DEEP RESIDUAL NETWORK WITH TRANSFER LEARNING FOR IMAGE-BASED STRUCTURAL DAMAGE RECOGNITION

Y. Q. Gao\textsuperscript{1}, K. B. Li\textsuperscript{2}, K. M. Mosalam\textsuperscript{3} and S. Gunay\textsuperscript{4}

ABSTRACT

In this data explosion epoch, structural health monitoring and rapid damage assessment after natural hazards have become new focuses in civil engineering. In this paper, state-of-the-art Deep Learning (DL) techniques were introduced and implemented for image-based structural damage recognition by using Convolutional Neural Networks (CNNs) and Deep Residual Network (ResNet). First, a new ImageNet, namely Structural ImageNet, was proposed for the benchmarking and assessment of progress in vision-based problems in structural engineering area. A hierarchy tree like evaluation procedure was planned for Structural ImageNet, and four simplified baseline detection tasks were designed and tested: component type, spalling condition, damage level and damage type. Selected from the database, a small experimental dataset was established with 2000 manually labeled images based on domain knowledge. Due to the relatively small number of labeled images, instead of training CNN from scratch, Transfer Learning (TL) approach was applied in two different ways with feature extractor (FE) and fine-tuning (FT). Four recognition experiments were conducted on different configurations of ResNet by both FE and FT. Our results show that both the FE and FT approach created highly accurate neural network classifiers. In the visualization and explanation section, we use local linear approximations of the model to interpret and explain the model’s behavior with respect to the images. In summary, it is demonstrated that deep CNN realizes structural damage recognition through image in a quite accurate way. Furthermore, establishment of a larger Structural ImageNet and application of more complicated tasks such as damage localization are proposed as seed of future work.

\textsuperscript{1}Graduate Student Researcher, Dept. of Civil Engineering, University of California, Berkeley, CA 94720, United State & Tsinghua Berkeley Shenzhen Institute (TBSI), Shenzhen, China (email: gaoyuqing@berkeley.edu)
\textsuperscript{2}Graduate Student Researcher, Dept. of Industrial Engineering & Operations Research Department, University of California, Berkeley, CA 94720, United State (email: kbl4ew@berkeley.edu)
\textsuperscript{3}Professor, Dept. of Civil Engineering, University of California, Berkeley, CA 94720, United State & Tsinghua Berkeley Shenzhen Institute (TBSI), Shenzhen, China (email: mosalam@berkeley.edu)
\textsuperscript{4}Project Scientist, Dept. of Civil Engineering, University of California, Berkeley, CA 94720, United State (email: selimgunay@berkeley.edu)
Deep Residual Network with Transfer Learning for Image-based Structural Damage Recognition

Y. Q. Gao\textsuperscript{1}, K. B. Li\textsuperscript{2}, K. M. Mosalam\textsuperscript{3} and S. Gunay\textsuperscript{4}

ABSTRACT

In this data explosion epoch, structural health monitoring and rapid damage assessment after natural hazards have become new focuses in civil engineering. In this paper, state-of-the-art Deep Learning (DL) techniques were introduced and implemented for image-based structural damage recognition by using Convolutional Neural Networks (CNNs) and Deep Residual Network (ResNet). First, a new ImageNet, namely Structural ImageNet, was proposed for the benchmarking and assessment of progress in vision-based problems in structural engineering area. A hierarchy tree like evaluation procedure was planned for Structural ImageNet, and four simplified baseline detection tasks were designed and tested: component type, spalling condition, damage level and damage type. Selected from the database, a small experimental dataset was established with 2000 manually labeled images based on domain knowledge. Due to the relatively small number of labeled images, instead of training CNN from scratch, Transfer Learning (TL) approach was applied in two different ways with feature extractor (FE) and fine-tuning (FT). Four recognition experiments were conducted on different configurations of ResNet by both FE and FT. Our results show that both the FE and FT approach created highly accurate neural network classifiers. In the visualization and explanation section, we use local linear approximations of the model to interpret and explain the model’s behavior with respect to the images. In summary, it is demonstrated that deep CNN realizes structural damage recognition through image in a quite accurate way. Furthermore, establishment of a larger Structural ImageNet and application of more complicated tasks such as damage localization are proposed as seed of future work.

