Fake News Detection with Hybrid CNN-LSTM

1st Kian Long Tan
Faculty of Information Science & Technology
Multimedia University
Jalan Ayer Keroh Lama, 75450, Melaka, Malaysia.
1181300023@student.mmu.edu.my

2nd Chin Poo Lee
Faculty of Information Science & Technology
Multimedia University
Jalan Ayer Keroh Lama, 75450, Melaka, Malaysia.
cplee@mmu.edu.my

3rd Kian Ming Lim
Faculty of Information Science & Technology
Multimedia University
Jalan Ayer Keroh Lama, 75450, Melaka, Malaysia.
kmlim@mmu.edu.my

2021 9th International Conference on Information and Communication Technology (ICoICT) | 978-1-6654-0447-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICoICT52021.2021.9527469

***Abstract*—In the past decades, information and communication technology has developed rapidly. Therefore, social media has become the main platform for people to share and spread information to others. Although social media has brought a lot of convenience to people, fake news also spread more rapidly than before. This situation has brought a destructive impact to people. In view of this, we propose a hybrid model of Convolutional Neural Network and Long Short-Term Memory for fake news detection. The Convolutional Neural Network model plays the role of extracting representative high-level sequence features whereas the Long Short-Term Memory model encodes the long-term dependencies of the sequence features. Two regularization techniques are applied to reduce the model complexity and to mitigate the overfitting problem. The empirical results demonstrate that the proposed Convolutional Neural Network -Long Short-Term Memory model yields the highest F1-score on four fake news datasets.**

Keywords—Fake news, fake news detection, machine learning, CNN, LSTM

I. INTRODUCTION

As the rapid development of technology, social media has gradually substituted the traditional news organization. This is because people can now get and share information via social media with just a click. It is no doubt that social media like Facebook, Twitter, Instagram and so like has brought a lot of convenience to people. Yet, in the century of information, people find it difficult to differentiate the credibility of the news whether it is fake or real. Consequently, it may harm the people who mistakenly believe the fake news. In view of this, fake news detection is required to help people differentiate the credibility of the news so that the readers would not be affected by the fake news.

In this paper, we propose a hybrid model consisting of CNN layers and LSTM layer, referred to as the CNN-LSTM model. The CNN layers learn the high-level representative features from the embedding vectors, subsequently the LSTM layer captures the long-term dependency of the embeddings. The CNN-LSTM model comprises embedding layer, convolutional layers, max pooling layers, LSTM layer, global max pooling layer and dense layers. The two regularization techniques are applied to solve the overfitting problem. Eventually, the experimental result and the loss plot have shown that the proposed model has better generalization capability and is less prone to overfitting problem.

This research is supported by the Fundamental Research Grant Scheme of the Ministry of Higher Education under the award number FRGS/1/2019/ICT02/MMU/03/7.

II. RELATED WORKS

The machine learning methods have shown promising results in a wide spectrum of applications [1-6]. This section reviews some techniques implemented for fake news detection. Ahmed et al. (2017) [7] used n-gram with two techniques, namely Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF)) for feature extraction. Six different classification algorithms were adopted for performance evaluation. The dataset was compiled from Reuters.com and kaggle.com. They performed data preprocessing, including stop words removal and stemming. Their experimental results show that the Linear Support Vector Machines with TF-IDF of unigram yields the highest accuracy.

Khan et al. (2019) [8] compared the performance of conventional machine learning methods and deep neural network models in fake news detection. In the data pre-processing, stemming, IP and URL removal, stop words removal and spelling correction were done. The n-gram, lexical and sentiment features were extracted to be used in each model. For deep neural network models, Global Vectors for Word Representation (GloVe) [9] was adopted as the representation vector for the words in the text. Among the conventional machine learning methods, Naive Bayes with bigram TF-IDF records the highest accuracy on the Combined Corpus. The authors indicated that all neural network models performed comparatively.

