Multimodal Emergent Fake News Detection via Meta Neural Process Networks

Yaqing Wang[§], Fenglong Ma[◇], Haoyu Wang[§], Kishlay Jha[†] and Jing Gao[§]
[§]Purdue University, West Lafayette, Indiana, USA
[◇]Pennsylvania State University, Pennsylvania, USA
[†]University of Virginia, Charlottesville, Virginia, USA
[§]{wang5075, jinggao, wang5346}@purdue.edu, [◇]fenglong@psu.edu, [†]kishlay@email.virginia.edu

KDD'21

210826 Chia-Chun Ho

Outline

Introduction

Problem Formulation

Preliminary Work

Methodology

Experiments

Case Study

Conclusions

Comments

Introduction

Task Challenges

- Despite the success of deep learning models with large amounts of labeled datasets, the algorithms still suffer in the cases where fake news detection is needed on emergent events.
- Adding the knowledge from newly emergent events requires to build a new model from scratch or continue to fine-tune the model on newly collected labeled data.
  - Be challenging, expensive, and unrealistic for real-world settings.
- How to leverage a small set of verified posts to make the model learn quickly to detect fake news on the newly-arrived events is a crucial challenge.

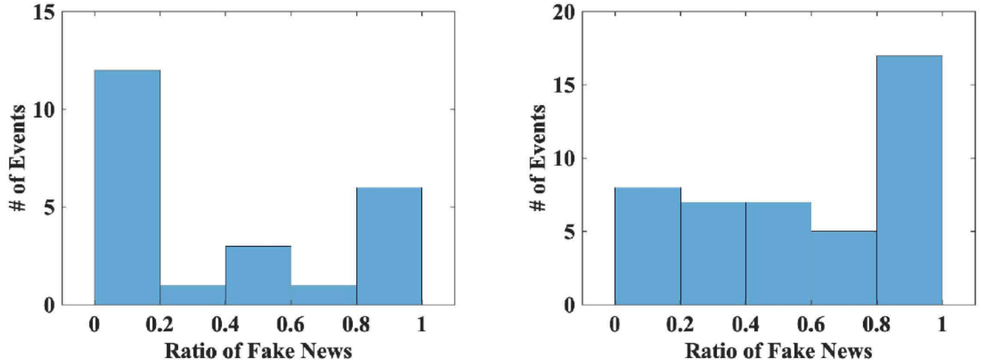
Introduction

Limitations of Current Techniques

- To overcome the challenge as just mentioned, the few-shot learning, which aims to leverage a small set of data instances for quick learning, is a possible solution.
- Basic idea of meta-learning is to leverage the global knowledge from previous tasks to facilitate the learning on new task.
- Existing methods is highly associated with an important assumption:
  - The tasks are from a similar distribution and the shared global knowledge applies to different tasks.

Introduction

Assumption of Meta-Learning Methods



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

- The assumption usually doesn't 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.
- Observed from the figure that the ratios of fake news on events are significantly different.
- The significant difference across events posts serious challenges on event heterogeneity, which cannot be simply handled by globally sharing knowledge.

Introduction

Neural process and its limitations

- Another research line of few-shot learning is neural processes, which conduct inference inference using a small set of data instances as conditioning.
- Even through neural processes show better generalizability,
  - they are based on a fixed set of parameters and usually suffer from the limitations like under-fitting,
  - thereby leading to unsatisfactory performance.

Introduction

Analysis of two research lines of models

- Two research lines of models are complementary to each other.
- The parameter adaptation mechanism in meta-learning can provide more flexibility to alleviate unfitting issues of the neural process.
- The neural processes can help handle the heterogeneity challenge for MAML by using a small set of data instances as conditioning instead of encoding all the information into parameter set.
- Although it's promising to integrate two popular few-shot approaches together, the incompatible operations on the given small set of data instances is main obstacle.

