Multi-modal Emergent Fake News Detection via Meta Neural Process Networks

Yaqing Wang[†], Fenglong Ma[◇], Haoyu Wang[†], Kishlay Jha[*], Jing Gao[†].
[†]**SUNY Buffalo,** [◇]**Pennsylvania State University,** [*]**University of Virginia**
{yaqingwa, hwang79, jing}@buffalo.edu,
fenglong@psu.edu,
kishlay@email.virginia.edu

Abstract

Fake news travels at unprecedented speeds, reaches global audiences and puts users and communities at great risk via social media platforms, especially for new emerging events. Due to the fact that it is hard to obtain sufficient verified labeled data, significant challenges are posed for existing detection approaches to stop the spread of fake news in the early stage. In order to address this challenge, we propose an end-to-end fake news detection framework named MetaFEND, which is able to learn quickly with a few verified posts using the proposed meta neural process networks for effective detection on newly arrived events. In particular, a label embedding module and a hard attention mechanism are proposed to enhance the learning of context information, which further helps meta neural process networks to improve the performance of detecting emergent fake news. Extensive experiments are conducted on multimedia datasets collected from Twitter and Weibo. The experimental results show our proposed MetaFEND model[1] can detect fake news on never-seen events effectively and outperform the state-of-the-art methods.

Introduction

The recent proliferation of social media has significantly changed the way in which people acquire information. According to the 2018 Pew Research Center survey, about two-thirds of American adults (68%) get news on social media at least occasionally. The fake news on social media usually take advantage of multimedia content which contain misrepresented or even forged images, to mislead the readers and get rapid dissemination. The dissemination of fake news may cause large-scale negative effects, and sometimes can affect or even manipulate important public events. Recent years have witnessed a number of high-impact fake news spread regarding terrorist plots and attacks, presidential election and various natural disasters. Therefore, there is an urgent need for the development of automatic detection algorithms, which can detect fake news as early as possible to stop the spread of fake news on those emergent events and mitigate its serious negative effects.

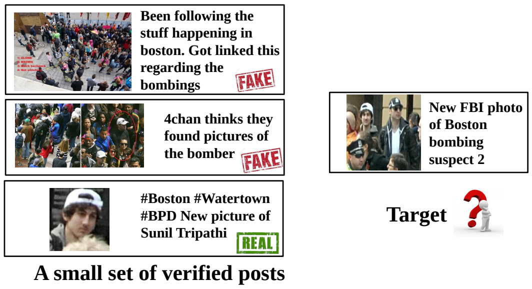

Been following the stuff happening in boston. Got linked this regarding the bombings
FAKE
4chan thinks they found pictures of the bomber
FAKE
#Boston #Watertown #BPD New picture of Sunil Tripathi
REAL
A small set of verified posts
New FBI photo of Boston bombing suspect 2
Target

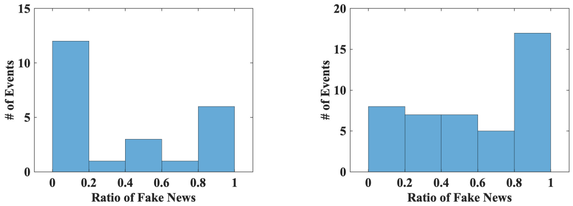
Figure 1: Fake news examples on an emergent event from Twitter.

Task Challenges. Thus far, various fake news detection approaches, including both traditional learning (Conroy, Rubin, and Chen 2015; Tacchini et al. 2017) and deep learning based models (Ruchansky, Seo, and Liu 2017; Ma et al. 2016; Ma, Gao, and Wong 2018, 2019; Wang et al. 2018; Popat et al. 2018) have been exploited to identify fake news. Although deep learning models have achieved an improvement in the performance of fake news detection, the algorithms achieved on old news data may exaggerate performance on never-seen events due to dynamic nature of news (Wang et al. 2018, 2019). Considering the great need to detect fake news emerged on newly arrived events, it poses significant challenges for fake news detection in the real setting. When an emergent event happens, we may only have a handful of related verified posts at hand (An example is shown in the Fig. 1). Thus, how to leverage *a small set of verified posts* to make the model learn quickly to detect fake news on the never-seen events is a crucial challenge.

Limitations of Current Techniques. The few-shot learning, which aims to leverage a small set of data instances for quick learning, is a possible solution. One promising research line of few-shot learning is **meta-learning** (Finn, Abbeel, and Levine 2017; Li et al. 2017), whose basic idea is to leverage the knowledge from previous tasks to facilitate the learning on new task. However, most of the existing meta-learning methods rely on an important assumption: the tasks are sampled from a similar distribution. This assumption usually does not hold in the fake news detection problem as *the writing style, content, vocabularies and even class distributions of news on different events usually tends to differ.* As can be observed from Figure 2a and Figure 2b, the class distributions of events are significantly different and imbalanced. The **event heterogeneity** cannot be handled by globally sharing knowledge among events and further exacerbates the unstable

[1] Code is available at http://tiny.cc/5q2lpz.

problem of meta-learning algorithms.



(a) Twitter

(b) Weibo

Figure 2: The number of events with respect to different percentages of fake news.

Another research line of few-shot learning is **neural processes** (Garnelo et al. 2018a,b; Kim et al. 2019), which conduct inference using a small set of data instances as conditioning and can generalize well to heterogeneous tasks. However, they suffer from the limitations like *ignoring categorical characteristics of label information and underfitting*. These two research line of models are complementary to each other. In order to integrate them together, we need to handle the incompatible operations on the given small set of data instances.

Our Approach. To address the aforementioned challenges, in this paper, we propose a novel meta neural process network (namely MetaFEND) for emergent fake news detection. MetaFEND is built upon gradient-based meta-learning algorithms and neural process models, which makes it able to learn rapidly with a few labeled examples as conditioning information. Besides, the proposed model can handle heterogeneous events via sharing global knowledge and using event-specific data instances as conditioning simultaneously. Furthermore, we incorporate two novel components - *label embedding* and *hard attention* - to tackle the inherited limitations of neural process models. Experimental results on two large real-world datasets show that the proposed model effectively detect fake news on new events with a handful of posts and outperforms the state-of-the-art approaches.

