Basu, A., Kaewrak, K., Petropoulakis, L., Di Caterina, G., & Soraghan, J. J.
(Accepted/In press).
Indoor home scene recognition through instance segmentation using a combination of neural networks.
Paper presented at 2022 IEEE World Conference on Applied Intelligence and Computing, Sonbhadra, India

Indoor Home Scene Recognition through Instance Segmentation Using a Combination of Neural Networks

Amlan Basu
Ph.D. Student
Department of EEE
University of Strathclyde, Glasgow
amlan.basu@strath.ac.uk

Keerati Kaewrak
Ph.D. Student
Department of EEE
University of Strathclyde, Glasgow
keerati.kaewrak@strath.ac.uk

Lykourgos Petropoulakis
Senior Knowledge Exchange Fellow
Department of EEE
University of Strathclyde, Glasgow
l.petropoulakis@strath.ac.uk

Gaetano Di Caterina
Lecturer
Department of EEE
University of Strathclyde, Glasgow
gaetano.di-caterina@strath.ac.uk

John J. Soraghan
Professor
Department of EEE
University of Strathclyde, Glasgow
j.soraghan@strath.ac.uk

Abstract - This work presents a technique for recognizing indoor home scenes by using object detection. The object detection task is achieved through pre-trained Mask-RCNN (Regional Convolutional Neural Network), whilst the scene recognition is performed through a Convolutional Neural Network (CNN). The output of the Mask-RCNN is fed in input to the CNN, as this provides the CNN with the information of objects detected in one scene. So, the CNN recognizes the scene by looking at the combination of objects detected. The CNN is trained using the various object detection outputs of Mask-RCNN. This helps the CNN learn about the various combinations of objects that a scene can have. The CNN is trained using 500 combinations of 5 different scenes (bathroom, bedroom, kitchen, living room, and dining room) of the indoor home generated by Mask-RCNN. The trained network was tested on 24,000 indoor home scene images. The final accuracy produced by the CNN is 97.14%.

Keywords - Indoor home scenes, Object recognition, Scene recognition, Instance Segmentation, Transfer learning, Mask-RCNN, CNN.

I. INTRODUCTION

The scene and object recognition have always been performed separately as they are considered separate tasks. This also applies to indoor home scenes and indoor object recognition. It can be argued that object recognition is not as complex as compared to scene recognition. In fact, in object detection, the neural network only has to identify one object at a time, whereas, in the case of a scene, a neural network has to extract all minute information present in that scene. Scene recognition tasks for indoor home scenes can become very complicated because of the similarity between different scenes. For example, if a chair is present in the dining room, the same kind of chair can also be present in the bedroom and living room. This makes the task of the neural network very complicated during the learning phase. Moreover, in case of neural networks, the whole scene is learnt by acquiring different patterns present in the scene. That pattern may even involve the patterns of windows, doors, curtains, etc. Such things are present in every scene of an indoor home. Therefore, this becomes major hinderance in learning indoor home scenes.

Some similarities between indoor home scenes are inevitable. Hence, this indicates that if the neural network is always expected to recognize the indoor home scene correctly, it must remember the objects, or combinations of objects, associated with that scene. So, to make this possible, object and scene recognition tasks have to be combined.

In this work, connecting the pre-trained Mask-RCNN [1] with a CNN for accurately performing indoor home scene recognition is proposed. A Mask-RCNN pretrained on the COCO dataset for object detection is used to perform instance segmentation, which in turn provides labels or colour values to every pixel of each object. This helps extract the information of all the objects present in the indoor home scene accurately. The database has 500 home indoor images (100 images from each of 5 scenes: bathroom, bedroom, dining room, living room, kitchen), which are taken from the Places365 dataset [2]. The images are evenly distributed over five different categories of indoor home scenes. The five different scenes are bathroom, bedroom, kitchen, living room, and dining room. These images are tested on a Mask-RCNN for object detection. The detection helps in producing 100 different combinations for each indoor home scene. This detection output is used as data to train a CNN specifically designed to learn the object combinations of indoor home scenes.

