Examining the Effects of Deep Learning Model Structure on Model Interpretability for Time-Series Classifications in Fire Research

Wai Cheong Tam[1,‡], Linhao Fan[2,‡], Qi Tong[1], Hongqiang Fang[1]

[1] Fire Research Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA
[2] School of Mechanics and Safety Engineering, Zhengzhou University, Zhengzhou, China

Email: waicheong.tam@nist.gov, ‡ joint first authors

Abstract. This present work utilizes an interpretability model to understand and explain the decisions of deep learning models. The use of DeepLIFT is proposed and attributions of a study case are obtained. Benchmarking against two other interpretability models, namely Grad-CAM and dCAM, is conducted. Results show that DeepLIFT can provide precise attributions to the model inputs in both temporal and spatial directions. A parametric study is also carried out to understand the effects of deep learning model structure on the attributions obtained from the interpretability model. Ten different convolutional neural network model structures are considered. Three important observations are made: 1) changes in the model structure have minor effects on the attributions in the temporal direction, but 2) they have negligible effects on attributions in the spatial direction, and 3) convolutional layers need to be fixed to avoid attribution discrepancies. By understanding the model decision and the resulting effects of the model structure, it is hoped that this work can contribute to the development of trustworthy deep learning models for the fire research community.

1. Introduction

Deep learning (DL) has become a robust model development approach for time-series classification tasks in fire risk prevention in residential homes [1,2] and smart firefighting [3-4]. In [1], a four-layer artificial neural network model was developed to detect pre-ignition conditions with enough time to prevent oil ignition on a kitchen cooktop. The proposed model outperformed the baseline rule-based model and had fewer false negatives. Kou and her colleagues [2] developed a gated recurrent neural network (GRU) model to detect the fire location and its intensity. The GRU model had an accuracy of about 95.4 % and it could make predictions in less than a second. In [3], Zhang et al. trained a forecasting model using long short-term memory with real experimental data. The authors were successful in accurately predicting flashover 20 s before it occurs. Li and her colleagues [4] used convolutional neural networks (CNN) and regularization techniques to develop a multi-class classification model. Even when dealing with complex 12-lead electrocardiogram data with motion artifacts, the CNN model used only a short segment of the data (i.e., 6 s or 12 s) and achieved an overall prediction accuracy of about 98 %. It can be seen that the use of DL will help to provide accurate real-time actionable information about the impending hazardous events in which fire losses, injuries, and deaths can be greatly reduced. However, the training process of the DL models lacks transparency. The models from [1-4] are rather Black-Boxes. For that, the implementation of these DL models in the fire safety/protection community, where human lives are usually involved, is limited.

Research efforts have been made to provide model transparency to understand what specific information the DL models have learned to make a decision. For time-series classifications, Fan et al.

[5] used a dimension-wise class activation map (dCAM) to obtain attributions[1] for the model inputs in both temporal and spatial directions and they observed that their proposed model was capable of making use of human-agreeable information (i.e., focusing on the last 15 s of temperature information in the room of fire origin) to predict the onset of flashover. For image classifications, Wang et al. [6] utilized a gradient-weighted class activation map (Grad-CAM) to identify the parts of an input image that most impact the classification. Their results revealed that the model was focusing on the flaming objects and the background, the color, and the brightness of the images did not have substantial effects on the model performance if the model was sufficiently trained. To the best of the authors' knowledge, the studies from [5,6] are the only available literature for understanding/interpreting the DL model decision in the fire research community. However, knowledge gaps exist. For example, 1) it is not clear what interpretability model is more suitable for multi-input time-series classification tasks and 2) what the effects of DL model structure on the attributions obtained from the interpretability model are. Therefore, this research work aims to address these two knowledge gaps. In this paper, the use of a new interpretability model (DeepLIFT) is explored. Also, attributions obtained from DeepLIFT are compared to those obtained from dCAM and Grad-CAM. Lastly, the effects of DL model structure on attributions obtained from DeepLIFT are investigated. The results of this study aim to provide more transparency for the DL model and help to build trustworthy DL models for multi-input time-series classifications in fire research.

2. Interpretability Model (DeepLIFT)

Deep Learning Important FeaTures (DeepLIFT) is used as the interpretability model for this study. DeepLIFT was proposed by Shrikumar et al. [7] to interpret DL-based genetic models. Its core idea is to calculate the attribution by measuring the difference in the activation value between the actual inputs and a reference input. Fig. 1 illustrates the attribution determination process using DeepLIFT.