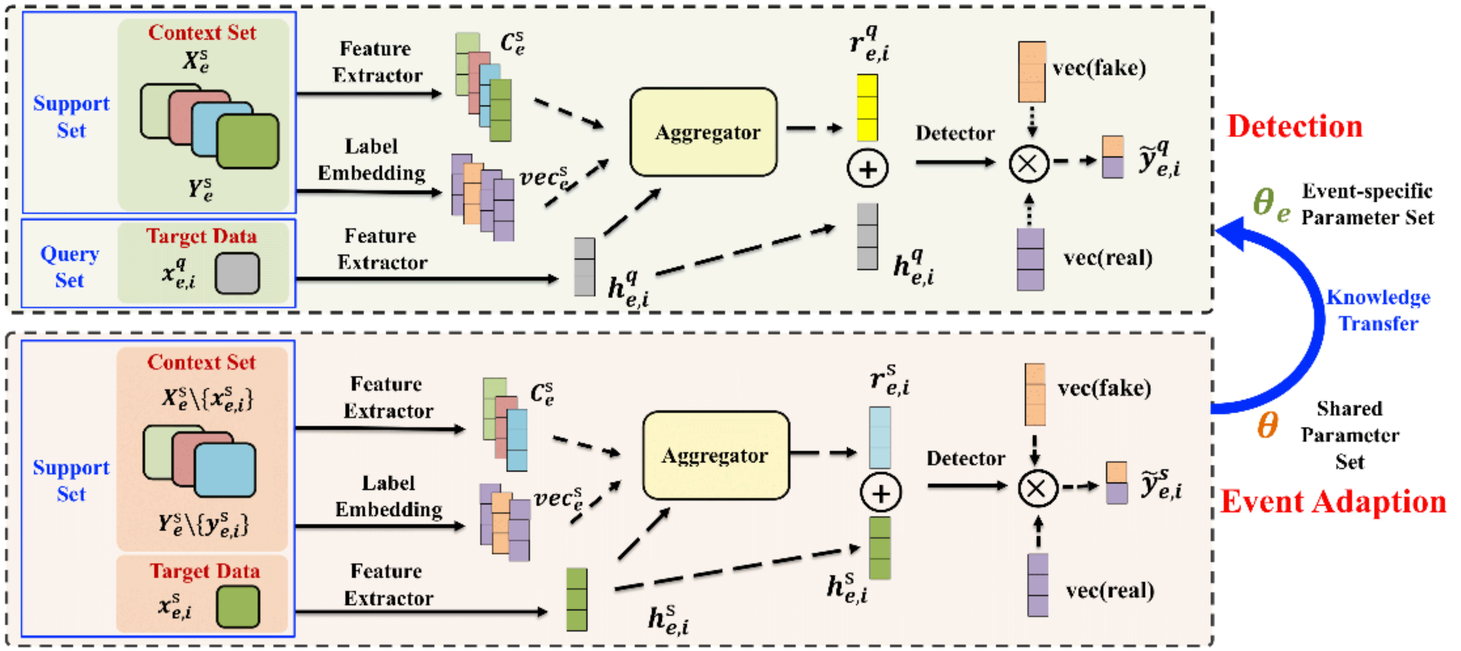
Introduction

Proposed Approach: MetaFEND

- To address the aforementioned challenges, proposed a novel meta neural process network (MetaFEND) for emergent fake news detection.
- MetaFEND unifies the incompatible operation from meta-learning and neural process via simple yet novel simulated learning task,
  - whose goal is to adapt the parameters to better take advantage of given support data points as conditioning.

Introduction

Proposed Approach: MetaFEND



Overview of MetaFEND

- Proposed to conduct leave-one-out prediction as shown in figure.
- Repeatedly use one of given data as target data and the rest are used as context set for conditioning on all the data support set.
- Therefore, MetaFEND can handle heterogeneous events via event adaption parameters and conditioning on event-specific data instances simultaneously.

Introduction

Components in MetaFEND

- Incorporate two novel components
  - **Label embedding**
    - Handle categorical characteristic of label information.
  - **Hard attention**
    - Extract the most informative instance as conditioning despite imbalanced class distribution of news events.

Introduction

Contributions of MetaFEND

- Recognize the challenges of fake news detection on emergent events and formulate the problem into a few-shot learning setting.
- Proposed MetaFEND to detect fake news on emergent events with a handful of data instances by fusing the meta-learning method and neural process models together via a simulated learning task design.
- Also propose label embedding and hard attention to handle categorical information and select the informative instance respectively.

Problem Formulation

Notation

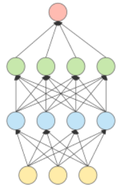
- Core idea of few-shot learning is to use episodic classification paradigm to simulate few-shot settings during model training.
- $\mathscr{E}$: set of news event, $e \in \mathscr{E}$: news event (which has a few labeled posts)
- In each episode during training stage, the labeled posts are partitioned into two independent sets:
  - Support set: $\left\{\mathbf{X}_e^s, \mathbf{Y}_e^s\right\} = \left\{x_{e,i}^s, y_{e,i}^s\right\}_{i=1}^{K}$, Query set: $\left\{\mathbf{X}_e^q, \mathbf{Y}_e^q\right\} = \left\{x_{e,i}^q, y_{e,i}^q\right\}_{i=K+1}^{N}$
- For each event e, the model leverages its corresponding K labeled posts as support set to conduct fake news detection on given event e.