II. RELATED WORKS

There are hardly any works that are totally focused on indoor home scene recognition, and most of such works are for more general indoor scene recognition. This can also

include offices, cafes, airports, etc. There is also hardly any work available representing indoor scene recognition through object detection using neural networks. Most of the available state-of-the-art works that propose scene recognition methods through object detection are not based on neural networks.

Indoor scene recognition task using Efficient Neural Network (EfficientNet) is proposed by M. Afif et al. [3]. EfficientNet is the combination of convolutional (Conv) layers and mobile bottleneck convolution layer (MBConv). Conv layers are used for feature extraction. MBConv layers are responsible for making the computation less complex by encoding the subspace's feature maps of lower dimensions. The training process of the neural network is done using transfer learning techniques. This work is also tested on an indoor home scene dataset (different from the one used in the proposed work) with bathroom, bedroom, kitchen, and living room scenes. The accuracy produced is very high (97%).

CodeBookless Model (CLM) [4] has also been used for indoor scene recognition and reported 20% of increased accuracy (90% overall) than the traditional methods of codebook construction. The method was also tested on indoor home scenes present in the Scene 15 dataset [5].

Multi-resolution CNN [6] is another neural network architecture specifically developed to solve the problems associated with large-scale scene recognition. The method has achieved good results on different large-scale benchmarks like Places [2], Places2 [2], LSUN [7], MIT67 [8], and SUN397 [9]. However, false-positive outputs still exist in this method.

Unified Convolutional Neural Network (Unified-CNN) proposed by H. Sun et al. [10], has parallel networks: one is dedicated to object detection, and the other is responsible for scene recognition. The parallel CNNs have a common input and common fully connected (FC) layers. However, the accuracy reported for both the task is very low (51.7% and 52.7% on scene and object respectively).

The partial use of neural networks can be understood from the work presented by X. Song et al. [11] in which Fast-RCNN [12] is used to first detect objects in different scenes. Then using the results obtained from Fast-RCNN, two techniques COOR (co-occurring frequency of object-to-object relation) and SOOR (sequential representation of object-to-object relation) are developed to establish object-to-object relationship. The established object relationships using COOR are then used to train SVM (support vector machine) for scene classification, whereas, in the case of SOOR, first the established object relationships are encoded using RNN (Recurrent Neural Network) and then MLP (Multi-Layer Perceptron) is used for scene classification. This whole technique still produced very low accuracy (from 50% to 66.9% on different datasets).

Some scene recognition techniques that do not involve neural networks are based on Scale Invariant Features Transform (SIFT) [13], Speed Up Robust Features (SURF) [14], etc. In Li Lj et al. [15] and Sudderth et al. [16], a scene recognition task is performed based on object recognition, which does not involve neural networks. An adaptive object detection method for indoor scene recognition that does not include a neural network is proposed by P. Espinace et al. [17, 18]. In these works, the probability of recognizing a scene accurately is too low, as a very smaller number of object combinations is used. These works have also failed to address the issue of the presence of similar objects in different rooms.

Except for the work proposed by M. Afif et al. [3] and X. Song et al. [11], none of the works have reported any technique completely dedicated to indoor home scene recognition. Even M. Afif et al. [3] have shown their performance on the indoor home scene dataset, but they trained their neural network on datasets that have mixed home and other indoor scenes (MIT67 [8] and Scene 15 dataset) instead of training only on indoor home scenes. So, there is no neural network that is trained and tested only on indoor home scene data. This may be attributed to the high similarity between indoor home scenes which can lead to very low accurate accuracy. The work presented in this letter addresses this issue and it implements an indoor home scene recognition using object detection and deep neural network, trained and tested on indoor home scene data. The overall structure has produced a system capable of highly accurate output (97.14% accuracy).

III. PROPOSED METHOD

The method proposed in this work involves a pre-trained Mask-RCNN (transfer learning [19] concept is utilized) capable of doing instance segmentation, which then helps achieve better object detection. The Mask-RCNN is made to identify objects in 500 different images of bathroom, bedroom, kitchen, living room, and dining room, with 100 images for each indoor home scene.