Introduction

Structural Health Monitoring (SHM) and rapid damage assessment after natural hazards and disasters are important focuses in Civil Engineering in recent years. Moreover, Structural response records and images are ubiquitous and available for real time analysis. Meanwhile, Artificial Intelligence (AI) and Machine Learning (ML) technologies are developing rapidly, especially in applications of Deep Learning (DL) in computer vision (CV). In addition, objective

\textsuperscript{1}Graduate Student Researcher, Dept. of Civil Engineering, University of California, Berkeley, CA, United State & Tsinghua Berkeley Shenzhen Institute (TBSI), Shenzhen, China (email: gaoyuqing@berkeley.edu)

\textsuperscript{2}Graduate Student Researcher, Dept. of Industrial Engineering & Operations Research, University of California, Berkeley, CA, United State (email: kbl4ew@berkeley.edu)

\textsuperscript{3}Professor, Dept. of Civil Engineering, University of California, Berkeley, CA 94720, United State & Tsinghua Berkeley Shenzhen Institute (TBSI), Shenzhen, China (email: mosalam@berkeley.edu)

\textsuperscript{4}Project scientist, Dept. of Civil Engineering, University of California, Berkeley, CA, United State (email: selimgunay@berkeley.edu)
of the implementation of ML and DL is to make computers perform labor-intensive repetitive tasks to facilitate learning from past experiences in an efficient way. Thus, it is timely to implement the state-of-art DL technologies to Civil Engineering applications and evaluate its potential power.

Convolutional Neural Networks (CNNs) have been at the heart of recent spectacular advances in DL. Compared with traditional CV and ML approaches, CNNs allows practitioners to avoid manually deriving features and, instead, opt for data-driven feature engineering via multilayer learning and representation learning. Modern advancements in computational power and the ubiquity of Graphic Processing Units (GPUs) allow modelers to design and train CNNs of increasing depth, such as VGG Net [1], Google Net [2] and Deep Residual Net [3]. The combination with vast amounts of data allow better representation learning in a data-driven fashion. Besides of great development of computer hardware, recent wide applications of CNNs are attributed to the boosting from ImageNet Large Scale Visual Recognition Challenge (ILSVRC), aka the ImageNet Challenge [4], since 2012. The ImageNet Challenge has been held for several years, aiming to evaluate algorithms for object detection and image classification. In this ImageNet, there are more than 1.2 million images collected from variety of domains and labeled with 1000+ classes. In the ImageNet Challenge of 2015, Residual Net (ResNet), the new CNN architecture developed by Microsoft research team [3] achieved the first place in all main competition tracks, and it indicated the greatly improvement of recognition accuracy than previous models and methods.

Transfer Learning (TL) is a powerful ML technique that attracts vast attention in both research and industry applications. It applies the knowledge from source domains to target domains which might be related but different [5], making several pre-trained models more useful towards other datasets. The major advantage of TL is that it can relax the requirement of training for a large number of data, through tuning part of the parameters from pre-trained model in source domain with few labeled data in target domain. Many experiments were conducted to demonstrate the efficiency and promising results of the application of TL [5-7]. In TL, there are usually two common strategies: feature extractor and fine-tuning, which will be discussed later in this paper.

Civil Engineering applications have not fully benefited yet from above data-driven computer vision technologies partially due to the fact that relevant labeled data are costly and time consuming to obtain, e.g. labeling structural damage data require significant amount of domain-specific professional knowledge in Structural Engineering. Therefore, TL and more general pre-trained models is a good way to mitigate the shortage of labeled structural data, but there are only a few studies on TL applied for vision-based SHM and reconnaissance efforts. Thus, inspired by the ImageNet effort and in the hopes of better applying TL to Civil Engineering applications, we propose to create a Structural ImageNet to vision problems in the Structural Engineering area, benchmark and assess the performance of ML on multiple structural damage recognition tasks.

Decision makers often mistrust highly accurate ML/DL models due to its blackbox nature. Model explainability and reproducibility are therefore important especially when the predictions are used in critical decision making settings. The answer is not always, because the accuracy of their model may be overestimated. Therefore, more insight should be placed on explaining model and understanding how CNN works. Many recent studies are focusing on this interpretation topic, and a variety of approaches were proposed, such as using class activation map (CAM) with heat-map plots to visualize the most active areas of CNN [8] using Local Interpretable Model-agnostic Explanations (LIME) with a contiguous patch of similar pixels (super-pixels) to show the interpretable representations [9] and etc.

The rest of the paper is organized as follows (1) propose the Structural ImageNet with four...
baseline recognition tasks in object level; (2) apply TL approach based on ResNet-50 and 101 through FE and FT to achieve well-performed classifiers according to different detection tasks; (3) follow LIME algorithm to explain the model’s performance and offer insights on the CNNs.

