COMP5331 Project Report
Group 14
Type: Implementation

Exploring Current Models on Visual Information for Fake News Detection

CHENG, Tat Fan GUPTA, Ayush LIU, Songyu
LUNG, Kun Hung YAN, Chiu Wai

November 22, 2020

Abstract

Fake News Detection is an important problem to solve in today's modern world. In this report we reviewed MVNN (Multi-domain Visual Neural Network), a novel CNN-RNN mixed model for fake new image detection. Specifically, we had implemented several versions of model of MVNN and baseline models to test, evaluate and compare their performance. We end up with a similar result to the original paper, which outperforms all of our baseline models tested. Furthermore, we also conducted an ablation study to investigate the strength and weakness of the MVNN model, and thus propose possible improvements. At the end of the report, we have the Appendix section containing the general information of our project, including project type, group number, group members and more.

1 Introduction

1.1 Background

In the era of digitalization, the overwhelming growth of social media and news websites has made the news much more accessible, interactive and vivid than before. On the contrary, the role of traditional ways to report news through television, radio and newspapers has become more and more insignificant. Further, the social networking platforms (like Twitter, Weibo and Facebook) serve as popular microblogs for people to share or even create news to their friends and followers. However, the social media and platforms are also the breeding ground for fake news and rumor. It is pointed out that fake news can defame celebrities, impose bias and mislead content in people's mind and even affect the results of political elections [1].

Notably, since the emotional impact of visual information is more than that of textual information, misrepresented or tampered images are often used in fake news to attract readers' attention. In certain cases, misleading real images are also used in fake news. Hence, it becomes increasingly important to utilize image features in the news content to improve fake news detection.

1.2 Previous work

Fake news is actually a long-existing problem in social media and researchers have designed various techniques to perform the detection from different perspectives. While the previous researches focused mainly on discriminating textual features for fake news detection [2], emerging studies showed the usefulness of images for rumor detection through supervised learning [3, 1, 4].

For example, a pivotal study, titled 'Multimodal fusion with recurrent neural networks for rumor detection on microblogs' was published in 2017 Association for Computing Machinery (ACM). It attempted to detect fake news through fusing the information from text, social context and images [3]. The authors proposed a novel RNN with an attention mechanism (att-RNN) to fuse multimedia features in news. The att-RNN model showed improvements in fake news detection accuracy using the Weibo and Twitter datasets. In particular, they highlighted the vital role of visual features in improving the detection accuracy, compared to other features like text and social context.

Regarding rumor detection based on image in news articles, a recent study published in 2019 IEEE International Conference on Data Mining (ICDM), entitled 'Exploiting Multi-domain Visual Information for Fake News Detection' [1], innovatively exploited visual features from different domains for rumor detection. In this paper, the authors utilize the Weibo dataset containing fake and real images crawled from official rumor exposing Weibo platform. They proposed a framework called Multi-domain Visual Neural Network (MVNN), which involves the Convolutional Neural Network (CNN) and the Recurrent Neural Network (RNN). Their approach utilizes the pixel domain and frequency domain information from the images, to classify them as real or fake, and outperforms existing methods with at least 9.2% accuracy.

1.3 Research problem and objectives

One major problem for rumor detection is the scarcity of benchmark dataset that label whether the news is fake or real, especially for the news images. This, to some extent, limits the development of rumor detection models based on visual features. Given the existing models such as MVNN that are capable of exploiting visual features for detecting fake news and the emerging news image datasets, this project aims to compare different models for fake news detection based on news images. Specifically, we intend to achieve the following objectives:

1. Use the MVNN model to extract visual features from different domains (e.g., pixel and frequency domains) and merge them using attention for

fake news detection.

2. Compare the performance of the MVNN model with other baseline models on real-world dataset.
3. Conduct an ablation study on the performance of the MVNN model to investigate which part contributes more to its final results
4. Attempt to improve the MVNN model by experimenting different fine architecture and hyperparameters where the paper did not specify in detail.

2 Implementation

2.1 Data set

A good data set for fake news detection via deep learning for image classification, should consist of both rumor and non rumor images, and be large enough for the task of Deep learning.