The output is the combination of objects detected by a Mask-RCNN represented as [57 58 60 61] where, in this case, the numbers represent “couch”, “chair”, “bed” and “dining table”, respectively. The combinations obtained, 500 overall, are then converted into a dataset. Each combination in the dataset is labelled as either bathroom, bedroom, kitchen, living room, or dining room. The obtained labelled data is then used to train the CNN. For training, 80% of the data is used and the remaining 20% is used for validation. This process is illustrated in Figure 1. The object output produced by Mask-RCNN is in sorted form (ascending order). Therefore, sorting is not required.

Fig.1. Training process for indoor home scene recognition using object detection

The CNN is developed to use it for learning the combination of objects and for classifying the scenes accordingly Since the obtained training data is 1D, also the CNN is one dimensional. As shown in Table 1, the CNN architecture contains three 1D convolutional layers: Conv1, Conv2, and Conv3. All convolutional layers have the same filter size of 3. The number of filters for convolutional layer 1, convolutional layer 2, and convolutional layer 3 is 16, 32, and 64, respectively. All convolutional layers have ReLU (Rectified Linear Unit) activation function, and no padding has been used in any convolutional layer. The stride in every convolutional layer is 1. After convolutional layer 3, there are three Fully Connected layers. The first and second Fully Connected layers have 4096 neurons with ReLU activation function. The third Fully Connected layer has neurons equal to the output, which is 5, with a softmax function.

Table I. CNN architecture specifications

Layers	No. of Filters/Neurons	Filter Size	Activation Function
Convolutional layer 1	16 filters	3	ReLU
Convolutional layer 2	32 filters	3	ReLU
Convolutional layer 3	64 filters	3	ReLU
Fully Connected layer-1	4096 neurons	-	ReLU
Fully Connected layer-2	4096 neurons	-	ReLU
Fully Connected layer-3	n-outputs (5)	-	Softmax

Once the CNN is trained, and weights are saved, then the testing is carried out, using the pre-trained Mask-RCNN and trained CNN shown in Figure 2. The output of the Mask-RCNN is connected to the input of the trained CNN, which produces the final output. There are 24,000 images in the Places365 dataset used for testing purposes. They are evenly distributed among bathroom, bedroom, kitchen, living room, and dining room. The Mask-RCNN produces an object combination of an indoor home scene. The combination then becomes the input for the CNN. Using this combination, the CNN predicts the indoor home scene.

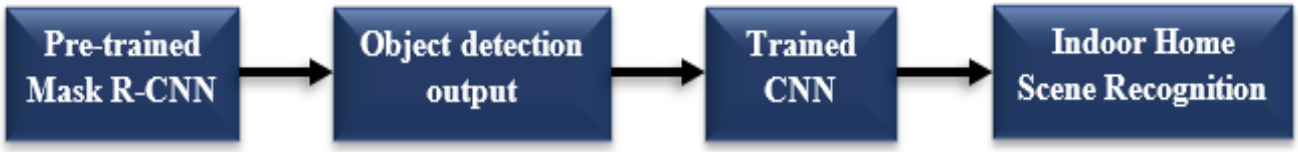


Fig. 2. Testing process using both Mask-RCNN and CNN

IV. RESULTS

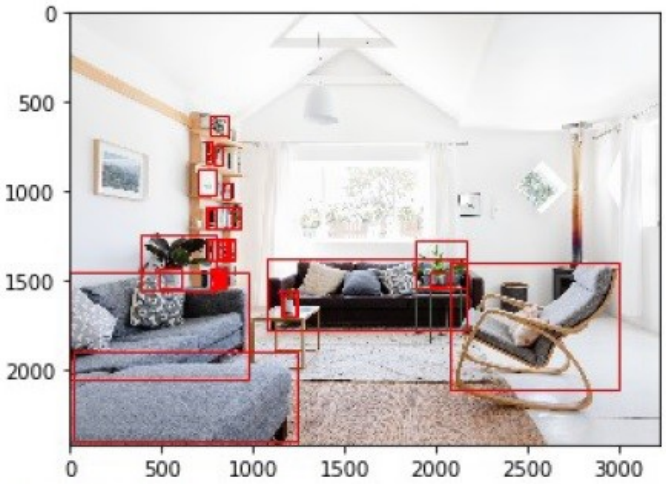
The proposed method is tested on 24,000 indoor home scene images distributed evenly among bathroom, bedroom, kitchen, living room, and dining room. All the images are taken from the Places365 dataset. The accuracy produced by the proposed method is 97.14%. This is the highest accuracy compared to other works on scene Some other neural networks which produced better accuracies on indoor scenes (includes indoor home scenes along with other indoor scenes) are ImageNet-GoogLeNet with 96.13% on Event8 dataset, Places365-VGG with 92.99% on SUN attribute dataset, Hybrid1365-VGG with 92.15% on Scene 15 dataset, and Places401-Deeper-BN-recognition. The accuracies produced by EfficientNet in M. Afif et al. are 95.6% on the MIT67 dataset (bathroom, bedroom, kitchen, and living room) and 97% on Scene 15 dataset (bedroom, kitchen, and living room). CML tested on the same Scene 15 dataset like M. Afifi et al. produced 90% accuracy.