Preliminary Work

Meta-Learning

Meta Learning

Machine Learning ≈ 根據資料找一個函數 f 的能力

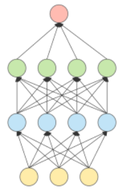


$f($  $) =$ “Cat”

Meta Learning

≈ 根據資料找一個找一個函數 f 的函數 F 的能力

$F($ 

$) = f^*$ 

https://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Meta1%20(v6).pdf

Preliminary Work

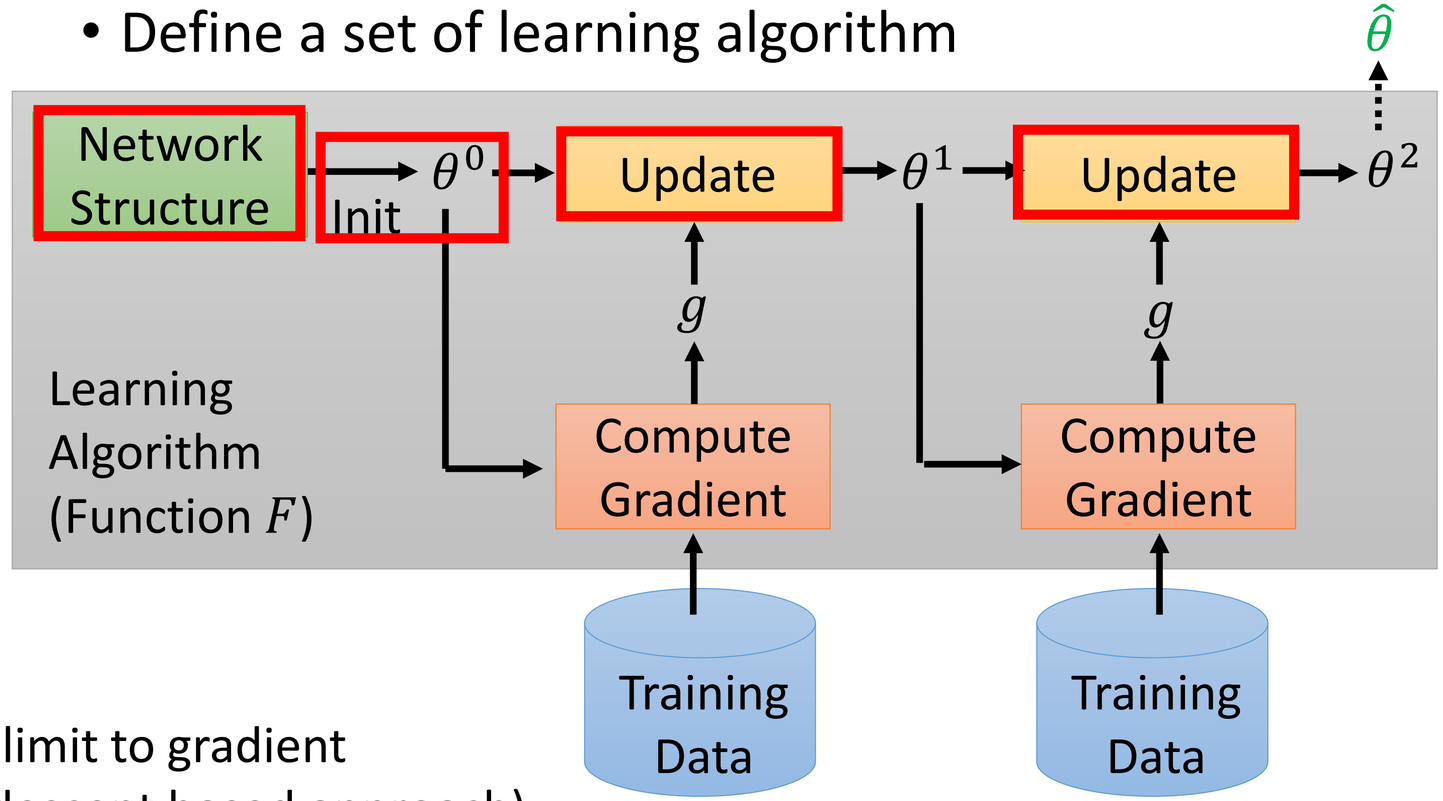
Meta-Learning

Meta Learning

Different decisions in the red boxes lead to different algorithms.

What happens in the red boxes is decided by humans until now.

- Define a set of learning algorithm



(limit to gradient descent based approach)