Bahad et al. (2019) [10] compared the performance of Convolutional Neural Networks (CNN), unidirectional LSTM and bidirectional LSTM. The stop words and punctuations were removed in the text preprocessing. They similarly adopted GloVe representation on the headline and body of the article. Some optimization methods were adopted to improve the performance. The empirical results demonstrate that the bidirectional LSTM outperforms the CNN method as it can better extract the local and position-invariant features thus making bidirectional LSTM performs well in long-range semantic dependency-based classification.

Thakur et al. (2020) [12] proposed a fake news detection model with two components, i.e., news extractor and stance detection. The news extractor identifies the keywords in the headlines and news articles. Two classification methods, namely Gradient Boosted Decision Tree and CNN were used in stance detection to detect whether the given text is fake or real news. Combining two classification methods show the best performance in fake news detection.

Authorized licensed use limited to: Indian Institute of Information Technology Lucknow. Downloaded on October 23,2024 at 13:26:27 UTC from IEEE Xplore. Restrictions apply.

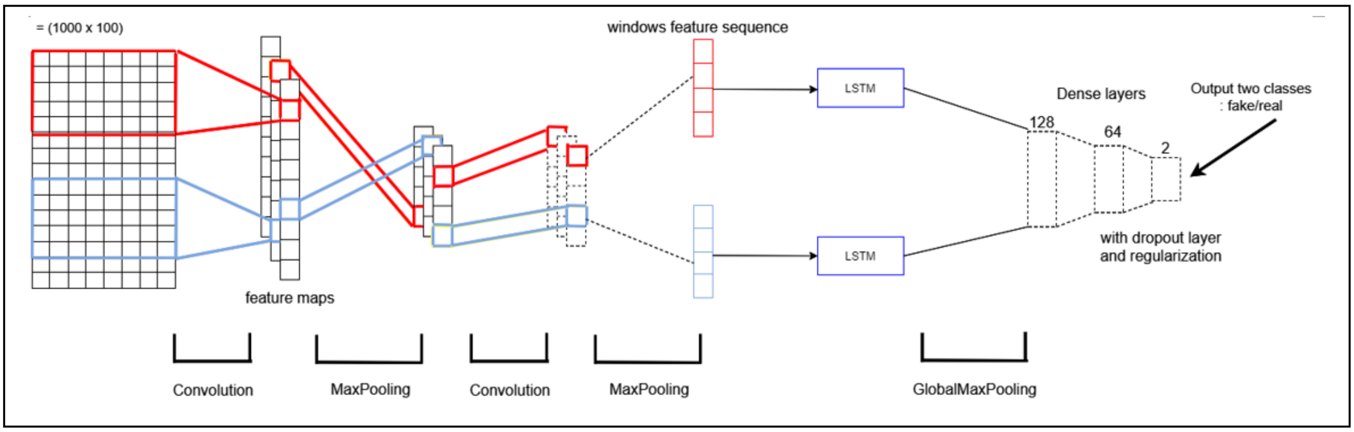

Fig. 1. The architecture of the CNN-LSTM model.

Kaur et al. (2020) [11] proposed a multi-level voting model that consists of three feature extraction techniques, namely TF-IDF, Count-Vectorizer and Hashing-Vectorizer. Using the features, twelve classifiers were included in the performance comparison. The pre-processing steps removed the duplicated data, non-categorized data, stop words and NAN values in the text. The classification using multi-level voting ensemble model that combines three classifiers obtains the highest accuracy compared to individual classifiers.

Majority of the existing methods consider the tokens in the news articles independently regardless of their position in the text sequence. Therefore, this paper proposes a CNN-LSTM model that encodes both the spatial and long-range semantic dependency of the text for fake news detection.

III. CNN-LSTM FOR FAKE NEWS DETECTION

This section explains the workflow of the fake news detection and the architecture of the proposed CNN-LSTM model. The datasets will firstly go through the preprocessing step to clean the noise and insignificant information in the text. Then, tokenization is performed where the texts are split into individual tokens. Subsequently, every token is represented in a vector. If the number of token in the text is less than the maximum token number, the sequence padding is applied to ensure the token vector has a fixed length. The sentiment classification is finally done by a CNN-LSTM model. The CNN-LSTM model integrates the strengths of CNN and LSTM. The CNN model extracts the high-level features from the token vector and the LSTM model captures the long-term dependency between the tokens. Fig. 1 and Fig. 2 depict the architecture of the CNN-LSTM model.