Inception with 86.7% on MIT67 dataset. Unified-CNN produced an accuracy of 51.7% in scene recognition tasks, whereas it had 52.7% accuracy for object detection. All the mentioned neural networks are CNNs trained on a very large dataset to acquire the highest possible accuracy. The mentioned accuracies are summarized in Table 2.

Table II. Comparison of accuracies of different neural networks used for scene recognition

Neural Network	Accuracy (%)
Proposed Method (Mask-RCNN + CNN) (5 home scenes)	**97.14**
M. Afif et al. EfficientNet (4 home scenes) [115]	97
M. Afif et al. EfficientNet (3 home scenes) [115]	95.6
ImageNet-GoogLeNet [105]	96.13
Places365-VGG [105]	92.99
Hybrid1365-VGG [105]	92.15
CLM [112]	90
Places401-Deeper-BN-Inception [106]	86.7
Unified-CNN [104]	51.7

Different random images were chosen to check the correctness of the developed method. There are many similarities and often duplication of objects in the indoor home scenes. Hence, the proposed method is tested whether it is able to clearly distinguish between the indoor home scenes and recognize them properly in the face of such similarities. Figure 3 is the scene of a living room. The red rectangles show the objects detected in the scene. The objects detected by Mask-RCNN are [57 58 59 74 76], where 57, 58, 59 74 and 76 represent chair, couch, potted plant, book and vase respectively. The values seen just below the combination of objects detected are the probabilities of scenes in the form of [bathroom bedroom dining-room kitchen living-room] which are coded as [0 1 2 3 4]. For example, in Figure 3 the output 4 in this list shows that the highest probability value is that for the living-room. Therefore, the prediction is correct. In the living-room scene, it can be seen that objects like flower vase, books, and chair are present. These objects can be present in any room. Still, CNN predicts the scene properly based on combinations learnt.



[57 58 59 74 76]
[[1.7224460e-08 8.5565279e-12 7.7887211e-04 2.5235431e-07 9.9922085e-01]]
[4]
livingroom

Fig. 3. Scene recognition through object recognition in living room

Figure 4 is the scene of a bedroom with red quadrilaterals showing the objects detected on it. The object combination obtained from Mask-RCNN in Figure 4 is [42 59 60 76], where 42, 59, 60 and 76 represent cup, potted plant, bed and vase respectively. The scene detected by CNN from the obtained object combination is accurate as the highest probability is given to 1, which is the bedroom.

[42 59 60 76]
[[1.2231063e-10 1.0000000e+00 2.2445845e-09 1.0146228e-11 6.8124009e-12]]
[1]
bedroom

Fig. 4. Scene recognition through object recognition in bedroom

Figure 5 is the scene of a bathroom with red quadrilaterals showing the objects detected in it. The object combination obtained from Mask-RCNN in Figure 5 is [40 59 62 72 76], where 40, 59, 62, 72 and 76 represent bottle, potted plant, toilet, sink and vase respectively. The scene detected by CNN from the obtained object combination is accurate as the highest probability is given to 0, which is a bathroom. In the bathroom scene, a flower vase and some objects related to the kitchen, like water kettles, are detected. Still, CNN predicts the scene properly. This shows that CNN has learnt the accurate object combination associated with the bathroom.