Our Contributions. The main contributions of this paper can be summarized as follows:

- We recognize the challenges of fake news detection on emergent events and formulate the problem into a few-shot learning setting. Towards this end, we propose an effective meta neural process framework to detect fake news on never-seen events with a handful of data instances.
- The proposed MetaFEND method fuses the gradient-based meta-learning method and neural process models together via a simulated learning task design together to enjoy the benefits of two models without suffering their limitations.
- We empirically show that the proposed method MetaFEND can effectively identify fake news on various events and significantly outperform the state-of-the-art models on two real-world datasets.

Related Work

In this section, we briefly review the work related to the proposed model. We mainly focus on the following two topics: fake news detection and few-shot learning.

Fake News Detection

Many fake news detection algorithms try to distinguish news according to their features, which can be extracted from social context and news content. (1) *Social context features* represent the user engagements of news on social media (Shu et al. 2017) such as the number of followers, hash-tag (#), propagation patterns (Wu, Yang, and Zhu 2015) and retweets. However, social context features are very noisy, unstructured and labor intensive to collect. Especially, it cannot provide sufficient information for newly emerged events. (2) *Textual features* are statistical or semantic features extracted from text content of posts, which have been explored in many literatures of fake news detection (Shu et al. 2017; Gupta et al. 2014; Castillo, Mendoza, and Poblete 2011). Unfortunately, linguistic patterns are not yet well understood, since they are highly dependent on specific events and corresponding domain knowledge (Ruchansky, Seo, and Liu 2017). To overcome this limitation, approaches like (Ma et al. 2016; Ma, Gao, and Wong 2018, 2019; Popat et al. 2018; Lu and Li 2020) propose to use deep learning models to identify fake news and have shown the significant improvements. (3) *Visual features* have been shown to be an important indicator for fake news detection (Jin et al. 2017c; Shu et al. 2017). The basic features of attached images in the posts are explored in the work (Gupta, Zhao, and Han 2012; ping Tian et al. 2013; Jin et al. 2017c).

In this paper, we consider multiple types of features simultaneously when identifying fake news on social media. To tackle *multi-modal fake news detection*, in (Jin et al. 2017a), the authors propose a deep learning based fake news detection model, which extracts the multi-modal and social context features and fuses them by attention mechanism. To detect fake news on newly emergent events, Wang et al. (Wang et al. 2018) propose an event-adversarial neural network which can capture event-invariant features for fake news detection. However, such method discards the event specification which is informative for fake news detection.

Few-Shot Learning

Meta-learning has long been proposed as a form of learning that would allow systems to systematically build up and re-use knowledge across different but related tasks (Vilalta and Drissi 2002). Meta-Learning approaches can be broadly classified into three categories: gradient-based, model-based and based on metric-learning. Optimization methods aim to modify the gradient descent based learning procedure for new task quick adaptation. In the optimization-based methods, MAML (Finn, Abbeel, and Levine 2017) is to learn model initialization parameters that are used to rapidly learn novel tasks with a small set of labeled data. Following this direction, besides initialization parameters, Meta-SGD (Li et al. 2017) learns step sizes and updates directions automatically in the training procedure. As tasks usually are different in the real setting, to handle task heterogeneity, HSML (Yao et al. 2019) customizes the global shared initialization to each cluster using a hierarchical clustering structure. However, the hierarchical relationship does not exist in news events.

Neural process approaches that combine stochastic process and neural network can offer an efficient method to modelling a distribution over functions conditioned on a context set. Conditional Neural Process (CNP) (Garnelo et al. 2018a) uses a neural network to take input-output pairs of support set as conditioning for inference. Another research work Neural Process (NP) (Garnelo et al. 2018b) shares similar principles but uses the stochastic neural network to better capture the uncertainty of functions. However, these two works suffer from the underfitting problem because they aggregate the context set by average or sum ignoring their differences in importance. To overcome this limitation, Attentive Neural Process (ANP) (Kim et al. 2019) incorporates attention mechanism into Neural Process and further achieves better performance. However, they employ soft-attention mechanism which is less effective as the size of support set increases. Furthermore, ANP directly concatenates the label values with feature representation and discard the categorical characteristics of label information.

Remark. Technically, different from existing work, our proposed framework maintains the parameter flexibility via the gradient-based meta learning procedure and inherits generalization ability from neural processes. Moreover, we incorporate label embedding component to handle categorical characteristics of label information and utilize hard attention to extract most informative context information. Thus, our proposed model enjoys the benefits of two model families without suffering their own disadvantages.

Preliminary

MAML. We first give an overview of Model-Agnostic Meta-Learning method (Finn, Abbeel, and Levine 2017), a representative algorithm of gradient-based meta-learning approaches, and take few-shot fake news detection as an example. Given a base learner f with θ as parameters, the event specific parameters θ_e are learned to make accurate fake news detection for event e. The goal of meat-learning is to learn from previous events, which is a well-generalized meta-learner $M(\cdot)$ to facilitate the training of the base learner in a new event with a few examples. To accomplish such a goal, the meta-learning are split into two stages: meta-training and meta-testing.

During the meta-training stage, the baseline learner f_θ will be adapted to specific event e as f_{θ_e} with the help of meta-learner $M(\cdot)$ on the support set $\mathcal{D}_e^s$, i.e., $\theta_e = M(\theta, \mathcal{D}_e^s)$. Such an event specific learner f_{θ_e} is evaluated on the corresponding query set $\mathcal{D}_e^q$. The loss $\mathcal{L}(f_{\theta_e}, \mathcal{D}_e^q)$ on $\mathcal{D}_e^q$ is used to update the parameters of baseline learner θ and meta-learner $M(\cdot)$. During the meta-testing stage, the baseline learner f_θ will be adapted to the testing event e' using the same procedure with meta-training stage, i.e., $\theta_{e'} = M(\theta, \mathcal{D}_{e'}^s)$, and make predictions for the $\mathcal{D}_{e'}^q$.