There were 2 major data sets when [1] was published: the Twitter dataset [5] and the Weibo dataset. Since the Twitter data set consisted only of about 350 images, it was too small for the task of deep learning. We still tried applying standard Deep learning models and they were unable to perform well on it. The Weibo data set instead, consisted of almost 13272 images, 5318 of non rumor (real) and 7954 rumor (fake) images. Thus, Weibo data set is considered large enough for Deep learning and it has been used in this project.

There are different pre-processing techniques which can be applied before training. For example, [3] crawled the Weibo dataset and performed data cleansing. As an another example, [1] only used one image for one post. Since News can be divided into different topics or categories. If most of training data are about the same topic, the model will undergo over-fitting. To prevent overfitting, some of the previous researchers clustered the images first. Such clustering pre-processing techniques can be adopted to improve the final performance. However, we obtained performance comparable with the [1], by implementing other pre-processing techniques such as data augmentation.

2.2 Methodology

Tampered and misleading images typically exhibit variations compared with ordinary images, in the aspects of both pixel domain and frequency domain [1]. At the physical layer, a fake-news image might appear in low quality due to the multiple upload/download of the composite image, resulting as heavier re-compression artifacts when compared to the real-news image. This characteristic can be extracted easily by inputting the image frequency information into a CNN-based network. At the semantic level, due to the pursuit of visual impacts and emotional provocations, fake-news images trend to exhibit differences to real-news images throughout visual factors of image levels, which can be reflected by pixel information. Building a multi-branch CNN-RNN network could help extract features in different semantic levels for further classification. As a result, we propose to adopt the Multi-domain Visual Neural Network (MVNN) introduced in previous sections. To process input from different domains, MVNN consists of 3 components, a sub-network in frequency domain, a sub-network in pixel domain, and a sub-network that fuses the result of the two previous sub-networks.

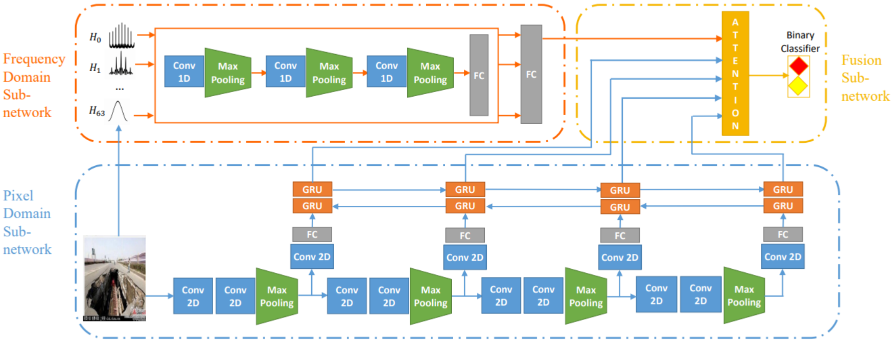

Figure 1: The MVNN model architecture proposed in [1].

2.2.1 Baseline models

Since the input is an image while the output is a binary value of being fake, our task can be regarded as an image classification problem. Therefore, we select some of the models that perform well in the field of image classification to be our baselines. Among the models, VGG-16 and VGG-19 [6] are popular models with their ability to extract the image features from pixels. Therefore, we adopt VGG-16 and VGG-19 pretrained on ImageNet as our baseline models, and fine-tune them with our training dataset.

2.2.2 Simple Sequential Model of CNN

The Simple Sequential Model of Convolutional Neural Network (CNN) in this project utilizes the architecture of 2D CNN models with ReLU activation function, combined with BatchNormalization, MaxPooling2D and Dropout layers for each set of CNN applied to the data. After these layers, a fully connected layer is used to map the feature vectors into output logits, in order to perform the binary classification task. These CNN takes images with dimension width, height and RGB channels as input data.

This Simple Sequential Model is experimented with different numbers of CNN layers, from 1 to 4. Before providing the data into this model, we perform data augmentation to increase the number of training images and ensure that the model learns the important features that can be observed in many scenarios instead of only learning non-important features. During the training of the model, we also adopted learning rate decay when there is no significant improvement in the validation accuracy. This helped the model to automatically accommodate the learning rate and improve its performance during the training process.