https://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Meta1%20(v6).pdf

Preliminary Work

Meta-Learning

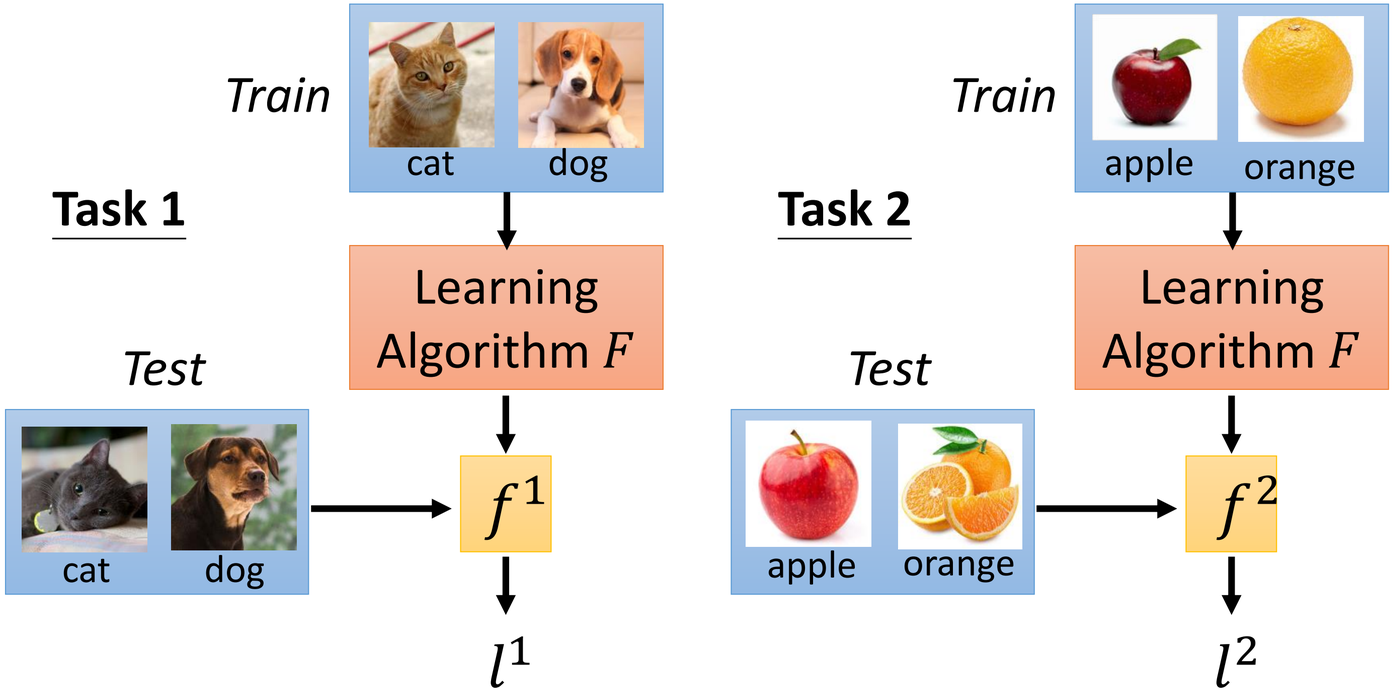
Meta Learning

$$L(F) = \sum_{n=1}^{N} l^n$$

N: N tasks

l^n: Testing loss for task n after training

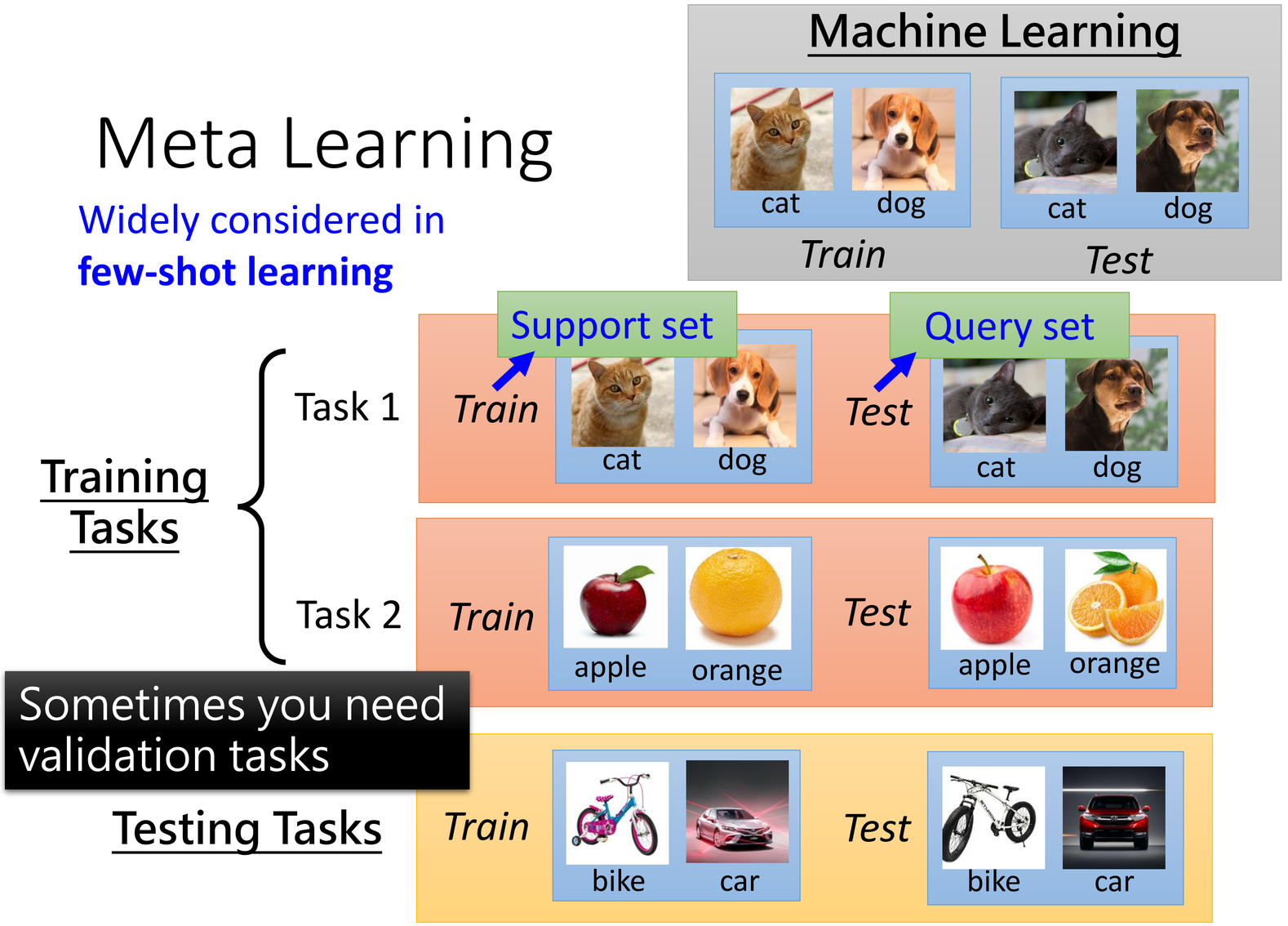
- Defining the goodness of a function F



https://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Meta1%20(v6).pdf

Preliminary Work

Meta-Learning



https://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Meta1%20(v6).pdf