MAML update parameter vector θ using one or more gradient descent updates on event e. For example, when using one gradient update:

$$\theta_e = M(f_\theta, \mathcal{D}_e^s) = \theta - \alpha \bigtriangledown_\theta \mathcal{L}(f_\theta, \mathcal{D}_e^s).$$

The model parameters are trained by optimizing for the performance of f_{θ_e} with respect to θ across events sampled from $p(\mathcal{E})$. More concretely, the meta-objective is as follows:

$$\min_\theta \sum_{e \sim \mathcal{E}} \mathcal{L}(f_{\theta_i}) = \sum_{e \sim \mathcal{E}} \mathcal{L}(f_{\theta - \alpha \bigtriangledown_\theta \mathcal{L}(f_\theta, \mathcal{D}_e^s)}, \mathcal{D}_e^q).$$

Limitations of MAML. The MAML can capture task uncertainty via one or several gradient updates. However, in fake news detection problem where events are heterogeneous, the event uncertainty is difficult to encode into parameters via one or several gradient steps.

Conditional Neural Process (CNP). The CNP includes four components: encoder, feature extractor, aggregator and decoder. The basic idea of conditional neural process is to make predictions with the help of support set $\mathcal{D}_e^s = \{x_{e,i}^s, y_{e,i}^s\}_{i=1}^N$ as context. The dependence of a CNP on the support set is parametrized by a neural network encoder, denoted as $g(\cdot)$. The encoder $g(\cdot)$ embeds each observation in the support set into feature vector, and the aggregator $q(\cdot)$ maps these feature vectors into an embedding of fixed dimension. In CNP, the aggregation procedure is a permutation-invariant operator like averaging or summation. The query data of interest $x_{e,k}^q \in \mathbf{X}_e^q$ is fed into feature extractor $h(\cdot)$ to get the feature vector. Then the decoder $f(\cdot)$ takes the concatenation of aggregated embedding and given target data $x_{e,i}^q \in \mathbf{X}_e^q$ as input and output the corresponding prediction as follows:

$$p(y_{e,i}^q | \mathbf{X}_e^s, \mathbf{Y}_e^s, x_{e,i}^q) = f(q(g(\mathbf{X}_e^s, \mathbf{Y}_e^s)) \oplus h(x_{e,i}^q)).$$

where $\oplus$ is concatenation operator.

Limitations of CNP. One widely recognized limitation of CNP is underfitting (Kim et al. 2019). For different context data points, their importance is usually different in the prediction. However, the aggregator of CNP treats all the support data equally and cannot achieve query-dependent context information. Moreover, it is difficult for a fixed parameter set to capture uncertainties of all the events.

Intuitions. The two research lines are complementary to each other. The parameter update mechanism in MAML can help alleviate unfitting issues of the neural process. The neural processes which leverage a context set, can help handle the heterogeneity challenge for MAML by retrieving information from context set instead of encoding all the information into parameter set. Such an intuition motivates us to propose Meta neural process framework.

Methodology

Problem Formulation

We model fake news detection on emergent news events as a few-shot learning problem. The zero-shot learning setting is not the focus of this paper and we leave it as future work. Let $\mathcal{E}$ denote a set of news events. On each news event $e \in \mathcal{E}$, we have two independent sets: support set $\{\mathbf{X}_e^s, \mathbf{Y}_e^s\}$ and query set $\{\mathbf{X}_e^q, \mathbf{Y}_e^q\}$. Our goal is to leverage the knowledge learned from past events and a few examples (i.e., support set $\{\mathbf{X}_e^s, \mathbf{Y}_e^s\}$) on event e to conduct effective fake news detection on the corresponding query set $\{\mathbf{X}_e^q, \mathbf{Y}_e^q\}$.

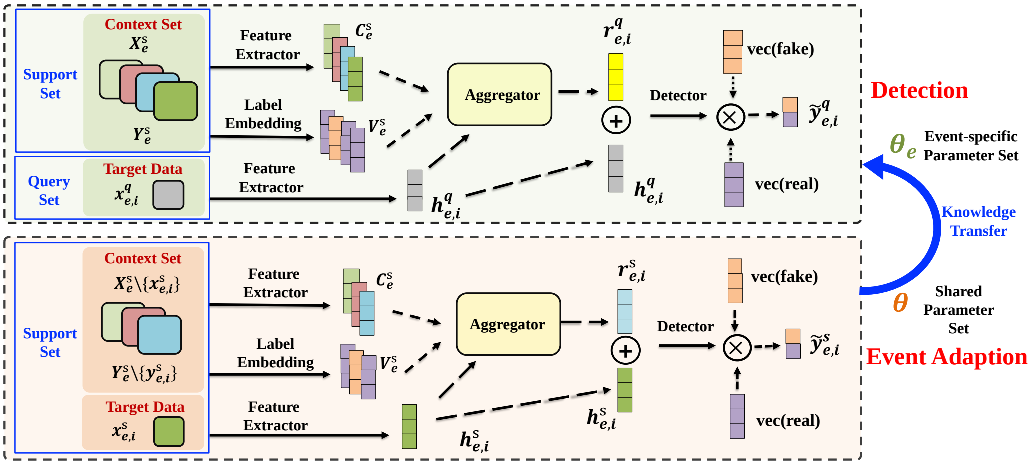


Figure 3: The proposed framework MetaFEND. The proposed framework has two stages: event adaption and detection. During the event adaption stage, the model parameter set θ is updated to event-specific parameter set θ_e. During the detection stage, the event-specific parameter set θ_e is used to detect fake news on event e. $\oplus$ denotes concatenation operation and $\otimes$ means element-wise product.

Meta-learning Neural Process Design

As shown in Figure 3, our proposed framework includes two stages: event adaptation and detection.

Event adaption. We take the i-th support data $\{x_{e,i}^s, y_{e,i}^s\}$ as an example, in the event adaption stage, the $\{x_{e,i}^s, y_{e,i}^s\}$ is used as target data and the rest of support set $\{\mathbf{X}_e^s, \mathbf{Y}_e^s\} \setminus \{x_{e,i}^s, y_{e,i}^s\}$ are used as context set accordingly. The context set $\{\mathbf{X}_e^s, \mathbf{Y}_e^s\} \setminus \{x_{e,i}^s, y_{e,i}^s\}$ and target data $x_{e,i}^s$ are fed into the proposed model to output the prediction. The loss can be calculated between the prediction $\hat{y}_{e,i}^s$ and the corresponding label $y_{e,i}^s$. For simplicity, we use θ to represent all the parameters included in the proposed model. Then, our event adaption objective function on the support set can be represented as follows:

$$\mathcal{L}_e^s = \sum_i \log p_\theta(y_{e,i}^s | \{\mathbf{X}_e^s, \mathbf{Y}_e^s\} \setminus \{x_{e,i}^s, y_{e,i}^s\}, x_{e,i}^s). \qquad (1)$$