**Structural ImageNet**

Inspired by the establishment of the ImageNet and the idea of TL, it is proposed to construct a Structural ImageNet which contains images relevant to Structural Engineering, such as buildings, bridges, columns, walls, etc., with both structural damaged and undamaged states, which can be used for recognition and vision problems in Structural Engineering. Analogous to ImageNet, in order to construct this Structural ImageNet, a large amount of images are needed. Up to now, over 20,000 structural images were collected from NISEE, DESIGN SAFE, EERI Learning from Earthquake Reconnaissance Archive, Google Image and Baidu Image to form the base of Structural ImageNet.

**Figure 1. Hierarchy tree of Structural ImageNet**

Analogous to the classification and localization tasks in ImageNet challenge, for the constructed Structural ImageNet, similar recognition tasks are expected to be designed for structural damage recognition and evaluation. Based on past experiences from reconnaissance [10-11], several issues affect the safety of structures, namely the type of the damaged component, the severity of damage in the component, and the type of damage. Due to the fact that images collected from the reconnaissance broadly vary, such as different distances from objects, camera angles, emphasized targets and etc., it is easier to think that they can be clustered in different levels, i.e. image taken from very close distance or only contains part of the component belongs to pixel level; major targets in images such as single or multiple components can be considered as object level; image that contains most part of the structures is clustered as structural level. Moreover, the corresponding evaluation criteria will differ for different levels, i.e. image in pixel level will be more interested in material type and damage status but image in structural level will more focus
on structural type or failure status. Therefore, we would like to propose a new processing method with a hierarchy tree structure, Fig.1. Instead of doing scene classification, (1) raw image will be clustered to different levels; (2) according its level, corresponding recognition tasks can be applied layer by layer following this hierarchy structure; (3) each node can be seen as one recognition task or classifier, and the output of each node can be seen as characteristic or feature of the image. As a pilot study, this paper focuses on one branch of the tree, starting from the object level to separate recognition tasks independently to alleviate the strong dependency of large-scale labeled data. Therefore, the following categories are considered: (1) binary classification task for component type identification (beam-column/wall), (2) binary classification task for spalling condition check (no spalling/spalling), (3) 3-classes classification task for damage level evaluation (no/minor/moderate to heavy damage), and (4) 4-classes classification task for damage type determination (no/flexural/shear/combined). In application, each image is labeled with 4 tags as the 4 attributes according to these tasks, where sample images are shown in Fig.2.

Figure 2. Sample images used in experiments: (1) component type identification: (a)-(c) are beam or column, and (d)-(f) are wall; (2) spalling condition: (g)-(i) are no spalling, and (j)-(l) are spalling; (3) damage level evaluation: (m) and (n) are no damage, (o) and (p) are minor damage, and (q) and (r) are moderate to heavy damage; (4) damage type determination: (s) and (t) are flexural damage, (u) and (v) are shear damage, (w) and (x) are combined damage

**Deep Residual Network**

Even though a deeper network has more trainable parameters than a shallow network, the increasing complexity might help solving more complicated problems and improving performance. However, gradients vanishing/exploding and degradation [3] issues will occur with increasing network depth, which make it hard to train. To address this, a new CNN architecture with shortcut connection, named Residual Net (ResNet), was developed [3]. Like its name, reformatting the layers with reference to the input through shortcut connection, the network will learn residual
function instead of learning unreferenced function which is the way as traditional CNN works. Those shortcuts act like highways and the gradients can easily flow back, resulting in faster training and support for stacking more layers. Mathematically, assume $H(x)$ is any desired mapping function, the traditional way assumes that convolutional operations fit $H(x)$; on the contrary, in ResNet, these convolutional operations are used to fit the residual function $F(x)$, and then mapping can be represented as the sum of function of input and residual terms $H(x) = F(x) + h(x)$. If the additive term $h(x)$ is original input, this shortcut connection is thought as identify mapping, Fig. 3(a). Besides, more complicated mapping was also designed for shortcut connection, like using convolutional shortcut, Fig. 3(b). More detailed derivation and validation were studied in [3]. Moreover, CNN structures shown in Fig. 3(a) and (b) are denoted as identity residual unit and convolutional residual unit respectively.

In this study, ResNet-50 and 101 were implemented as the baseline recognition detectors, both of which consist of stacking multiple identity and convolutional residual units. The detailed CNN configuration is shown in Fig. 4. The first residual unit in one convolutional block (conv block) is set as convolutional unit and latter units are all identity. The only difference in ResNet-50 and 101 is that ResNet-101 has more identity residual units in conv block 4, which might help with extracting mid to high level features. Interested readers can find more information in [3].