A. *Preprocessing*

The purpose of text pre-processing is to remove the insignificant information in the text dataset. Some pre-processing steps were applied, including unnecessary column removal, punctuations and stop words removal. The examples of unnecessary columns are ID, URL, etc. The punctuations and stop words are excluded to improve the performance of subsequent processes. The text is then standardized into lowercase to offer case insensitivity in the method.

B. *Tokenization and Sequence Padding*

Tokenization breaks the text into smaller units called tokens or words. The tokens will then form a word index vector where every index represents a token, as shown in Fig.2. The length of the word index vector is set to 1000 to ease computation. Zero padding is performed to fill the blanks if the number of tokens is less than 1000.

```
Layer (type)                  Output Shape
==============================================
embedding_1 (Embedding)       (None, 1000, 100)
dropout (Dropout)             (None, 1000, 100)
conv1d (Conv1D)               (None, 1000, 128)
max_pooling1d (MaxPooling1D)  (None, 200, 128)
dropout_1 (Dropout)           (None, 200, 128)
conv1d_1 (Conv1D)             (None, 200, 128)
max_pooling1d_1 (MaxPooling1  (None, 40, 128)
lstm (LSTM)                   (None, 40, 80)
global_max_pooling1d (Global  (None, 80)
dense (Dense)                 (None, 128)
dropout_2 (Dropout)           (None, 128)
dense_1 (Dense)               (None, 64)
dropout_3 (Dropout)           (None, 64)
dense_2 (Dense)               (None, 2)
==============================================
```

Fig. 2. The layers in the CNN-LSTM model.

Authorized licensed use limited to: Indian Institute of Information Technology Lucknow. Downloaded on October 23,2024 at 13:26:27 UTC from IEEE Xplore. Restrictions apply.

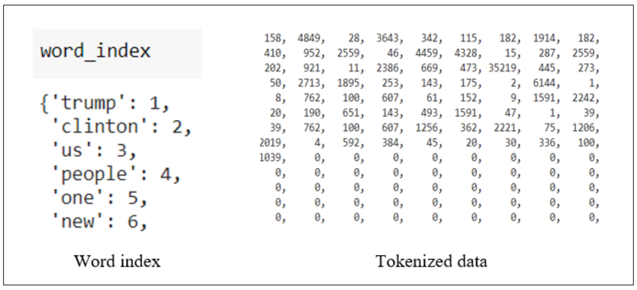

Fig. 3. The tokenized words (left) and the vector representation (right).

C. CNN-LSTM

The proposed CNN-LSTM model integrates the CNN model and LSTM model to capture the sequential relations in the text. The CNN model is good at extracting the high-level features of the text whereas the LSTM model learns the long-range dependencies from the high-level features. The proposed CNN-LSTM model consists of an embedding layer, two convolutional layers, two max pooling layers, one LSTM layer, one global max pooling layer and a series of dense layers. Some enhancements are also introduced into the CLSTM model to avoid overfitting problem and to optimize the training process.

1) Embedding layer

The embedding layer converts the tokens into the corresponding word embeddings matrix based on the word index vector. In word embeddings, the semantically similar words have closer vector values. Fig. 4 presents some examples of word embeddings projected onto the geometric space. In this paper, the length of the word index vector is 1000 and each index represents a word. To have a better weight initialization, the weight values of the words are adopted from the pretrained GloVe with dimension 100. Hence, the output size of the embedding layer is 1000 × 100. Fig. 4 shows the sample of the word embedding matrix where sequence length is 1000 and dimension size is 100. It means that each row represents a word using 100 vector values of weight.