Preliminary Work

Meta-Learning

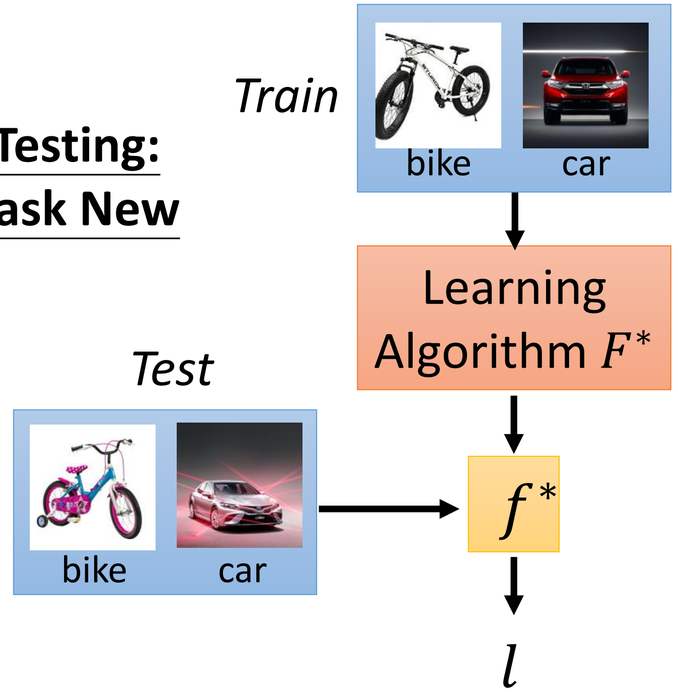
Meta Learning

- Defining the goodness of a function F

$$L(F) = \sum_{n=1}^{N} l^n$$

- Find the best function F^*

$$F^* = arg \min_F L(F)$$



https://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Meta1%20(v6).pdf

Preliminary Work

MAML (Model-Agnostic Meta-learning)

- Proposed at ICML'17
- MAML is a representative algorithm of gradient-based meta-learning approaches.