[40 59 62 72 76]
[[9.9999952e-01 2.2787933e-15 1.2968398e-20 4.2241561e-07 2.3824250e-13]]
[0]
bathroom

Fig. 5. Scene recognition through object recognition in bathroom

Figure 6 is the scene of a dining room with red quadrilaterals showing the objects detected in it. The object combination obtained from Mask-RCNN in Figure 4 is [46 57 59 61 76], where 46, 57, 59, 61 and 76 represent bowl, chair, potted plant, dining table and vase respectively. The scene detected by CNN from the obtained object combination is accurate as the highest probability is given to 2, which is the dining room. In the dining room scene, a flower vase, hanging photo frame, and some objects related to the kitchen like bowl/plate are detected. The chairs and table, which can be present in any scene, are the main objects of a dining room that can easily confuse the CNN. Nonetheless, the CNN can predict the scene properly. This shows that CNN has learnt the accurate object combination associated with the dining room.

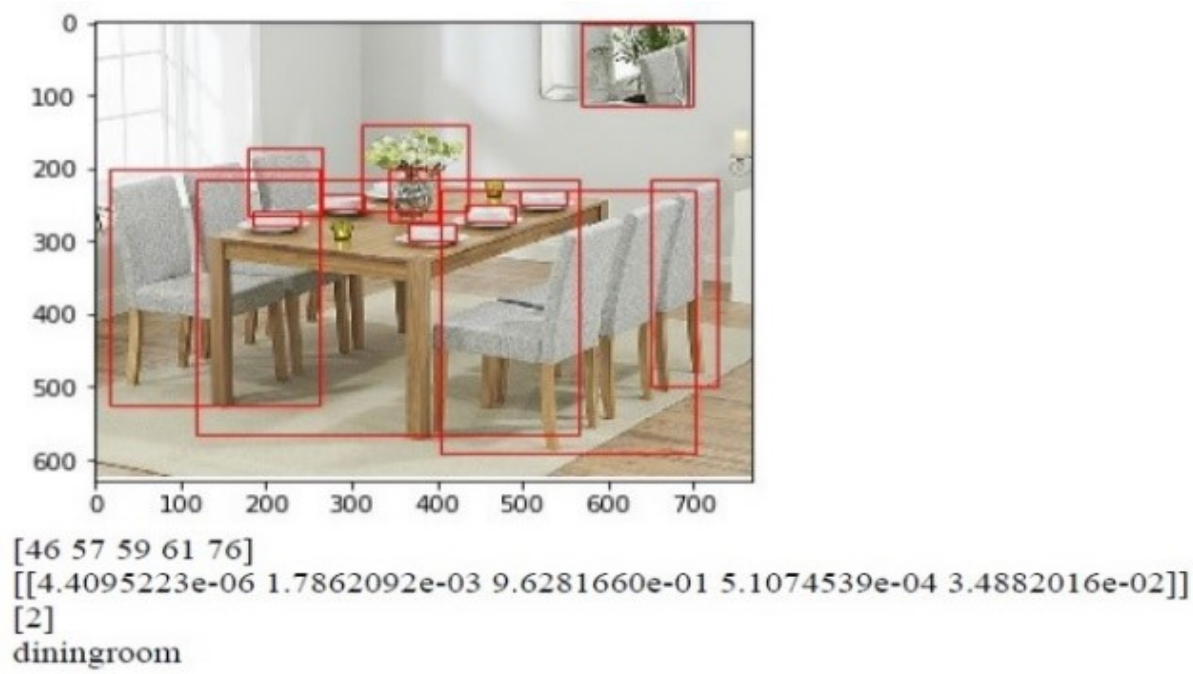


Fig. 6. Scene recognition through object recognition in the dining room

Figure 7 is the scene of a kitchen with red quadrilaterals showing the objects detected on it. The object combination obtained from Mask-RCNN in Figure 4 is [1 40 64 69 70 73 75], where 1, 40, 64, 69, 70, 73 and 75 represent background objects, bottle, laptop, microwave, oven, refrigerator and clock respectively. The scene detected by CNN from the obtained object combination is accurate as the highest probability is given to 3, which is the kitchen. The interesting part of the shown kitchen scene is that it has a laptop as well. This is quite unusual because a laptop is not the object that belongs to the kitchen scene. Wall clock also is detected in the kitchen, which can be found in any other room. Still, CNN from the obtained combination of objects accurately detects the kitchen. This is because CNN has properly learnt the actual combination of objects that make up a kitchen.