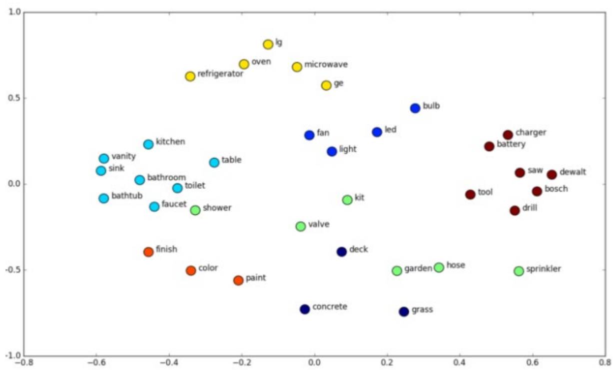

Fig. 4. Examples of word embeddings in geometric space.

2) 1D Convolutional layer

Convolutional layer is one of the core building blocks of CNN model. Convolutional layer extracts the representative high-level sequence feature from the input. A convolution is a linear operation that multiplies the input with a set of weights, known as a filter or a kernel, to produce a feature map as the output. In this model, the convolution operations involve a set of 128 filter with the size of 5 × 5.

Next, the feature map is passed through a nonlinear activation function, which is Rectified Linear Unit (ReLU) function. The benefit of ReLU function is that it only activates the neurons with value ≥ 0 in the operation. Hence, reducing the overfitting and makes the computation more efficient.

3) Max pooling layer

The max pooling layer obtains the maximum values in the feature map for de-noising and dimension reduction. This model used a filter size of 5 × 5 in the max pooling layer to reduce the output size to 20% before passing to the next layer.

4) LSTM layer

The LSTM layer aims to propagate the historical information via the chain of neural networks as well as to store the information of previous input. There are 3 gates in the LSTM, i.e., forget gate (f_t), input gate (I_t) and output gates (O_t). Given the input of the LSTM layer, the forget gates determine the information that is unnecessary and forgettable. The input gates decide which information is important and needs to be memorized. Whereas the output gates decide what value in the memory cell should be the output. The equations below show the gate variants in this model.

$$f_t = \sigma\left(w_f * (V_f h_{t-1}) + b_f\right) \tag{1}$$

$$i_t = \sigma\left(w_i * (V_i h_{t-1}) + b_i\right) \tag{2}$$

$$o_t = \sigma\left(w_o * (V_o h_{t-1}) + b_o\right) \tag{3}$$

$$c_t = tanh\left(w_C * (V_c h_{t-1}) + b_c\right) \tag{4}$$

where w and V refers to the weight in each element, h_{t-1} is the hidden state at the time step t-1, the cell block that controls the update value is denoted by c, and b refers to the bias in the current state.

5) Global max pooling layer

The global max pooling layer flattens the 3D tensor feature map into 2D tensor feature map before feeding it into the subsequent layer.

6) Dense layer

The dense layer is also known as the fully connected layer. The dense layer compiles the data that was extracted from the previous layers to form a final output. The sigmoid function is applied to compute the probability of the classes, i.e., fake news and real news.

7) Enhancements

An over complicated network model is likely to suffer from the overfitting problems. Two regularization techniques are leveraged to prevent the overfitting problem, which are dropout layer and L2 regularization. The purpose of applying these techniques is to reduce the complexity of the network.

In this work, a dropout layer with probability of 0.3 is inserted. It means that 30% of neurons in the hidden layer will be dropped out in every update. The L2 regularization also reduces the weight size in every update. Thus, when the

Authorized licensed use limited to: Indian Institute of Information Technology Lucknow. Downloaded on October 23,2024 at 13:26:27 UTC from IEEE Xplore. Restrictions apply.

number of hidden neurons becomes smaller, the complexity of the network is also reduced.

Apart from that, the Adam optimizer is adopted to optimize the gradient descent process. Adam optimizer adaptively tunes the learning rate for each weight by taking into consideration the momentum to avoid getting stuck in the local minima. The Adam optimizer is effective in accelerating the gradient descent process.

In every training epoch, a loss function is used to calculate the loss. Since the fake news detection is a binary classification problem, i.e., to categorize the news articles into fake or real, the binary cross entropy loss function is used, as below:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \tag{5}$$

where y refers to the class label, $p(y)$ denotes the predicted probability of i-th sample.