Preliminary Work

MAML

MAML

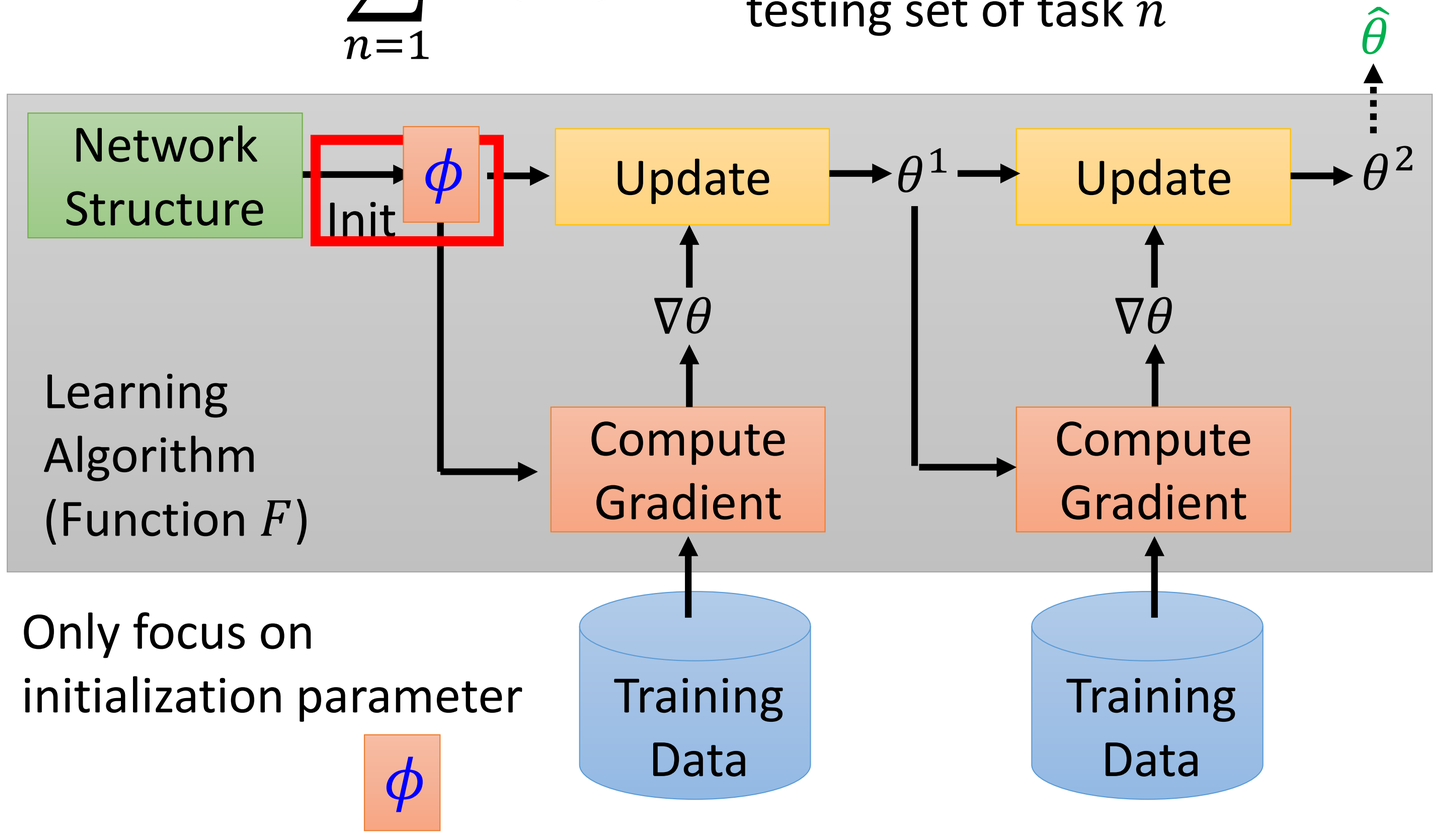
$\hat{\theta}^n$: model learned from task n

$\hat{\theta}^n$ depends on ϕ

Loss Function:

$$L(\phi) = \sum_{n=1}^{N} l^n(\hat{\theta}^n)$$

$l^n(\hat{\theta}^n)$: loss of task n on the testing set of task n



https://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Meta1%20(v6).pdf

Preliminary Work

MAML

MAML

$\hat{\theta}^n$: model learned from task n

$\hat{\theta}^n$ depends on ϕ

Loss Function:

$$L(\phi) = \sum_{n=1}^{N} l^n\left(\hat{\theta}^n\right)$$

$l^n\left(\hat{\theta}^n\right)$: loss of task n on the testing set of task n

How to minimize $L(\phi)$? Gradient Descent

$$\phi \leftarrow \phi - \eta \nabla_{\phi} L(\phi)$$

Model Pre-training

Widely used in transfer learning

Loss Function:

$$L(\phi) = \sum_{n=1}^{N} l^n(\phi)$$

https://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Meta1%20(v6).pdf

Preliminary Work

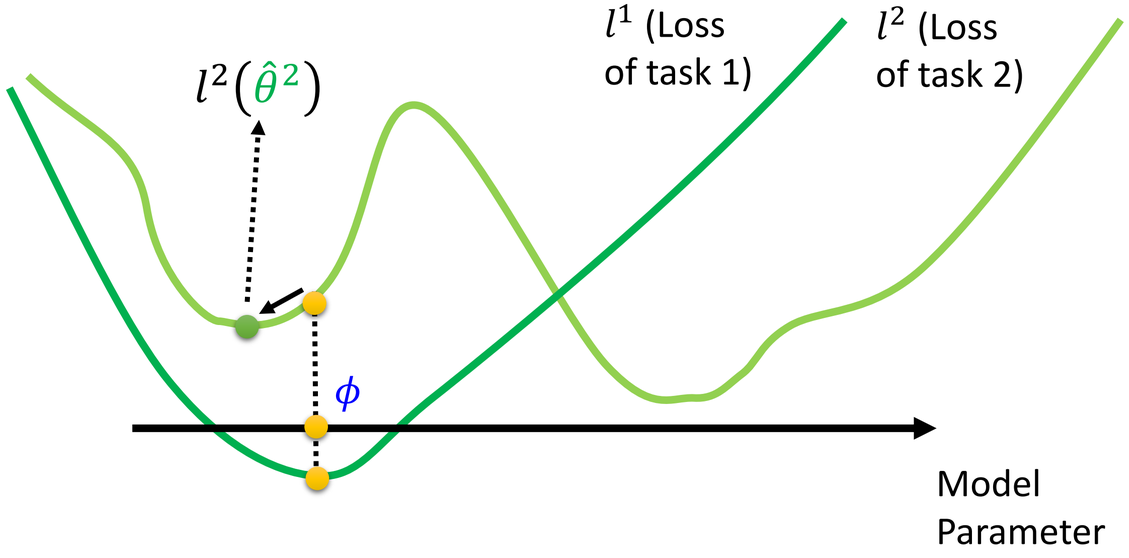
MAML

Model Pre-training

$$L(\phi) = \sum_{n=1}^{N} l^n(\phi)$$

找尋在所有 task 都最好的 ϕ

並不保證拿 ϕ 去訓練以後會得到好的 $\hat{\theta}^n$



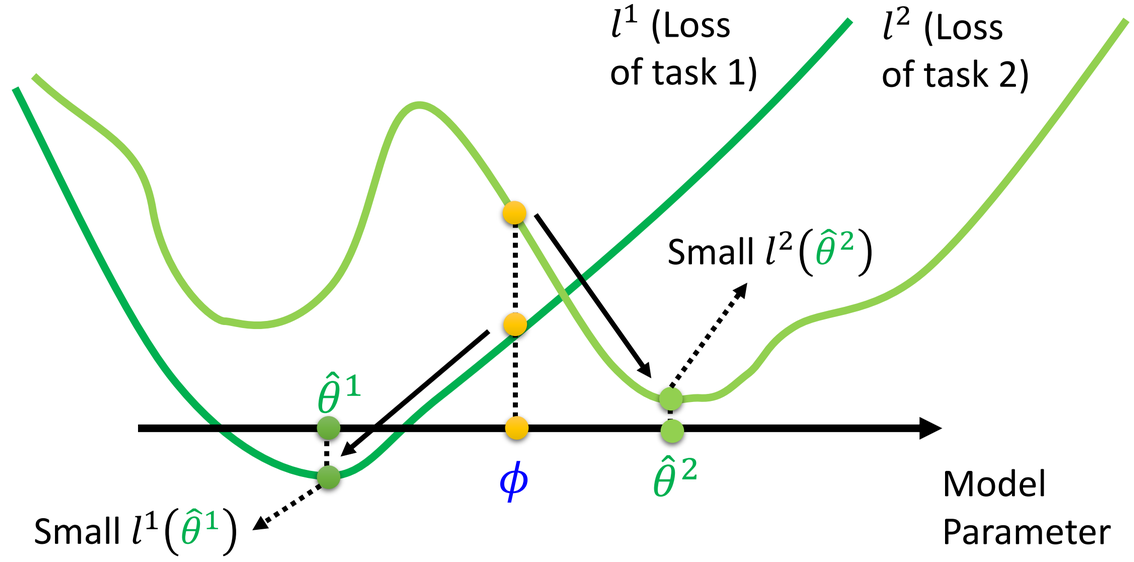
https://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Meta1%20(v6).pdf

MAML

$$L(\phi) = \sum_{n=1}^{N} l^n(\hat{\theta}^n)$$

我們不在意 ϕ 在 training task 上表現如何

我們在意用 ϕ 訓練出來的 $\hat{\theta}^n$ 表現如何



https://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Meta1%20(v6).pdf

Preliminary Work

MAML

MAML

$\hat{\theta}^n$: model learned from task n

$\hat{\theta}^n$ depends on ϕ

Loss Function:

$$L(\phi) = \sum_{n=1}^{N} l^n(\hat{\theta}^n)$$

$l^n(\hat{\theta}^n)$: loss of task n on the testing set of task n

How to minimize $L(\phi)$? Gradient Descent

$$\phi \leftarrow \phi - \eta \nabla_\phi L(\phi)$$

Find ϕ achieving good performance **after training** 潛力

Model Pre-training

Widely used in transfer learning

Loss Function:

$$L(\phi) = \sum_{n=1}^{N} l^n(\phi)$$

Find ϕ achieving good performance 現在表現如何

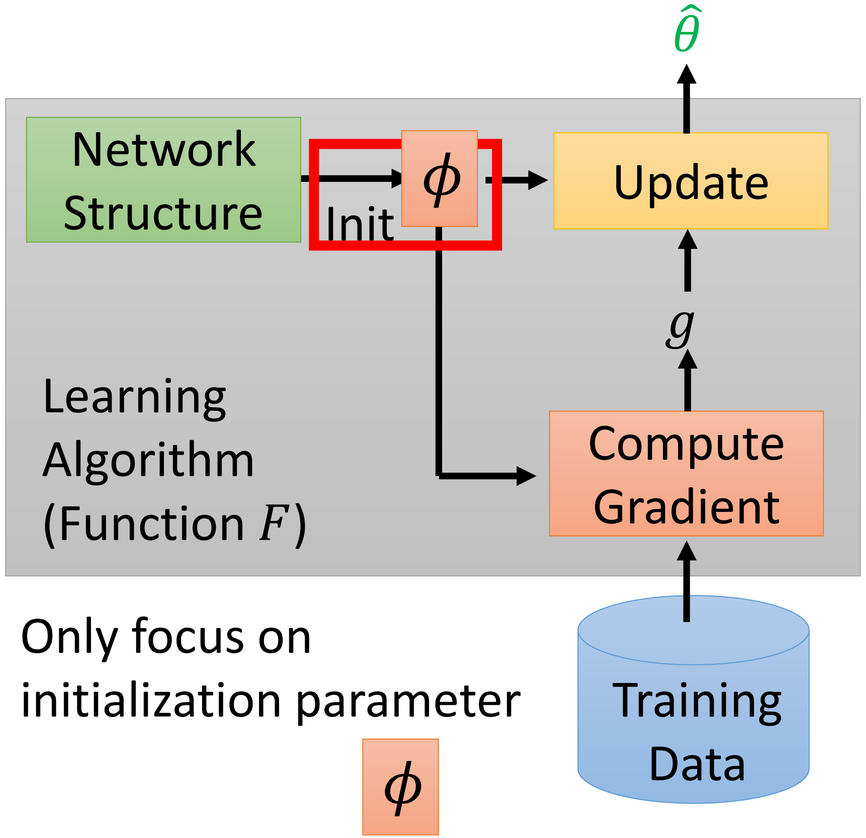
https://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Meta1%20(v6).pdf