We then update parameters θ one or more gradient descent updates on $\mathcal{L}_e^s$ for event e. For example, when using one gradient update:

$$\theta_e = \theta - \alpha \bigtriangledown_\theta \mathcal{L}_e^s. \qquad (2)$$

Detection stage. The proposed model with event-specific parameter set θ_e will take query set $\mathbf{X}_e^q$ and entire support set $\{\mathbf{X}_e^s, \mathbf{Y}_e^s\}$ as input and output predictions $\hat{\mathbf{Y}}_e^q$ for query set $\mathbf{X}_e^q$. The corresponding loss function in the detection stage can be represented as follows:

$$\mathcal{L}_e^q = \log p_{\theta_e}(Y_e^q | X_e^s, Y_e^s, X_e^q) \qquad (3)$$

Through this meta neural process, we can learn an initialization parameter set which can rapidly learn to use given context input-outputs as conditioning to detect fake news on newly arrived events.

Neural Network Architecture. From Figure 3, we can observe that the network structures used in these two stages are the same, including feature extractor, label embedding, aggregator and detector. The feature extractor is a basic module which can take posts as input and output corresponding feature vectors. Label embedding component is to capture semantic meanings of labels. Then we use an aggregator to aggregate these information into a fixed dimensional vector, namely context embedding, which is used as reference for fake news detection. Thereafter both the context embedding and target feature vector are fed into detector to output a vector. The final prediction is based on the similarities between this output vector and label embeddings. In the following subsections, we will use event adaption to introduce the details of each component in our proposed model. For simplicity, we omitted superscript s and q in the illustrations about components.

Feature Extractor

From Figure 3, we can observe that feature extractor is a basic module to process raw input. To handle multimedia news, our feature extractor consists of two parts: textual feature extractor and visual feature extractor.

Textual feature extractor. In this paper, we choose convolutional neural network (Kim 2014), which is proven effective in the fake news detection (Wang et al. 2018, 2019), as textual feature extractor. The input of the textual feature extractor is unstructured news content, which can be represented as a sequential list of words. For the t-th word in the sentence, we represent it by the word embedding vector which is the input to the convolutional neural network. After the convolutions neural network, we feed the output into a fully connected layer to adjust the dimension to d_f dimensional textual feature vector.

Visual feature extractor. The attached images of the posts are inputs to the visual feature extractor. In order to efficiently extract visual features, we employ the pretrained VGG19 (Simonyan and Zisserman 2014) which is used in the multi-modal fake news works (Wang et al. 2018; Jin et al. 2017a). On top of the last layer of VGG19 network, we add a fully connected layer to adjust the dimension of final visual feature representation to the same dimension of textual feature vector d_f. During the joint training process with the textual feature extractor, we freeze the parameters of pre-trained VGG19 neural network to avoid overfitting.

Semantic Label Embedding across Events

Categorical characteristic of label information. The context information includes posts and their corresponding labels. The existing works like CNP (Garnelo et al. 2018a) and ANP (Kim et al. 2019) usually simply concatenate the input features and numerical label values together as input to learn a context embedding via a neural network. Such operation discards the fact that label variables are categorical. Moreover, this operation tends to underestimate the importance of labels as the dimension of input features is usually significantly larger than that of single dimensional numerical value.

Label embeddings. To handle categorical characteristic, we propose to embed labels into fixed dimension vectors inspired by word embedding (Mikolov et al. 2013). We define two embeddings $\mathbf{vec}(\mathbf{fake})$ and $\mathbf{vec}(\mathbf{real})$ for the labels of fake news and real news respectively. For example, given the i-th post $x_{e,i}$ on event e, the corresponding label is fake and its label embedding vector is $\mathbf{vec}(\mathbf{fake})$, and we denote the label embedding of $x_{e,i}$ as $\mathbf{v}_{e,i}$. As the raw input unavoidably contains noisy information, we concatenate the label embedding with high-level learned feature representation of input. Specifically, given a context data $x_{e,k} \in \mathbf{X}_e$ the feature extractor takes raw input features of context posts as input and outputs a feature representation $\mathbf{c}_{e,k} \in \mathbf{C}_e$ as output. Then the label embedding $\mathbf{v}_{e,k} \in \mathbf{V}_e$ is concatenated with $\mathbf{c}_{e,k}$ to form a vector as $\mathbf{C}_e \oplus \mathbf{V}_e = \{\mathbf{c}_{e,k} \oplus \mathbf{v}_{e,k}\}_{k=1}^{K}$.

Trainable metrics. To ensure that the label embedding can capture the semantic meanings of corresponding labels, we propose to use embeddings $\mathbf{vec}(\mathbf{fake})$ and $\mathbf{vec}(\mathbf{real})$ as metrics and output predictions are determined based on metric matching. More specifically, the similarities between output $\mathbf{o}_{e,i}$ from our model and label embeddings $\mathbf{vec}(\mathbf{fake})$ and $\mathbf{vec}(\mathbf{real})$ are first calculated based on element-wise mulitpication as follows:

$$similarity(\mathbf{o}_{e,i}, \mathbf{vec}(\mathbf{fake})) = \|\mathbf{o}_{e,i} \circ \mathbf{vec}(\mathbf{fake})\|\,, \tag{4}$$

then two similarity scores are concatenated together to form a two-dimensional score vector. We use *softmax* to map such two-dimensional score vector into $[0,1]$ as probabilities of posts being fake or legitimate. The trainable label metric can generalize easily to new events with the help of adaptation step according to Eq. 2.

Aggregator & Detector

To construct context embedding for target data, we need to design an aggregator which satisfies two properties: permutation-invariant and target-dependent. The attention mechanism can compute weights of each observations in context set with respect to the target and aggregates the values with these weight to form the value accordingly. Thus, such aggregation procedure is permutation-invariant and also target-dependent.