IV. Dataset

Four publicly available datasets are used in the performance evaluation of the proposed CNN-LSTM model.

a) Fake or real news dataset

The news articles in the fake or real news dataset were collected from Kaggle website, New York Times, WSJ, Bloomberg, NPR and The Guardian. The total article in this dataset is 6335 in which 3171 articles are real news and 3164 articles are fake news.

b) ISOT dataset

The ISOT dataset [13,14] contains 21417 true news articles collected from Reuter.com and 23481 fake news articles collected from the unreliable websites. The true news articles are categorized into two topics whereas the fake news articles are classified into six topics. These articles were collected from 2016 to 2017.

c) Kaggle – Getting real about fake news dataset

There are 20015 news articles in the Kaggle – Getting real about fake news dataset [15], out of which 11941 are fake news and 8074 are real news. The real and fake news articles were crawled from the reliable and unreliable websites respectively.

d) Combined news dataset

The dataset was aggregated from multiple Kaggle fake news dataset. The dataset contains 74012 news articles with 36969 real news articles and 37043 fake news articles.

V. Results and Analysis

In the experiments, the datasets are separated into 80% training data, 10% validation data and 10% testing data. The evaluation metric used is F1-score, defined as below:

$$F1score = \frac{2 * precision * recall}{precision + recall} \tag{6}$$

Table I presents the comparison results of the proposed CNN-LSTM model and some state-of-the-art methods. In the table, the datasets are represented in letters where D1: Fake or Real News Dataset, D2: ISOT Dataset, D3: Kaggle – Getting Real about Fake News Dataset and D4: Combined News Dataset. The experimental results demonstrate that the proposed CNN-LSTM model outshines the other methods in comparison on all 4 datasets.

The proposed CNN-LSTM model comprises CNN layers and LSTM layer. Specifically, the convolutional layers and max pooling layers extract useful sequence feature maps to be learned in the following layers. The LSTM layer learns the long-term dependency of the sequence feature. The dense layers interpret the relations between the features and class label. The final layer computes the probability distributions of the real or fake news class. The proposed enhancements are also effective in reducing the network complexity and alleviating the overfitting issues.

Fig. 5 shows the training and validation loss over epochs of the CNN-LSTM model. Noticeably, the training and validation loss of the CNN-LSTM model decrease steadily in each epoch. Hence, demonstrating that the CNN-LSTM model is able to learn well, and the regularization techniques are effective in mitigating the overfitting problem.

TABLE I. The Comparison Results of CNN-LSTM and State-of-the-Art Methods.

Method	**F1-score (%)**			
	D1	***D2***	***D3***	***D4***
CNN	91	98	94	96
LSTM	81	96	92	96
CLSTM	91	99	95	95
Bi-LSTM	81	98	89	96
BERT [16]	83	95	81	72
Modified CLSTM	**92**	**99**	**95**	**96**

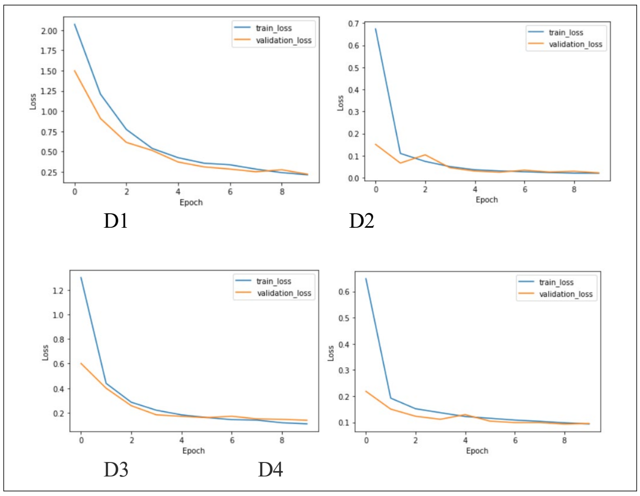

Fig. 5. The train and validation loss at every epoch of the CNN-LSTM model.