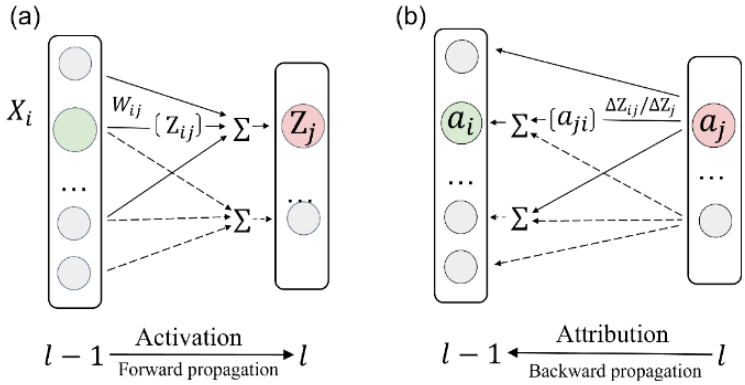

Fig. 1. a) The forward propagation process to obtain the activation value and b) the backward propagation to determine the attribution in DeepLIFT.

As shown in Fig. 1, the determination process requires two steps: i) a forward propagation and ii) a backward propagation. Fig. 1a shows the forward propagation process. Using the normalized inputs/features, X, from neuron i in the previous layer (l – 1), the activation value, Z, for neuron, j, in layer l can be obtained. Mathematically, the activation value is calculated as:

$$Z_j = \sum_i^{M_{l-1}} Z_{ij} = \sum_i^{M_{l-1}} f(W_{ij} X_i^{(l-1)}) \quad (1)$$

where M_{l-1} is the total number of neurons that are connected to neuron j in the layer (l - 1), f is a function representing the neural network operations and the mathematical operations for the nonlinear fittings and $W_{ij}$is the weight corresponding to neurons i and j which is being determined during the model training process. By aggregating all activation values over neuron i related to j, the final

[1] Attribution provides a score or ranking over features, conveying the (relative) importance of each feature to the model's output.

activation value Z_j for neuron *j* is determined. The same process is carried out to obtain the activation values for all layers and it should be noted that the activation values will be used to determine the attribution values in the backward propagation process.

Figure 1b shows the backward propagation process. It can be seen from the figure that the information flow from the backward propagation process is completely opposite to that in the forward propagation process. The attribution values to neurons *i* in the (*l* - 1) layer are obtained based on the attribution values from neurons *j* in the "previous" layer (*l*) and the corresponding contribution factors. Mathematically, the attribution value, *a*, between neurons *j* and *i* is described as:

$$a_{ji}^{(l-1)} = \frac{\Delta Z_{ij}}{\Delta Z_j} \cdot a_j^{(l)} \tag{2}$$

where $\Delta Z_{ij}/\Delta Z_j$ is the contribution factor between neurons *j* and *i* which regulates the importance of a_{ji} in the (*l* - 1) layer based on a_j in layer (*l*). For notation purposes, a represents the intermediate attribution values in various layers whereas *A* represents the final attribution results to the inputs. The sum of $\Delta Z_{ij}/\Delta Z_j$ over all neuron i is unity. ΔZ_{ij} is the measure of difference between the activation value, Z_{ij}, and the activation value obtained based on a reference input value, *r*. Mathematically, the change of activation value between neuron *i* and *j*, Z_{ij}, is given by:

$$\Delta Z_{ij} = Z_{ij} - \tilde{Z}_{ij} \tag{3}$$

where Z_{ij} is the activation value evaluated using Eqn. 1. $\tilde{Z}_{ij}$ is the activation value based on the reference value and it is determined using the following expression:

$$\tilde{Z}_{ij} = f(W_{ij} r_i^{(l-1)}) \tag{4}$$

with r_i being the reference value in the (*l* - 1) layer. In principle, the reference value is used to compute the state of neuron *i* when it receives arbitrary information.

To obtain the overall difference in the activation value of neuron *j*, the following expression is used:

$$\Delta Z_j = \sum_i^{M_{l-1}} \Delta Z_{ij} \tag{5}$$

Similarly, the overall attribution for neuron i in the (*l* - 1) layer, a_i, is given as:

$$a_i^{(l-1)} = \sum_j^{N_l} a_{ji}^{(l-1)} \tag{6}$$

where N_l is the total number of neurons that are connected to neuron *i* in the *l* layer. The attribution evaluation process using Eqn. 2 to Eqn. 6 is carried from the last layer of the model to its first layer. In the last layer, the model outputs (i.e., the probability of each class) are used as the attribution values. With that, the attribution results to the inputs can be determined and they are highly correlated to the model outputs.