![Figure 3. Two kinds of shortcut connection in ResNet](image)

![Figure 4. Configuration of ResNet-50 (101)](image)

**Transfer Learning**

In some engineering fields, ML has already achieved a great success in recent years, but it heavily relies on the support of large amount of labeled data, while useful data sometimes are very expensive in real applications. In the absence of large quantity of data, TL is an effective tool to maximize the utilization of existing data.

Compared with general TL approach, the details of using TL in CNN are somewhat
different. In the CNN, parameters in shallower layers represent simple low-level features, such as color, texture and edges while parameters in deeper layers attempt to capture more complicated and abstract high-level features [12]. Therefore, the major objective of TL in CNN is to make use of parameters in a well-trained model from the dataset in source domain to help with training using the dataset in the target domain. If the two datasets are similar, some low-level features of CNN are proposed to be similar, which can be shared, and high-level features can be tuned by TL. Thus, usually in CNN, TL plays the similar roles of feature-representation-transfer and parameter-transfer.

In general, there are two different strategies while conducting TL in deep CNN, feature extractor (FE) and fine-tuning (FT). FE works as a feature extraction tool which greatly decreases the training time. Instead of training from bottom to top through large number of convolutional layers (conv layers) multiple times, only features extracted by CNN will be fed into a shallow fully connected neural network or even linear classifiers, which are fast to train. Because the expensive convolutional operation in feedforward procedure only operates once, FE spends much less time than regular CNN training. Different with FE, FT will retrain some parts of the CNN. Compared with regular CNN training from scratch, the parameters in some parts of the layers are fixed where their backward calculations are skipped and only parameters in unfixed layers are adjust by the gradient descent algorithm in backpropagation. Thus, it is also less time consuming than the regular CNN training but more than FE.

**Experiments**

In our experiments, two ResNet architectures (50 layers and 101 layers) were implemented, and both we experiment with both FE and FT. Similar with ImageNet challenge, the input size of images is 224x224. But in order to avoid inconsistency of labels, “random crop method” [1, 3] used in preprocessing procedure was abandoned, and the images were just scaled to target size without significant distortion, because the original aspect ratio was nearly 1~1.05. The evaluation criterial are training and testing accuracy, defined as the number of correct predictions, i.e. sum of true positive and true negatives divided by the number of data for training and test respectively. The total number of labeled images is 2000, and they were empirically split as training and test sets with 1600 and 400 images respectively. The label ratios in both sets keep the same.

The implementation of experiments was conducted on TensorFlow platform and performed on Alienware 15R3 with single GPU. (CPU: Intel(R) Core i7-7700HQ @2.80GHz, RAM:16.0 GB and GPU: Nvidia Geforce GTX 1060)

**Feature extractor experiment**

While in implementing FE, all parameters were fixed in conv blocks, and the output tensor before dense layer (Fig. 4) was extracted as features. Then these features were taken as new input of a two-layers fully-connected network [13] with 256 neurons and another Softmax layer. The weights were initialized from a normal distribution with zero mean and $10^{-2}$ variance (biases are initialized with zero) and optimized the categorical cross-entropy loss using Adam method [14] based on backpropagation, through assigning three key parameters $\beta_1$, $\beta_2$ and $\epsilon$ as empirical values $0.9$, $0.999$ and $10^{-8}$. The learning rate was set to $10^{-3}$ and manually changed to $10^{-4}$ after 200 epoch training.

The computation results are shown in Fig. 5, all tasks converge fast and performance would not improve much even though adjusting to lower learning rate after 200 training epochs. This
observation was attributed to fixed features extracted by conv blocks. Another observation is that the ResNet-50 and 101 share similar performance, both prediction accuracy of two binary cases reach to nearly 80% and 66% for damage level evaluation without overfitting (training accuracy higher than test accuracy). But, for damage type determination both accuracies are unsatisfactory, which are just slightly higher than random guess of 4-classes classification. This issue is attributed to less information and inadequate features extracted by conv blocks, and the performance improvement is expected by increasing data, using different CNN architectures and fine-tuning part of CNN layers.

FE only consumes small amount of computational time, roughly 28, 48 seconds for feature extraction process of ResNet-50 and 101 respectively and only 0.2 second for both in each training epoch. Since performances of FE for ResNet-50 and 101 have no much difference, ResNet-50 with FE can be placed as preliminary analysis for component type identification, spallings check and damage level evaluation. For damage type case, more studies should be conducted to mitigate the misclassification risks.