Attention mechanism. In this paper, we use scaled dot-product attention mechanism (Vaswani et al. 2017). This attention function can be described as mapping a query and a set of key-value pairs to an output, where the query $\mathbf{Q}$, keys $\mathbf{K}$, values $\mathbf{V}$, and output are all vectors. For simplicity, we omitted e in the following illustrations. In our problem, for the target data x_i and the context set $\mathbf{X} = \{x_k\}_{k=1}^{K}$ on event e. We use the the target feature vector $\mathbf{h}_i \in R^{1\times d}$ after linear transformation as query vector $\mathbf{Q}_i$, the context feature vector $\mathbf{C} = [c_1, ..., c_K] \in R^{K\times d}$ after linear transformation as the Key vector $\mathbf{K}$, and the concatenation of context feature and label embedding $\mathbf{C} \oplus \mathbf{V} = [c_1 \oplus v_1, ..., c_K \oplus v_K] \in R^{K\times 2d}$ after linear transformation as value vector $\mathbf{V}$. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by dot-product function of the query with the corresponding key. More specifically, attention function can be represented as follows:

$$\mathbf{a}_i = softmax(\frac{\mathbf{Q}_i\mathbf{K}^T}{\sqrt{d}}) \tag{5}$$

$$Attention(\mathbf{Q}_i, \mathbf{K}, \mathbf{V}) := a_i\mathbf{V}. \tag{6}$$

Limitation of Soft-Attention. The attention mechanism with soft weight values is categorized into **soft-attention**. However, soft-attention cannot trim irrelevant data effectively when we have a large context set or a context set with an imbalanced class distribution for target data.

Hard-Attention. To overcome the limitation of soft-attention, we propose to select the most related context data point instead of using weighted average. To enable argmax operation to be differentiable, we use Straight-Through (ST) Gumbel SoftMax (Jang, Gu, and Poole 2016) for discretely sampling the context information given target data. In our problem, for the i-th target data point x_i with context set $\mathbf{X} = \{x_k\}_{k=1}^{K}$, the class probabilities refer to the weight vector $\mathbf{a}_i = [a_{i,1}, ..., a_{i,K}]$ from dot-product attention mechanism according to Eq. 5. We can generate K-dimensional sample vectors $\mathbf{P}_i = [p_{i,1}, p_{i,2}.., p_{i,K}]$ as follows:

$$p_{i,k} = \frac{\exp((\log(a_{i,k}) + g_k)/\tau)}{\sum_k^K \exp((\log(a_{i,k}) + g_k)/\tau)} \tag{7}$$

where τ is a temperature parameter, $g = -\log(-\log(\mu))$ is the Gumbel noise and μ is generated by a certain noise distribution (e.g., $u \sim \mathcal{N}(0,1)$). As the softmax temperature τ approaches 0, samples from the Gumbel Softmax distribution become one-hot and the Gumbel-Softmax distribution becomes identical to the categorical distribution.

The hard-attention can trim the irrelevant data points and select the most related data point, denoted as $\mathbf{c}_{e,k}^{s} \oplus \mathbf{v}_{e,k}^{s} \in R^{2d}$. Besides the hard-attention mechanism, the aggregator includes an additional fully connected layer on top of hard-attention to adjust the dimension. The $\mathbf{c}_{e,k}^{s} \oplus \mathbf{v}_{e,k}^{s}$ is fed into this fully connected layer to output context embedding $r_{e,i}^{q} \in R^{d}$.

Detector. The detector is a fully-connected layer which takes target feature vector and context embedding as inputs and outputs a vector that has the same dimensionality as that of the label embedding. Then, we can calculate the similarity scores according to Eq. 4. More specifically, in the event adaption stage, the context embedding $r^s_{e,i}$ and target feature vector $h^s_{e,i}$ are concatenated. Then the detector takes $r^s_{e,i} \oplus h^s_{e,i} \in R^{2d}$ as input and produces a output vector $o^s_{e,i} \in R^d$. The similarities between output vector $o^s_{e,i}$ and label embedding vec(fake), vec(real) are calculated accordingly. The two similarity scores are then mapped into $[0, 1]$ as probabilities via *softmax*.

Experiments

Datasets

To fairly evaluate the performance of the proposed model, we conduct experiments on datasets collected from two real-world social media datasets, namely Twitter and Weibo. The detailed description of the datasets are given below:

	Twitter	Weibo
# of fake News	6,934	4,050
# of real News	5,683	3,558
# of images	514	7,606

Table 1: The Statistics of the Real-World Datasets.

The **Twitter dataset** is from MediaEval Verifying Multimedia Use benchmark (Boididou et al. 2015), which is used in (Wang et al. 2018; Jin et al. 2017b) for detecting fake content on Twitter. The **Weibo dataset**[2] is used in (Jin et al. 2017a; Wang et al. 2018; Qi et al. 2019) for detecting multimodal fake news. In the two datasets above, we only keep the events which are associated with more than 20 posts for support and query data split purpose. To validate performance of the models on never-seen events, we ensure that the training and testing sets do not contain any common event. We adopt Accuracy, F1 Score, Precision and Recall as evaluation metrics.

Baselines

To validate the effectiveness of the proposed model, we choose baselines from multi-modal models and the few-shot learning models. For the multi-modal models, we fine-tune them on given examples from events in the testing data for a fair comparison. In the experiments, we have the 5-shot and 10-shot settings. In our problem, 5-shot setting refers to that 5 labeled posts are provided as support set.

Fine-tune models. All the multi-modal approaches take the information from multiple modalities into account, including VQA (Antol et al. 2015), att-RNN (Jin et al. 2017a) and EANN (Wang et al. 2018). In the fine-tune setting, the training data includes labeled support data and labeled query data. In the testing stage, the pre-trained models are first fine-tuned on the labeled support data of given event, and then make predictions for testing query data.

Few-shot learning models. We use CNP (Garnelo et al. 2018a), ANP (Kim et al. 2019), MAML (Finn, Abbeel, and Levine 2017) and Meta-SGD (Li et al. 2017) as few-shot learning baselines.

The details of baselines and the implementations of the proposed framework can be found in Appendix.

Performance Comparison

Table 2 and Table 3 show the performance of different approaches on the Twitter and Weibo datasets. We can observe that the proposed framework MetaFEND achieves the best results in terms of most of the evaluation metrics in both 5-shot and 10-shot settings.