Fig. 7. Scene recognition through object recognition in kitchen

V. Conclusion

This article presents a state-of-the-art technique that involves neural networks for indoor home scene recognition through instance segmentation. The technique proposed in this letter solves the problem associated with indoor home scenes which is the high similarity often experienced between different scenes. The combination of Mask-RCNN (pre-trained) and CNN help in identifying the indoor home scenes by eliminating the problem of the presence of similar objects in different indoor home scenes. Pre-trained Mask-RCNN helped in generating 500 different combinations of objects. These 500 combinations helped in training the CNN so that it can learn the scene combinations properly. Then the combination of the pretrained Mask-RCNN and CNN was tested on 24000 previously unseen indoor home scenes with very high accuracy of 97.14%. This is because, the CNN identifies the indoor home scene on the basis of the produced combination of objects by the Mask-RCNN, using the already learnt combinations. By comparison, other works, e.g. [2] - [5], use larger-scale datasets for training purposes yet they still produce many false-positive results. The proposed method performs the task in a much simpler and faster way. Training the CNN to identify the indoor home scene using a combination of objects produced by the Mask-RCNN, takes only 15 minutes for the 500 object combinations used in this work.

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2017: IEEE, pp. 2980-2988.

[2] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence,* 2017.

[3] M. Afif, R. Ayachi, Y. Said, and M. Atri, "Deep Learning Based Application for Indoor Scene Recognition," *Neural Processing Letters,* pp. 1-11, 2020.

[4] P. Wu, Y. n. Li, F. Yang, L. Kong, and Z. Hou, "A CLM-based method of indoor affordance areas classification for service robots," *Jiqiren/Robot,* vol. 40, no. 2, pp. 188-194, 2018.

[5] N. Ali *et al.*, "A hybrid geometric spatial image representation for scene classification," *PloS one,* vol. 13, no. 9, p. e0203339, 2018.

[6] L. Wang, S. Guo, W. Huang, Y. Xiong, and Y. Qiao, "Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs," *IEEE Transactions on Image Processing,* vol. 26, no. 4, pp. 2055-2068, 2017.

[7] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365,* 2015.

[8] A. Quattoni and A. Torralba, "Recognizing indoor scenes," 2009: IEEE, pp. 413-420.

[9] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," 2012: IEEE, pp. 2751-2758.

[10] H. Sun, Z. Meng, P. Y. Tao, and M. H. Ang, "Scene recognition and object detection in a unified convolutional neural network on a mobile manipulator," 2018: IEEE, pp. 1-5.

[11] X. Song, S. Jiang, B. Wang, C. Chen, and G. Chen, "Image representations with spatial object-to-object relations for RGB-D scene recognition," *IEEE Transactions on Image Processing,* vol. 29, pp. 525-537, 2019.

[12] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015 2015, pp. 1440-1448.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016, pp. 779-788.

[14] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *arXiv preprint,* vol. 1612, 2016.

[15] L.-J. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," 2009: IEEE, pp. 2036-2043.

[16] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Learning hierarchical models of scenes, objects, and parts," 2005, vol. 2: IEEE, pp. 1331-1338.

[17] P. Espinace, T. Kollar, A. Soto, and N. Roy, "Indoor scene recognition through object detection," 2010: IEEE, pp. 1406-1413.

[18] P. Espinace, T. Kollar, N. Roy, and A. Soto, "Indoor scene recognition by a mobile robot through adaptive object detection," *Robotics and Autonomous Systems,* vol. 61, no. 9, pp. 932-947, 2013.

[19] J. West, D. Ventura, and S. Warnick, "Spring research presentation: A theoretical foundation for inductive transfer," *Brigham Young University, College of Physical and Mathematical Sciences,* vol. 1, no. 08, 2007.