3. Results and Discussion

A flashover forecasting model is developed for a single-story ranch house[2] with a living room (LR), a kitchen (K), a dining room (D), a hallway (H), and three bedrooms (B1, B2, and B3). CFAST [9] is used to obtain the temperature data at a location that is 0.02 m away from the ceiling[2]. 17,657 numerical experiments with a wide range of fire conditions are conducted. Three fire locations are considered and they are on the floor and at the center of 1) LR, 2) K, and 3) D. The fire growth rate ranges from 3.290-e5 kW/s^2 and 4.139e-2 kW/s^2. The temperature profiles of each compartment for a flashover instance (i.e., 300 s temperature data) from a fire case in the LR are shown in Fig. 2a. The

[2] Due to space limitation, readers can refer to [8] to see the complete building layout and the sensor locations. Histograms for fire locations, fire growth, and door and window openings are provided in the Appendix A.

labeling temperature is not shown in the figure for simplicity but the upper layer gas temperature of 600 °C [10] in the room of fire origin (i.e., LR in this case) is used to determine the onset of flashover. Data preprocessing is carried out to obtain the training, validation, and testing subsets. Ten convolutional neural network (CNN) models are trained. Table 1 shows the model structures with i) different numbers of CNN layers, ii) kernel size, and iii) different numbers of fully connected layers. The number of kernels is always 16. Strip size of 1 is used for all CNN layers and global maximum pooling is used in the last CNN layer. Motivated by [11], 60 %, 20 %, and 20 % of the data are used for training, validation, and testing, respectively. Early-stopping with a patience number of 20 is used to avoid overfitting. The average accuracy, precision, and recall of these models are 90.8 %, 89.4 %, and 92.5 % with a standard deviation of 0.6 %, 1.0 %, and 0.9 %, respectively. These models are used to conduct a parametric study to understand the effects of the DL model structure on the interpretability model attributions which will be shown in Sec 3.1 and 3.2.

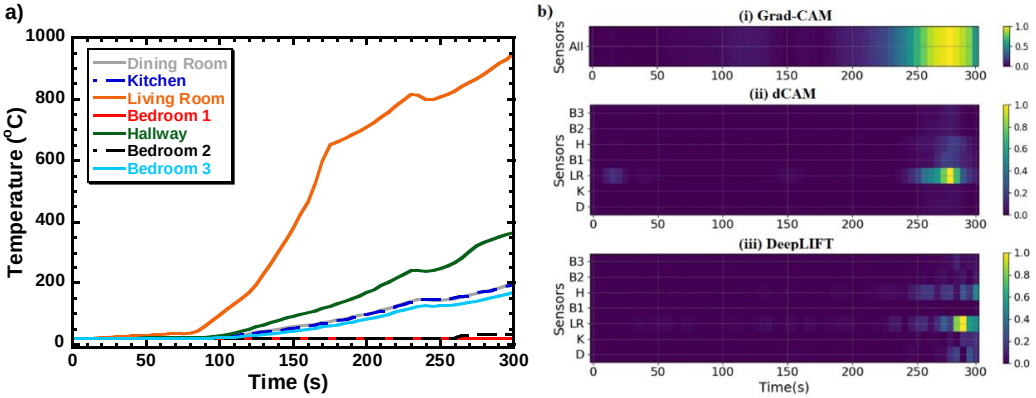

Fig. 2. a) Temperature profiles for a flashover instance and b) the corresponding attributions obtained from (i) Grad-CAM, (ii) dCAM, and (iii) DeepLIFT.

Table 1. Model structure summary for the ten different CNN models.

	CNN Model									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CNN layers	6	6	6	3	9	6	6	6*	6*	6*
Kernel size	3	5	7	5	5	5	5	5	5	5
Fully connect layers	1	1	1	1	1	2	3	1	2	3

* Freezing CNN layers from training to fix the learned features.

Figure 2b shows the attributions corresponding to the temperature profiles shown in Fig. 2a and the attributions are obtained from three different interpretability models: Grad-CAM [6], dCAM [5], and DeepLIFT. The value of the attributions is ranging from 0 to 1. The higher the value, the more relevant/important the responding temperature element is for the model prediction.