Twitter. On the Twitter dataset in 5-shot setting, compared with CNP, ANP incorporates the attention mechanism and hence can achieve more informative context information. Due to the heterogeneity of events, it is not easy for Meta-SGD to learn a shareable learning directions and step size across all events. Thus, Meta-SGD's performance is lower than MAML's in terms of accuracy. Compared with all the baselines, MetaFEND achieves the best performance in terms of most the metrics. Our proposed model inherits the advantages of MAML to learn a set of parameters which can rapidly learn to detect fake news with a small support set. Moreover, MetaFEND can use the support data as conditioning set explicitly to better capture the uncertainty of events and thus it is able to achieve more than 5% improvement compared with MAML in terms of accuracy. In the 10-shot setting, as the size of give support data increases, the soft attention mechanism of ANP unavoidably incorporates the irrelevant data points. In contrast, the proposed model MetaFEND employs the hard-attention mechanism to trim irrelevant data points from context set and significantly outperforms all the baselines in terms of all the metrics.

Weibo. Compared with the Twitter data, the Weibo dataset has different characteristics. Such differences can be observed from the number of unique images. On the Weibo dataset, most of the posts are associated with different images. Thus, we can evaluate the performance of models under the circumstance where support datasets do not include direct clues with query set. As EANN is not designed to capture event-specific features, it achieves the lowest accuracy among fine-tune models in 10-shot setting. For the few-shot models, ANP and CNP achieves better performance compared with gradient-based meta-learning methods MAML and Meta-SGD. This is because the parameter adaptation may not be effective when support data set and query set do not share the same patterns. Compared with ANP in 5-shot setting, our proposed method MetaFEND achieves 4.39% improvement in terms of accuracy and 5.51% improvement in terms of F1 score. The reason is that our MetaFEND can learn a base parameter which can rapidly learn to use a few examples as reference information for fake news detection. Thus, our proposed model enjoys the benefits of neural process and meta-learning model families.

Ablation Study

We show ablation study to analyze the role of Hard-Attention and label embedding components. The **case studies** of how the model works are shown in the Appendix.

Soft-Attention v.s. Hard-Attention. To intuitively illustrate the role of hard-attention mechanism in the proposed model, we conduct the following experiments. For the proposed model MetaFEND, we design its variant by replacing hard-attention by soft-attention. Then we repeatedly run the new designed model on the Twitter dataset five times in 5-shot

[2] https://github.com/yaqingwang/EANN-KDD18

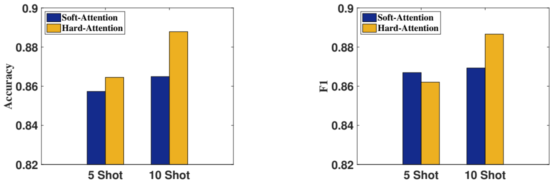
Category	Method	5-Shot				10-Shot			
		Accuracy	F1 Score	PR	Recall	Accuracy	F1 Score	PR	Recall
Fine-tune	VQA	0.7362 ± 0.0183	0.7669 ± 0.0123	0.7233 ± 0.0252	0.8172 ± 0.0204	0.7349 ± 0.0261	0.7469 ± 0.0297	0.7082 ± 0.0183	0.7915 ± 0.0530
	attRNN	0.6304 ± 0.0209	0.6025 ± 0.0463	0.6985 ± 0.0038	0.5341 ± 0.0714	0.6314 ± 0.0200	0.5660 ± 0.0525	0.6837 ± 0.0452	0.4938 ± 0.0927
	EANN	0.7001 ± 0.0358	0.7295 ± 0.0286	0.7016 ± 0.0369	0.7618 ± 0.0405	0.7056 ± 0.0100	0.6777 ± 0.0080	0.7776 ± 0.0445	0.6039 ± 0.0351
Few-shot	CNP	0.7142 ± 0.0258	0.7258 ± 0.0357	0.7389 ± 0.0316	0.7197 ± 0.0782	0.7247 ± 0.0361	0.7211 ± 0.0574	0.7191 ± 0.0218	0.7306 ± 0.1024
	ANP	0.7708 ± 0.0292	0.7965 ± 0.0381	0.7502 ± 0.0256	**0.8564 ± 0.0912**	0.7425 ± 0.0076	0.7516 ± 0.0127	0.7439 ± 0.0071	0.7605 ± 0.0318
	MAML	0.8224 ± 0.0154	0.8297 ± 0.0176	0.8599 ± 0.0993	0.8236 ± 0.1031	0.8522 ± 0.0064	0.8498 ± 0.0170	0.8627 ± 0.0678	0.8523 ± 0.0937
	Meta-SGD	0.7413 ± 0.0231	0.7535 ± 0.0256	0.7610 ± 0.0189	0.7471 ± 0.0402	0.7463 ± 0.0246	0.7457 ± 0.0274	0.7423 ± 0.0297	0.7518 ± 0.0514
	MetaFEND	**0.8645 ± 0.0183**	**0.8621 ± 0.0132**	**0.9454 ± 0.0559**	0.7950 ± 0.0237	**0.8879 ± 0.0127**	**0.8866 ± 0.0109**	**0.8962 ± 0.0578**	**0.8834 ± 0.0517**

Table 2: The performance comparison of different methods on the Twitter data under 5-shot and 10-shot settings.

Category	Method	5-Shot				10-Shot			
		Accuracy	F1 Score	PR	Recall	Accuracy	F1 Score	PR	Recall
Fine-tune	VQA	0.7693 ± 0.0071	0.7588 ± 0.0045	0.8042 ± 0.0370	0.7217 ± 0.0394	0.7780 ± 0.0143	0.7636 ± 0.0177	0.8179 ± 0.0138	0.7164 ± 0.0262
	attRNN	0.7607 ± 0.0163	0.7436 ± 0.0296	0.8050 ± 0.0125	0.6942 ± 0.0605	0.7809 ± 0.0058	0.7769 ± 0.0035	0.7930 ± 0.0133	0.7617 ± 0.0075
	EANN	0.7643 ± 0.0084	0.7451 ± 0.0056	0.8188 ± 0.0195	0.6838 ± 0.0067	0.7749 ± 0.0195	0.7656 ± 0.0128	0.8046 ± 0.0451	0.7327 ± 0.0241
Few-shot	CNP	0.7747 ± 0.0519	0.7701 ± 0.0466	0.7982 ± 0.0671	0.7459 ± 0.0409	0.7881 ± 0.0157	0.7807 ± 0.0198	0.8092 ± 0.0107	0.7547 ± 0.0319
	ANP	0.7785 ± 0.0167	0.7600 ± 0.0361	0.8346 ± 0.0415	0.7074 ± 0.0890	0.7652 ± 0.0184	0.7373 ± 0.0278	0.8379 ± 0.0277	0.6608 ± 0.0493
	MAML	0.7468 ± 0.0075	0.7416 ± 0.0033	0.7638 ± 0.0214	0.7216 ± 0.0181	0.7587 ± 0.0033	0.7341 ± 0.0086	0.8211 ± 0.0315	0.6663 ± 0.0365
	Meta-SGD	0.7173 ± 0.0181	0.6951 ± 0.0228	0.7605 ± 0.0225	0.6409 ± 0.0329	0.7334 ± 0.0235	0.7142 ± 0.0280	0.7704 ± 0.0241	0.6663 ± 0.0361
	MetaFEND	**0.8128 ± 0.0075**	**0.8019 ± 0.0127**	**0.8575 ± 0.0132**	**0.7541 ± 0.0316**	**0.8292 ± 0.0013**	**0.8237 ± 0.0028**	**0.8523 ± 0.0048**	**0.7971 ± 0.0095**

