outperforms all other methods on the RGB modality ($66.44\%$) by a significant margin ($5.34\%$). For the depth modality, similar to the SUN RGB-D results, our method performs worse than the other methods. Using fused features, our method outperforms the state-of-the-art ($69.70\%$) by a margin of $2.2\%$ over the next best method. It should also be noted that none of the top methods produce the best results on both RGB-D datasets. For example, the methods offered by (Song et al. 2019) and (Du et al. 2018) claim the best accuracy on the NYU Depth V2 dataset ($67.5\%$) but the method by (Yuan, Xiong, and Wang 2019) has the best accuracy on the SUNRGB-D dataset ($55.1\%$). Our approach results in better performance than all other methods on both datasets. Consistent with the ablation study for the SUNRGB-D dataset, our ablation study for the NYU Depth V2 dataset (Table 4) shows that most of the performance gains result from better classification of the RGB features.

Once the centroids have been created, we can also visualize the different clusters within a scene category by examining the images that compose them (Figure 6). We note that the images with the same layout are in fact in the same cluster while images with different layouts are in different clusters. This demonstrates an additional advantage of our approach in model interpretability.
Conclusion
This paper has introduced a new approach for RGB-D indoor scene classification based on clustering of CNN generated features. Our method uses the Places365-CNN features and then applies a variant of agglomerative clustering on the training data of each scene category to generate unique centroids representing different scene layouts for each scene category. To classify unlabeled images, the $n$ closest centroids are used in a weighted voting procedure that selects the best category for the image. Our approach produces state-of-the-art performance on the SUN RGB-D and NYU Depth V2 datasets. Furthermore, our centroids-based approach requires less training time than other methods and represents the scene layouts as cluster centroids improving model interpretability.

Our work offers new avenues of research for indoor scene classification. As the data shows, the success of our approach is hindered by our use of the depth features. In the future, we intend to use better depth features rather than simply using the Places365-CNN RGB features. Ultimately we hope that our method will be used in applications such as improved robot localization. Towards this goal we intend to publicly release our code.

References
[Aljalbout et al. 2018] Aljalbout, E.; Golkov, V.; Siddiqui, Y.; and Cremers, D. 2018. Clustering with deep learning: Taxonomy and new methods. CoRR abs/1801.07648.

[Arandjelovic et al. 2016] Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. Netvlad: Cnn architecture for weakly supervised place recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Cheng et al. 2017] Cheng, Y.; Cai, R.; Li, Z.; Zhao, X.; and Huang, K. 2017. Locality-sensitive deconvolutional networks with gated fusion for rgb-d indoor semantic segmentation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Chollet and others 2015] Chollet, F., et al. 2015. Keras. https://github.com/fchollet/keras

[Doshi, Kira, and Wagner 2015] Doshi, J.; Kira, Z.; and Wagner, A. 2015. From deep learning to episodic memories: Creating categories of visual experiences. In Third Annual Conference on Advances in Cognitive Systems.

[Du et al. 2018] Du, D.; Xu, X.; Ren, T.; and Wu, G. 2018. Depth images could tell us more: Enhancing depth discrim-
Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision* 115(3):211–252.

[Silberman et al. 2012] Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*.

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

[Song et al. 2019] Song, X.; Jiang, S.; Herranz, L.; and Chen, C. 2019. Learning effective rgb-d representations for scene recognition. *IEEE Transactions on Image Processing* 28(2):980–993.

[Song, Jiang, and Herranz 2017] Song, X.; Jiang, S.; and Herranz, L. 2017. Combining models from multiple sources for rgb-d scene recognition. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 4523–4529.

[Song, Lichtenberg, and Xiao 2015] Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In The *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Szummer and Picard 1998] Szummer, M., and Picard, R. W. 1998. Indoor-outdoor image classification. In *Proceedings 1998 IEEE International Workshop on Content-Based Access of Image and Video Database*, 42–51.

[Wang et al. 2016] Wang, A.; Cai, J.; Lu, J.; and Cham, T.-J. 2016. Modality and component aware feature fusion for rgb-d scene classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Wang et al. 2017] Wang, Z.; Wang, L.; Wang, Y.; Zhang, B.; and Qiao, Y. 2017. Weakly supervised patchnets: Describing and aggregating local patches for scene recognition. *IEEE Transactions on Image Processing* 26(4):2028–2041.

[Yang, Parikh, and Batra 2016] Yang, J.; Parikh, D.; and Batra, D. 2016. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5147–5156.

[Yuan, Xiong, and Wang 2019] Yuan, Y.; Xiong, Z.; and Wang, Q. 2019. Acm: Adaptive cross-modal graph convolutional neural networks for rgb-d scene recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01):9176–9184.

[Zhou et al. 2014] Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. 487–495.

[Zhou et al. 2017] Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Zhu, Weibel, and Lu 2016] Zhu, H.; Weibel, J.-B.; and Lu, S. 2016. Discriminative multi-modal feature fusion for rgbd indoor scene recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.