![Figure 5. History of accuracy in Feature extractor](image)

**Fine-tuning experiment**

Compared with FE, FT retrains later parts of the CNN. For simplicity, the conv blocks were kept instead of just retraining fewer conv layers. Since the major difference of ResNet-50 and 101 is number of identity residual units in block 4, in this experiment only conv block 4 and 5 were tuned for both ResNet-50 and 101 considering the concerns of overfitting due to small training set. Since forward propagation is inevitable and performed at each epoch, online data augmentation can be applied by a random selection of several transformations such as zoom in/out, flipping horizontally or vertically, rotation with random angles within 5 degrees, and translation in random directions (up/down/left/right). Stochastic gradient descent (SGD) [13] was applied
with momentum of 0.9 and an exponential decay learning schedule, whose initial decay value and decay rate are $10^{-4}$ and $10^{-6}$ respectively.

![Figure 6. History of accuracy in Fine-tuning](image)

The computation results are shown in Fig. 6. FT greatly improves the prediction performance for both ResNet. The difference between two nets can be observed but not prominent. The accuracies of first three cases reach over 90%, and even for damage type recognition, test accuracy increases over 70% within 20% overfitting. This observation also indicates for current small dataset, ResNet-50 already holds the complexity for recognition, additive with computational efficiency (roughly 82 seconds per epoch for ResNet-50 and 160 seconds for ResNet-101), ResNet-50 can be a more efficient choice.

Another interesting observation is the plateau in test accuracy history at the beginning, which is attributed to incompatible initial parameter values and large learning rate. With online data augmentation and decaying learning rate, the model escaped the plateau within several training epochs, and finally achieves very promising results.

**Visualization and Explanation**

While in analysis, an explanation for model performance will be more convincing than just listing the recognition accuracy. In image classification, an interpretable representation is the presence or absence of super-pixel in target images [9], so the main idea is to judge whether this small area of image will influence the classification result. The following are several key steps while LIME is implemented for fine-tuned Resnet-50: (1) Make segmentations by quick shifting clustering method [15] as super pixels or instances; (2) perturb instances as new samples by add random noise or hiding some parts; (3) use perturbed new samples to do classification based on pre-trained ResNet-50 and obtain relative weights; (4) according to previous computation highlight
instances with positive weights towards a specific class, as it gives the explanation why the model chooses for this class label; (5) include the domain knowledge to judge whether the model is reasonable.

(a) Component type identification

(b) Spalling condition check

(c) Damage level evaluation

(d) Damage type determination

Figure 7. Explaining classification prediction made by ResNet-50 (Note: Green highlighted areas represent the super-pixels have positive influence on results, vice versa for red.)

Fig. 7 presents several successful explanations towards four tasks. It is worth noting that for binary cases Fig. 7 (a) and (b), the highlighted parts also illustrate the boundaries for column or spalling areas, which is exact the way engineers judging the image. For multi-classes case Fig. 7 (c) and (d), even though how the network makes decision on damage level and type is unknown, the highlighted parts cover most damaged parts, indicating that the network still focuses on the right target which makes the model more reasonable. In addition, LIME for fine-tuned ResNet-101 was also conducted, which presents similar performance but more refined and accurate for boundaries. Thus, ResNet type CNN architectures not only achieve satisfactory prediction accuracy but also be well explainable property with visualization.

Conclusions

This paper firstly proposes the concept and establishment of Structural ImageNet, and then several recognition tasks are designed correspondingly. The ideas of Transfer Learning and Deep Residual Convolutional Neural Network are introduced and implemented to pre-defined tasks. Both ResNet-50 and 101 compute fast under feature extractor but with limited accuracies especially for multi-classes, but the performance can be improved greatly through fine-tuning
some parts of the network. Very promising recognition accuracies are achieved and also can be
well-explained by the state-of-the-art LIME algorithm, which make the model more reasonable.

Besides these four baseline tasks, more complex recognition can be pursued, especially for
damage localization and quantification. More real engineering project extensions are expected,
such as combing with prevailing drone/robot-based detection to help with decision making in real-
time structural damage recognition and evaluation in post-disaster reconnaissance.

Acknowledgments

The authors would like to acknowledge the funding support of TBSI, and appreciate the
assistance from graduate students Luka Liang, Yujia Xu and Renjie Wu.

References

7. Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using
9. Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: Explaining the predictions of any classifier. In
Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
10. Li B, Mosalam KM. Seismic performance of reinforced-concrete stairways during the 2008 Wenchuan
11. Mosalam KM, Takhirov SM, Park S. Applications of laser scanning to structures in laboratory tests and field
surveys. Structural Control and Health Monitoring. 2014 Jan 1;21(1):115-34.
computer vision 2014 Sep 6 (pp. 818-833). Springer, Cham.