There are three major observations. Firstly, as seen in Fig. 2b(i), Grad-CAM is only able to capture attributions in the temporal direction. The attributions are identical for all compartments. For that, Grad-CAM cannot provide any useful insights about the temperature importance in the spatial direction. Secondly, although dCAM does provide distinct attributions for both the temporal and spatial directions, the attributions are problematic. As seen in Fig. 2b(ii), false attributions are observed at around 275 s in B1. Since the temperature in B1 remains at relatively room temperature, this temperature information should not be useful for model decisions. Finally, attributions from Grad-CAM, dCAM, and DeepLIFT all indicate that the LR temperatures towards the end are the most important data for the model prediction. Physically, this observation makes sense because human judgment would also rely on this LR temperature segment. Moreover, it is worth mentioning that the attributions from DeepLIFT offer additional insights (i.e., the H temperature staring from around 250 s and the K and the D temperature towards the end). Based on these observations, DeepLIFT is a more

reliable interpretability model for the multi-input time-series classification task. In the subsections below, the effects of the model structure on the attributions obtained from DeepLIFT are examined.

3.1. Effects from CNN Layers

Figures 3a-c show the attributions for Model 1-3 that have a kernel size of three, five, and seven, respectively, and all models have six CNN layers and one fully connected layer. Based on the attributions, it can be seen in the figures that when the kernel size increases, the model tends to make use of more temperature elements, particularly in the LR and H temperature profiles, to make a prediction. Physically, the model with a larger kernel size can extract features simultaneously from a larger number of input elements. In addition, with a larger kernel size (i.e., Model 3 with a kernel size of seven), the model also pays attention to the LR temperature at around 50 s in which the temperature begins to raise and the D and H temperature towards the end of the data segments. For that, there appear to be more high-value attributions with a larger kernel size.

Figures 3d-f show the attributions for Model 4, 2, and 5, respectively, with three, six, and nine CNN layers and they all have a kernel size of five and one fully connected layer. As compared to the cases with increasing kernel size, similar behaviors are observed. With the increasing number of CNN layers, the model tends to pay additional attention to the LR temperature at around 50 s and the end of the temperature profile for the D and H. The results suggest that by increasing the complexity of the model (i.e., the kernel size or the number of CNN layers), the model has more learning capabilities to make use of additional temperature information to make a decision. Yet, it is worth noting that the overall trend of the attributions from all models is consistent and they all indicate that the LR and the H temperature are both important.

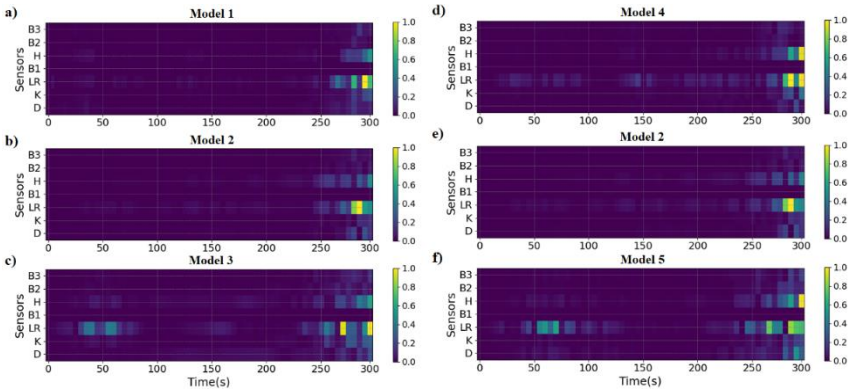

Fig. 3. Attributions of a flashover instance from Model a) 1, b) 2, c) 3, d) 4, e) 2, and f) 5.

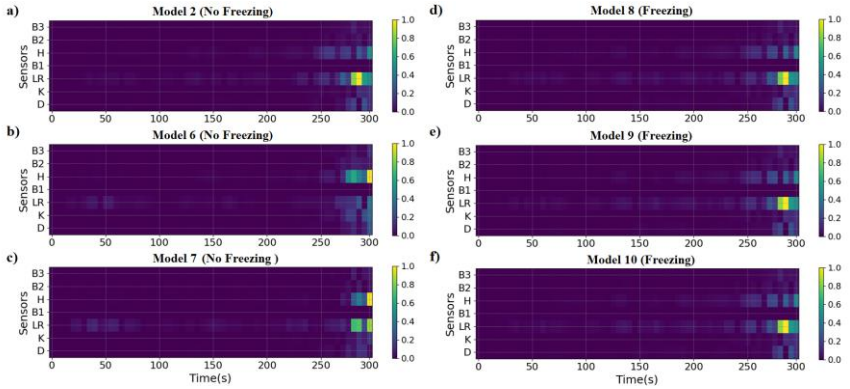

Fig. 4. Attributions of a flashover instance from Model a) 2, b) 6, and c) 7 without freezing the CNN layers and d) 8, e) 9, and f) 10 with the CNN layers being frozen.