VI. Conclusion

Technology has undoubtedly made the communication faster and easier. This has brought side effects where the fake

Authorized licensed use limited to: Indian Institute of Information Technology Lucknow. Downloaded on October 23,2024 at 13:26:27 UTC from IEEE Xplore. Restrictions apply.

news also spread more rapidly. Therefore, fake news detection is essential to help people recognizing the fake news. In this paper, a CNN-LSTM model is proposed for fake news detection. The CNN layers play the role of extracting high-level representative features while the LSTM layer encodes the long-term dependencies of the features. The regularization techniques are effective in mitigating the overfitting problem. The empirical results exhibit that the proposed CNN-LSTM model outperforms the methods in comparison on four fake news datasets.

REFERENCES

[1] Lee, C. P., Lim, K. M., & Woon, W. L. (2010, July). Statistical and entropy based multi purpose human motion analysis. In *2010 2nd International Conference on Signal Processing Systems* (Vol. 1, pp. V1-734). IEEE.

[2] Mogan, J. N., Lee, C. P., Lim, K. M., & Tan, A. W. (2017, November). Gait recognition using binarized statistical image features and histograms of oriented gradients. In *2017 International Conference on Robotics, Automation and Sciences (ICORAS)* (pp. 1-6). IEEE.

[3] Yu, T. J., Lee, C. P., Lim, K. M., & Razak, S. F. A. (2019). AI-Based Targeted Advertising System. *Indonesian Journal of Electrical Engineering and Computer Science*, *13*(2), 787-793.

[4] Lee, C. P., Woon, W. L., & Lim, K. M. (2010). Statistical and entropy based human motion analysis. *KSII Transactions on Internet and Information Systems (TIIS)*, *4*(6), 1194-1208.

[5] Mogan, J. N., Lee, C. P., & Tan, A. W. (2017, May). Gait recognition using temporal gradient patterns. In *2017 5th International Conference on Information and Communication Technology (ICoIC7)* (pp. 1-4). IEEE.

[6] Tan, Y. S., Lim, K. M., Tee, C., Lee, C. P., & Low, C. Y. (2020). Convolutional neural network with spatial pyramid pooling for hand gesture recognition. *Neural Computing and Applications*, 1-13.

[7] Ahmed, H., Traore, I., & Saad, S. (2017). Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. *First International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, *10618*, 169–181. https://doi.org/10.1007/978-3-319-69155-8

[8] Khan, J. Y., Khondaker, M. T. I., Iqbal, A., & Afroz, S. (2019). *A Benchmark Study on Machine Learning Methods for Fake News Detection*. 1–14. http://arxiv.org/abs/1905.04749

[9] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[10] Bahad, P., Saxena, P., & Kamal, R. (2019). Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. *Procedia Computer Science*, *165*(2019), 74–82. https://doi.org/10.1016/j.procs.2020.01.072.

[11] Kaur, S., Kumar, P., & Kumaraguru, P. (2020). Automating fake news detection system using multi-level voting model. *Soft Computing*, *24*(12), 9049–9069. https://doi.org/10.1007/s00500-019-04436-y

[12] Thakur, A., Shinde, S., Patil, T., Gaud, B., & Babanne, V. (2020). MYTHYA: Fake News Detector, Real Time News Extractor and Classifier. *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, *Icoei*, 982–987. https://doi.org/10.1109/ICOEI48184.2020.9142971

[13] Ahmed H, Traore I, Saad S. (2017) "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127- 138).

[14] Ahmed H, Traore I, Saad S. "Detecting opinion spams and fake news using text classification", Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018..

[15] Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., & Yu, P. S. (2018). *TI-CNN: Convolutional Neural Networks for Fake News Detection*. http://arxiv.org/abs/1806.00749

[16] Jwa, H., Oh, D., Park, K., Kang, J. M., & Lim, H. (2019). exBAKE: Automatic fake news detection model based on Bidirectional Encoder Representations from Transformers (BERT). *Applied Sciences (Switzerland)*, *9*(19), 1–9. https://doi.org/10.3390/app9194062

Authorized licensed use limited to: Indian Institute of Information Technology Lucknow. Downloaded on October 23,2024 at 13:26:27 UTC from IEEE Xplore. Restrictions apply.