2.2.3 Frequency Domain Sub-network

The Frequency Domain Sub-network is a simple 1D Convolutional Neural Network (CNN) structurally. Instead of processing pixel information or high-level image features in typical CNNs, this component attempts to discover any abnormal modification in the input images.

JPEG images produce traces during re-compression [7], which can be detected by reproducing the 8x8 DCT (Discrete Cosine Transform) blocks. Specifically, if a JPEG image is downloaded, tampered and saved back as a new JPEG file, the pattern shown in the 8x8 DCT blocks exhibits discrepancy comparing to raw images. Such difference can help in distinguishing raw image and tampered image. The Frequency Domain Sub-network is implemented specifically for this behavior. Note that retrieval of the frequency domain information is also required during JPEG compression, and we implement such procedure as pre-processing of our input data.

As the input dimension needs to be consistent, we reshape the input image into dimension $(3, 128, 128)$, in YCbCr color space. According to [7], Cb and Cr channels do not provide as much information as the Y channel does, so we only keep the Y channel of the image. Then the image is split into 8x8 patches, with each patch processed with 1D Block DCT, producing 64 weights on the cosine pattern with different vertical and horizontal frequencies. Since the Y channel of our input image has dimension $(128, 128)$, which can be divided right into 256 8x8 patches, the expected dimension after the pre-processing will be $(64, 256)$. Then the input will be fed into 3 consecutive 1D-convolution layers, each followed by a max pooling. Finally, the features go through a fully connected layer, yielding the logits indicating the predicted classification.

2.2.4 Pixel Domain Sub-network

The Pixel Domain Sub-network is a multi-branch CNN-RNN model which extracts the features of each semantic level of the image and feed them into an RNN model.

Since the sub-network handles images in the pixel domain, the input are expected to be matrices with dimension $(3, H, W)$, where H and W are the height and width of the images respectively, while the 3 channels are the RGB color intensities of the corresponding pixel.

It is recommended to extract both high-level and low-level image features to fully capture the semantic characteristic of fake news images [1]. Therefore the CNN part of the sub-network can be divided into 4 branches, with each block containing a 3×3 and 1×1 convolution layer and a max pooling layer. Apart from stacking together, the output of each branch is also processed through a 1×1 convolutional layer followed by a fully connected layer outputting one dimensional vectors.

After all the processing above, the 4 sets of image features $v = \{v_1, v_2, v_3, v_4\}$ will be concatenated into a sequence of length 4 and directed to the input of a standard bidirectional GRU network, while the GRU unit composes of the reset gate, update gate, and the hidden gate. Since Bi-GRU contains a forward GRU GRU_f along with a backward GRU GRU_b, the final semantic feature is formulated by $L = \{l_1, l_2, l_3, l_4\}$, where $l_t = [GRU_f(v_t), GRU_b(v_t)]$ for $t \in [1, 4]$.

$$\overrightarrow{l_t} = GRU_f(v_t); \overleftarrow{l_t} = GRU_b(v_t) \tag{1}$$

$\overrightarrow{l_t}; \overleftarrow{l_t}$ are both hidden states.

2.2.5 Fusion Sub-Network

The Fusion Sub-Network is a classifier model that fuses the output feature vectors from the previous two sub-network by an attention mechanism, then it classifies if the input image is a real-news or fake-news image.

To leverage the output vector l_0 from Frequency Domain Sub-Network and $\{l_1, l_2, l_3, l_4\}$ from Frequency Domain Sub-Network, denoted as levels, MVNN develops an automatic weight for each of the feature vector through the attention mechanism. Each of the feature vectors will be scored base on the information within, and a high level feature vector u can be computed balancing the five vectors. Similar to the additive attention proposed in [8], the process can be formulated by the following:

$$\mathcal{F}(l_i) = v^T \tanh\left(W_f l_i + b_f\right), i \in [0, 4]$$

$$\alpha_i = \frac{\exp(\mathcal{F}(l_i))}{\sum_j \exp(\mathcal{F}(l_j))}$$

$$u = \sum_i \alpha_i l_i$$

where v^T is a weight vector learned, $\mathcal{F}(\cdot)$ is the score function evaluating the significance of each feature vector based on their context, α_i is the final weight assigned to each vector, and u is the weighted output of the 5 input vectors.