3.2. Effects from Fully Connected (FC) Layers

Figures 4a-c show the attributions for Model 2, 6, and 7 which have one, two, and three fully connected layers, respectively, with identical number of CNN layers and kernel size. Changes to the attributions are observed by adding additional fully connected layers. However, the change in the fully connected layers should not lead to dramatic changes in the attributions because the FC layers are only used to combine the features obtained from the CNN layers to form the predictions. To confirm this hypothesis, Model 2, 6, and 7 are being retrained and the CNN layers are being frozen. This means that the CNN layers will not be trained and only the FC layers are being retrained. Figs. 4d-f show the attributions for Model 8, 9, and 10. It can be seen that the attributions are more consistent and this is because the CNN layers learned similar features. The result indicates that the change in attributions in both temporal and spatial directions is driven by the CNN layers.

4. Conclusion

Attributions for a flashover instance from Grad-CAM, dCAM, and DeepLIFT are obtained. As compared to Grad-CAM and dCAM, results show that DeepLIFT is more suitable for multi-time series classification tasks. DeepLIFT is capable of providing more reliable and more detailed attributions in both temporal and spatial directions which is useful for understanding what input data the model tends to focus on while making the prediction. Being able to interpret the DL model decision is important to gain the trust of the stakeholders and/or the end-users and to develop a trustworthy machine learning system to address fire safety concerns. In addition, results from a parametric study show that if the DL model is adequately trained where there is no overfitting and underfitting, the effects from the change of the model structure on the attribution for a specific flashover instance are observed in both temporal and spatial directions. However, it should be noted that additional studies are needed to verify the accuracy of the attributions. This work is underway and it will be presented in the near future.

References

[1] Mensch, A.E., Hamins, A., Tam, W.C., Lu, Z.J., Markell, K., You, C. and Kupferschmid, M., 2021. Sensors and machine learning models to prevent cooktop ignition and ignore normal cooking. *Fire technology*, pp.1-24.

[2] Kou, L., Wang, X., Guo, X., Zhu, J. and Zhang, H., 2021. Deep learning based inverse model for building fire source location and intensity estimation. *Fire Safety Journal*, *121*, p.103310.

[3] Zhang, T., Wang, Z., Wong, H.Y., Tam, W.C., Huang, X. and Xiao, F., 2022. Real-time forecast of compartment fire and flashover based on deep learning. *Fire Safety Journal*, *130*, p.103579.

[4] Li, J., Brown, C., Dzikowicz, D.J., Carey, M.G., Tam, W.C. and Huang, M.X., 2023. Towards real-time heart health monitoring in firefighting using convolutional neural networks. *Fire Safety Journal*, *140*, p.103852.

[5] Fan, L., Tam, W.C., Tong, Q., Fu, E.Y. and Liang, T., 2023. An explainable machine learning based flashover prediction model using dimension-wise class activation map. *Fire Safety Journal*, *140*, p.103849.

[6] Wang, Z., Zhang, T. and Huang, X., 2024. Explainable deep learning for image-driven fire calorimetry. *Applied Intelligence*, *54*(1), pp.1047-1062.

[7] Shrikumar, A., Greenside, P. and Kundaje, A., 2017, July. Learning important features through propagating activation differences. In *International conference on machine learning* (pp. 3145-3153).

[8] Tam, W.C., Fu, E.Y., Li, J., Peacock, R., Reneke, P., Ngai, G., Leong, H.V., Cleary, T. and Huang, M.X., 2023. Real-time flashover prediction model for multi-compartment building structures using attention based recurrent neural networks. *Expert Systems with Applications*, *223*, p.119899.

[9] Peacock, R.D., Reneke, P.A. and Forney, G.P., 2017. CFAST–consolidated model of fire growth and smoke transport (version 7) volume 2: user's guide. *NIST Technical Note 1889v2*.

[10] Peacock, R.D., Reneke, P.A., Bukowski, R.W. and Babrauskas, V., 1999. Defining flashover for fire hazard calculations. *Fire Safety Journal*, *32*(4), pp.331-345.

[11] Tam, W.C., Fu, E.Y., Li, J., Huang, X., Chen, J. and Huang, M.X., 2022. A spatial temporal graph neural network model for predicting flashover in arbitrary building floorplans. *Engineering Applications of Artificial Intelligence*, *115*, p.105258.