Table 3: The performance comparison of different methods on the Weibo data under 5-shot and 10-shot settings.

and 10-shot settings respectively and report the average of accuracy and F1 score values. The results are show in the Figure 4. From Figure 4a, we can observe that accuracy scores of “Hard-Attention” in 5-shot and 10-shot settings are greater than those of “Soft-Attention” respectively. However, from Figure 4b, F1 score of “Soft-Attention” is greater that of “Hard-Attention” in 5-shot setting. This is because that “Soft-Attention” achieves high recall value compared with that of ‘Hard-Attention”. The results show that the “Soft-Attention” and ‘Hard-Attention” have their own advantages in 5-shot setting. As the number of support set increases, hard-attention mechanism does not have the limitation of soft-attention mechanism which unavoidably incorporates unrelated data points and significantly outperforms the soft-attention in terms of accuracy and F1 score. Thus, we can conclude that hard-attention mechanism can take effectively advantage of support set, and the superiority is more significant as we enlarge size of support set.

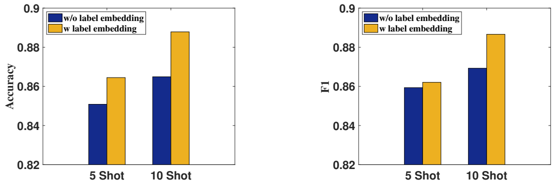


(a) Accuracy

(b) F1

Figure 4: The performance comparison for the models Soft-Attention and Hard-Attention.

w/o Label Embedding v.s. w/ Label Embedding. To analyze the role of label embedding in the proposed model, we conduct the following experiments on the Twitter dataset in 5-shot and 10-shot settings five times respectively. For the proposed model MetaFEND, we design it corresponding reduced model by replacing label embedding with label value 0 or 1. Accordingly, we change the multiplication between output with label embedding to a binary-class fully connected layer to directly output the probabilities. Figure 5 shows the results in terms of accuracy and F1 score. In Figure 5, “w/o label embedding” denotes that we remove the label embedding, and “w label embedding” denotes the original approach. From Figure 5, we can observe that both the accuracy and F1 scores of “w label embedding” are greater than “w/o label embedding” in 5-shot and 10-shot settings. It can also be observed from Figure 5, the improvement of “w label embedding” upon “w/o label embedding” is larger as the size of support set increases from 5 to 10. This further shows that the label embedding can help encode more unique semantic information of classes from different events with the help of a larger support set.



(a) Accuracy

(b) F1

Figure 5: The performance comparison for the models w/o label embedding and w embedding.

Conclusions

In this work, we study the problem of fake news detection on emergent events. The major challenge of fake news detection stems from newly emerged events on which existing approaches only showed unsatisfactory performance. In order to address this issue, we propose a novel fake news detection framework, namely MetaFEND, which can rapidly learn to detect fake news for unseen events with a few labeled examples. The proposed framework works by integrating four components including the feature extractor, the label embedding, the aggregator and the detector. The proposed framework can enjoy the benefits of meta-learning and neural process model families without suffering their own limitations. Extensive experiments on two large scale dataset collected from popular social media platforms show that our proposed model can outperform the state-of-the-art models.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.

Boididou, C.; Andreadou, K.; Papadopoulos, S.; Dang-Nguyen, D.-T.; Boato, G.; Riegler, M.; Kompatsiaris, Y.; et al. 2015. Verifying Multimedia Use at MediaEval 2015. In *MediaEval*.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5: 135–146. ISSN 2307-387X.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, 675–684. ACM.

Conroy, N. J.; Rubin, V. L.; and Chen, Y. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52(1): 1–4.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1126–1135. JMLR. org.

Garnelo, M.; Rosenbaum, D.; Maddison, C.; Ramalho, T.; Saxton, D.; Shanahan, M.; Teh, Y. W.; Rezende, D.; and Eslami, S. A. 2018a. Conditional Neural Processes. In *International Conference on Machine Learning*, 1704–1713.

Garnelo, M.; Schwarz, J.; Rosenbaum, D.; Viola, F.; Rezende, D. J.; Eslami, S.; and Teh, Y. W. 2018b. Neural processes. *arXiv preprint arXiv:1807.01622* .

Gupta, A.; Kumaraguru, P.; Castillo, C.; and Meier, P. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, 228–243. Springer.

Gupta, M.; Zhao, P.; and Han, J. 2012. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, 153–164. SIAM.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* .

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017a. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of the 2017 ACM on Multimedia Conference*, 795–816. ACM.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017b. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, 795–816.

Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; and Tian, Q. 2017c. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia* 19(3): 598–608.

Kim, H.; Mnih, A.; Schwarz, J.; Garnelo, M.; Eslami, A.; Rosenbaum, D.; Vinyals, O.; and Teh, Y. W. 2019. Attentive neural processes. *arXiv preprint arXiv:1901.05761* .

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* .

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Li, Z.; Zhou, F.; Chen, F.; and Li, H. 2017. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835* .

Lu, Y.-J.; and Li, C.-T. 2020. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. *arXiv preprint arXiv:2004.11648* .

Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B. J.; Wong, K.-F.; and Cha, M. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks. In *IJCAI*, 3818–3824.

Ma, J.; Gao, W.; and Wong, K.-F. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion Proceedings of the The Web Conference 2018*, 585–593.

Ma, J.; Gao, W.; and Wong, K.-F. 2019. Detect rumors on Twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference*, 3049–3055.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

ping Tian, D.; et al. 2013. A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering* 8(4): 385–396.

Popat, K.; Mukherjee, S.; Yates, A.; and Weikum, G. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416* .

Qi, P.; Cao, J.; Yang, T.; Guo, J.; and Li, J. 2019. Exploiting Multi-domain Visual Information for Fake News Detection. *arXiv preprint arXiv:1908.04472* .

Ruchansky, N.; Seo, S.; and Liu, Y. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797–806. ACM.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19(1): 22–36.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Tacchini, E.; Ballarin, G.; Della Vedova, M. L.; Moret, S.; and de Alfaro, L. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506* .

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vilalta, R.; and Drissi, Y. 2002. A perspective view and survey of meta-learning. *Artificial intelligence review* 18(2): 77–95.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, 849–857.

Wang, Y.; Yang, W.; Ma, F.; Xu, J.; Zhong, B.; Deng, Q.; and Gao, J. 2019. Weak Supervision for Fake News Detection via Reinforcement Learning. *arXiv preprint arXiv:1912.12520* .

Wu, K.; Yang, S.; and Zhu, K. Q. 2015. False rumors detection on sina weibo by propagation structures. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, 651–662. IEEE.

Yao, H.; Wei, Y.; Huang, J.; and Li, Z. 2019. Hierarchically structured meta-learning. *arXiv preprint arXiv:1905.05301* .

Baselines

- **VQA** (Antol et al. 2015). Visual Question Answering (**VQA**) model aims to answer the questions based on the given images.
- **att-RNN** (Jin et al. 2017a). att-RNN is the state-of-the-art model for multi-modal fake news detection. It uses attention mechanism to fuse the textual, visual and social context features. In our experiments, we remove the part dealing with social context information, but the remaining parts are the same.
- **EANN** (Wang et al. 2018). EANN is one of the state-of-the-art models for fake news detection. It consists of three components: feature extractor, event discriminator and fake news detector. It captures shared features across different events of news to improve generlziation ability. EANN has identical feature extractor with our model. On top of the feature extractor, a fully connected layer take the extracted feature to make binary classification.
- **CNP** (Garnelo et al. 2018a). Conditional neural process is the state-of-the-art model for few-shot learning. It combines neural network and gaussian process by using a small set of input-output pairs as context to output predication for given input of data.
- **ANP** (Kim et al. 2019). Attentive neural process belongs to the family of neural process which outputs prediction based on concatenation of learned distribution of context, context features and given input. .
- **MAML** (Finn, Abbeel, and Levine 2017). Model-aganostic Meta-learning is a representative optimization-based meta-learning model. The mechanism of MAML is to learn a set of shared model parameters across different tasks which can rapidly learn novel task with a small set of labeled data.
- **Meta-SGD** (Li et al. 2017). Meta-SGD is one of the state-of-the-art meta learning method for few-shot learning setting. Besides a shared global initialized parameters as with MAML, it also learns step sizes and update direction during the training procedure.

Implementations

In the proposed model, the 300 dimensional FastText pre-trained word-embedding weights (Bojanowski et al. 2017) are used to initialize the parameters of the embedding layer. The window size of filters varies from 1 to 5 for textual CNN extractor. The hidden size d_f of the fully connected layer in textual and visual extractor and dimension d are set as 16 which is searched from options $\{8, 16, 32, 64\}$. τ decays from 1 to 0.5 as the suggested way in (Jang, Gu, and Poole 2016). The gradient update step is set to 1 an inner learning rate β is set to 0.1 for fine-tune models: MAML, Meta-SGD and our proposed framework MetaFEND. We implement all the deep learning baselines and the proposed framework with PyTorch 1.2 using NVIDIA Titan Xp GPU. For training models, we use Adam (Kingma and Ba 2014) in the default setting. The learning rate α is 0.001. We use mini-batch size of 10 and training epochs of 400.

Case Study

In order to illustrate the challenges of emergent fake news detection and how our model handles challenges, we show one example in 5-shot learning setting as case study in Fig. 6. As it can be observed, the four of five news examples in the support set are real news. Due to imbalanced class condition in the support set, it is difficult for Soft-Attention to provide correct prediction for news of interest in the query set. More specifically, Fig. 6 shows the attention score values (red color) between examples in support set and query set based on multi-modal features. Although the first example with largest attention score value is most similar to news example in the query set, the majority of context information is from the other four examples due to imbalanced class distribution. Such an imbalanced class distribution leads to incorrect prediction for Soft-Attention. The Hard-Attention mechanism can achieve correct result by focusing on the most similar sample in the support set. Through this example, we can also observe the necessity of event adaption stage. The posts and images for the same event are very similar and difficult to distinguish. Without event adaption stage, the model cannot capture informative clues to make correct predictions.

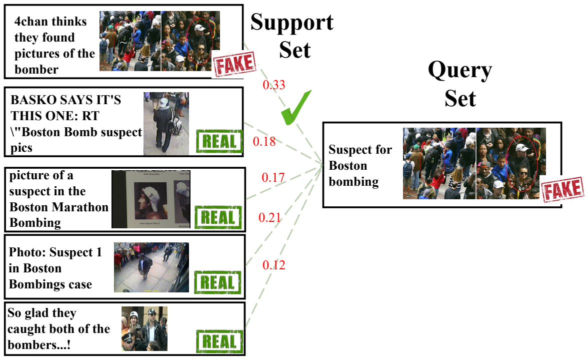

4chan thinks they found pictures of the bomber
FAKE
Support Set
0.33
✓
BASKO SAYS IT'S THIS ONE: RT \"Boston Bomb suspect pics
REAL
0.18
Query Set
Suspect for Boston bombing
FAKE
picture of a suspect in the Boston Marathon Bombing
REAL
0.17
0.21
Photo: Suspect 1 in Boston Bombings case
REAL
0.12
So glad they caught both of the bombers...!
REAL

Figure 6: Fake news examples missed by Soft-Attention but spotted by Hard-Attention