Finally, u is passed as a feature vector to a fully connected layer with softmax activation, predicting the input image to two resulting space: real-news image and fake-news image, and gain the probability distribution: $p = softmax(W_c u + b_c)$ where W_c denotes the weight matrix and b_c is the bias term.

3 Evaluation and Results

Images from the Weibo dataset (both real and fake) are split into 80% for training and 20% for validation. We utilized the following metrics for our testing: classification report (containing precision, recall, f1-score and support for both

Table 1: Results of Baseline models of VGG-16 and VGG-19

Type of model	Precision	Recall	f1-score	Accuracy	Cohen Kappa
VGG-16	69%, 73%	57%,82%	62%,78%	72%	0.40
VGG-19	59%, 79%	76%,64%	66%,71%	69%	0.38

the classes), confusion matrix (providing the details of precision and recall of our model for each of the classes), accuracy, balanced accuracy and the Cohen Kappa score (providing a mathematical comparison of our model with any random model, to show how well our model outperforms a random model) where the classes are class 0 (non-rumor/real images) and class 1 (rumor/fake images).

Unless otherwise specified, most of the training are conducted under the Google Colab enviornment, with a GPU instance of Nvidia Tesla K80 with 12GB RAM. Adam optimizer is used with a learning rate of 0.0001. We set the batch size to be 32. Though it is not illustrated in 1, we use ReLU activation and batch normalization between every fully connected layer and convolutional layer in each of our MVNN sub-networks.

3.1 Baseline models

We perform a train-test split on the Weibo dataset which 80% data is used for training while 20% data is used for testing. The Table 1 shows the performance of our selected baseline models in the validation dataset. The two values in precision, recall and f1-score corresponds to that of the prediction on class 0 (real-news image) and class 1 (fake-news image).

We found that both VGG-16 and VGG-19 gave almost similar results of overall accuracy of 72% and 69%, average f1 scores of 70% and 68.5%, and cohen kappa scores of 0.40 and 0.38. Then, we used them as our baseline models for this project. Further improvement in these baseline models can probably be done by performing more hyper-parameter tuning.

3.2 Simple Sequential Model of CNN

The results that we obtained using our sequential model of different CNN layers are shown in Table 2. The two values in precision, recall and f1-score corresponds to that of the prediction on class 0 (real-news image) and class 1 (fake-news image). Across various models with different number of CNN layers, we found the scores as shown in the Table 2.

We found that 2 CNN layers gave the best results with respect to considering 1,2,3,4 CNN layers. This serves as further motivation in MVNN model, to use 2 CNN layers at a time for each RNN unit. Further improvement in this Sequential model can probably be done by performing more hyper-parameter tuning.

Table 2: Results of our developed Sequential model of CNN

Num of CNN layers	Precision	Recall	f1-score	Accuracy	Cohen Kappa
1	61%, 72%	56%,76%	59%,74%	66%	0.332
2	68%, 77%	67%,78%	68%,77%	73%	0.451
3	58%, 78%	71%,66%	64%,72%	70%	0.357
4	58%, 78%	71%,66%	64%,72%	68%	0.356

Table 3: Results of Full MVNN model

Precision	Recall	f1-score	Accuracy	Cohen Kappa
76%, 87%	80%,84%	78%,85%	82.2%	0.631

3.3 MVNN

We now conduct the full MVNN model (consisting of the frequency domain, pixel domain and the attention mechanism). Table 3 has the validation results.

Our implementation of the MVNN model shows a promising accuracy of 82.2% and a Cohen Kappa score of up to 0.631, showing that the MVNN model is useful for fake news detection, comparing to our baseline models. Since the MVNN model consists of three parts: the frequency, the pixel and the attention mechanism, the relative importance of these components can offer insight into where the predictive power comes from and how may we simplify the model. The 5 child models' performances are reported in the following subsections.

3.4 MVNN w/o pixel domain

In this child model, we remove the pixel sub-network as well as the attention mechanism so that only the frequency sub-network with a binary classifier is adopted to perform fake news detection. The validation results are in Table 4.

Our implementation of MVNN without pixel domain show a similar accuracy with the result on the paper (73.7%). The apparent decay in accuracy is expected since the frequency domain of an image only reveals special property under restricted conditions.

Table 4: Results of MVNN w/o pixel domain

Precision	Recall	f1-score	Accuracy	Cohen Kappa
65%, 76%	61%,79%	63%,77%	72%	0.404

Table 5: Results of MVNN w/o frequency domain

Precision	Recall	f1-score	Accuracy	Cohen Kappa
79%, 83%	73%,87%	76%,85%	81.2%	0.603

Theoretically, to exhibit a detectable property due to recompression, the input images need to fulfill the following conditions: 1) JPEG compression scheme is used; 2) The image is compressed more than twice in total. However in the Weibo Dataset, not all images are compressed in JPEG format, also, some untampered but misused images are also categorized as fake news. Under this assumption and the empirical result, we can assert that the frequency domain sub-network helps in a certain degree in detecting modified JPEG image, but it fails to perform a complete study on the nature of the image, due to the lack of pixel information and high-level features. This result agrees with the decision to combine the two sub-networks using an attention mechanism rather than fixed weights, as the frequency analysis is almost ineffective in some particular images (e.g. PNG images). In such cases, the model needs to learn mechanism to relate to more significant features such as the pixel information.

3.5 MVNN w/o frequency domain

In contrast to the previous subsection, we now examine the performance of another child model with the frequency sub-network removed, but retaining the pixel domain and the attention mechanism. The validation results are shown in Table 5.

The MVNN without frequency sub-network demonstrates an accuracy over 80%, which is almost as good as the full MVNN model. Note that such accuracy already exceeds prior methods as compared in [3]. This may suggest that the predictive power for fake news detection based on images comes from the pixel sub-network in a great extent. Also, it reflects the effectiveness of the 4-branch CNN-RNN structure.

3.6 MVNN w/o attention

By concatenating the feature vectors generated by the frequency and pixel sub-network, we replace the original attention mechanism with a simple fully-connected layer to perform a simple classifier. The outcomes are shown in Table 6.

It is obvious that the result is very close to that by the full MVNN model, suggesting a marginal usefulness of the attention mechanism in this case. This is also expected due to the performance decay in the frequency domain sub-network. As the CNN-RNN model outperforms the frequency domain sub-network, the learnt weighting between two output will show usual bias to the

Table 6: Results of MVNN wøattention

Precision	Recall	f1-score	Accuracy	Cohen Kappa
75%, 86%	79%,83%	77%,85%	81.8%	0.620

Table 7: MVNN w/o GRU

Precision	Recall	f1-score	Accuracy	Cohen Kappa
76%, 85%	78%,83%	77%,84%	80.9%	0.606

pixel domain sub-network, which degrades the advantages of the attention mechanism.

3.7 MVNN w/o GRU

It is of interest to know how useful the bidirectional GRU components (in the pixel sub-network) would be. Thus in this child model, the outputs of the pixel sub-network would be v_1, v_2, v_3 and v_4, instead of l_1, l_2, l_3, l_4 originally from the GRU components. Results in Table 7.

The accuracy is just 1.3% lower than that of the full MVNN model (82.2%), the considerably smaller Cohen Kappa score is found ($\tilde{0}.61$), compared to the score from the full model ($\tilde{0}.65$). The result may suggest that the GRU units could improve the fake news detection since the bidirectional GRUs can update their weights by considering the lower-level and higher-level image features from the CNN model. To some extent, the GRU components might further mine a more general pattern of the fake news images.

3.8 MVNN w/o branches

In our reference paper, the authors also test the performance of the MVNN model without the ”branches” (and theus also the bidirectional GRUs) in the pixel sub-network. Results of such a child model are shown in Table 8.

Since the accuracy is similar to that of the MVNN without GRU model, we believe that the ”branches” are useful when it is coupled with the GRU

Table 8: MVNN w/o branches Results

Precision	Recall	f1-score	Accuracy	Cohen Kappa
76%, 86%	79%,83%	78%,84%	81.6%	0.620

components.

4 Conclusion

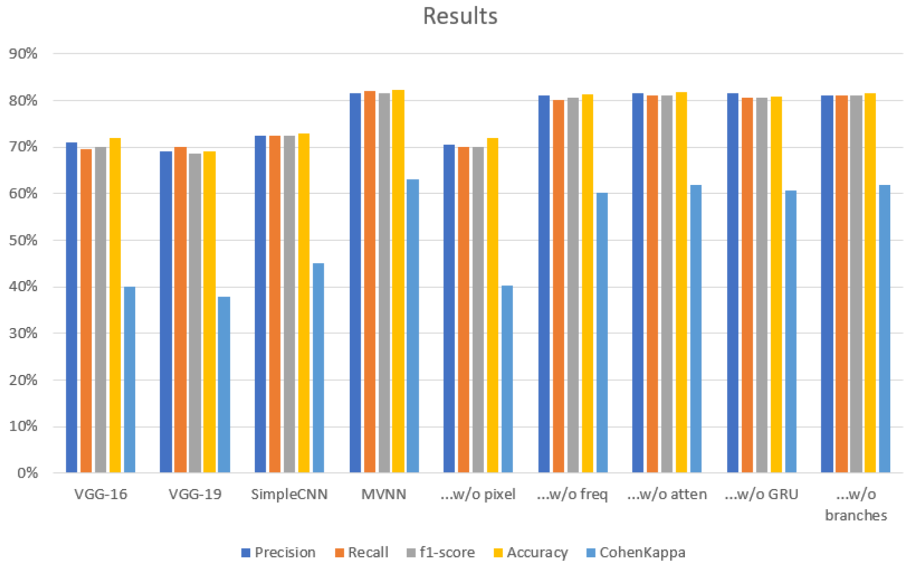

Figure 2: The overview result of Models

In this paper, we aimed to utilize the information from image part of news for fake new detection. We reviewed and re-implemented the Multi-domain Visual Neural Network (MVNN) model, its 5 child models and several baseline models related to fake news detection. In order to compare the performance of MVNN model with other models used in this area, such as in [3], our group gathered some real world data sets and conducted the tests. After summarizing the results of the models from the last section, we obtained the graph in Figure 2.

From Figure 2, MVNN can achieve a high accuracy (over 80%) for fake news detection, which performs overwhelmingly better than the other baseline models such as VGG-16 and VGG-19. In addition, we conclude that the pixel domain of an image shows a stronger correlation to the MVNN model.

However, during the implementation we found that the reason of using GRU is lack of comparison with stronger units and the choice of data set is limited. Therefore, we suggest a number of possible changes that can be attempted to improve the current MVNN:

1. Stronger units: The use of GRU can be tested and compared with more sophisticated architecture like LSTM or even Transformer-based architecture [9]

2. Dataset: The text part can be included since fake news detection depends a lot in the actual text describing the image

3. More preprocessing: Drop the assumption on JPEG image (which is found to be not the most helpful) so that technique like data augmentation can be applied

4. Transfer learning: Existing convolution parameters can be adopted instead of training from scratch with randomly initialized weights

References

[1] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 518–527, 2019.

[2] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*, pages 1103–1108. IEEE, 2013.

[3] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816. ACM, 2017.

[4] Sarthak Jindal, Raghav Sood, R. Singh, Mayank Vatsa, and Tanmoy Chakraborty. Newsbag: A benchmark multimodal dataset for fake news detection. In *SafeAI@AAAI*, 2020.

[5] Detection and visualization of misleading content on Twitter. Boididou, christina and papadopoulos, symeon and zampoglou, markos and apostolidis, lazaros and papadopoulou, olga and kompatsiaris, yiannis. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[7] Yi-Lei Chen and Chiou-Ting Hsu. Detecting recompression of jpeg images via periodicity analysis of compression artifacts for tampering detection. *IEEE Transactions on Information Forensics and Security*, 6:396–406, 06 2011.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.