Preliminary Work

MAML

MAML

- Fast … Fast … Fast …
- Good to truly train a model with one step. ☺
- When using the algorithm, still update many times.
- Few-shot learning has limited data.



$$L(\phi) = \sum_{n=1}^{N} l^n(\hat{\theta}^n)$$

$$\phi \leftarrow \phi - \eta \nabla_{\phi} L(\phi)$$

Considering one-step training:

$$\hat{\theta} = \phi - \varepsilon \nabla_{\phi} l(\phi)$$

https://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Meta1%20(v6).pdf

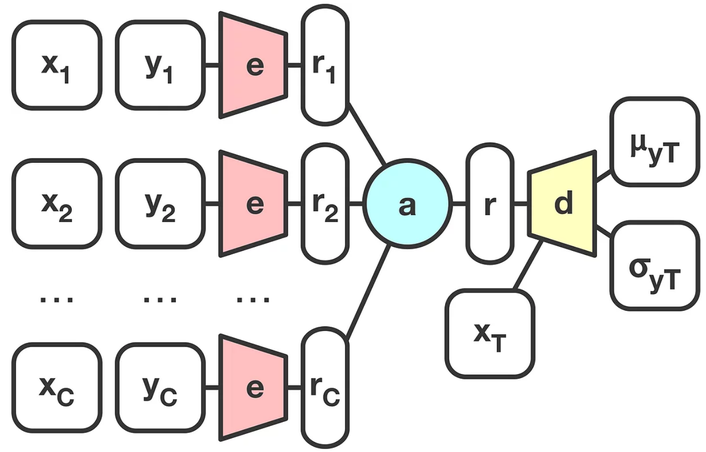
Preliminary Work

Limitations of MAML

- MAML can capture task uncertainty via one or several gradient updates.
- However, in fake news detection problem, when events are heterogeneous, the event uncertainty is difficult to encode into parameters via one or several gradient steps.
- Moreover, even support and query data from same event, there's no guarantee that they are all highly related to each other.
- In such a case, the parameter adaptation on fake news detection loss on support set may be misleading for some posts.

Preliminary Work

Conditional Neural Process (CNP)

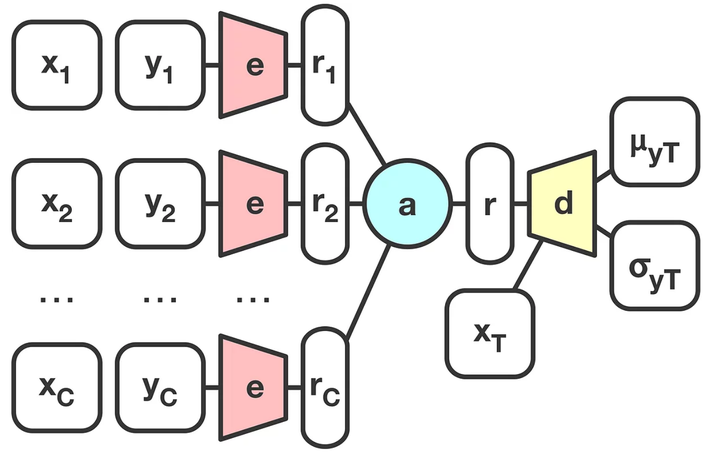


https://colab.research.google.com/github/deepmind/neural-processes/blob/master/conditional_neural_process.ipynb

- DeepMind proposed at ICML'18
- The basic idea of CNP is to make predictions with help of support set as context.
- Includes four components:
  - The neural network encoder embeds each observation in support set into feature vector.
  - The aggregator maps these feature vectors into an embeddings of fixed dimension.
  - The query data is fed into feature extractor to get the feature vector.
  - Then the decoder takes the concatenation of aggregated embedding and given target data as input and output the corresponding predictions.

Preliminary Work

Limitations of CNP

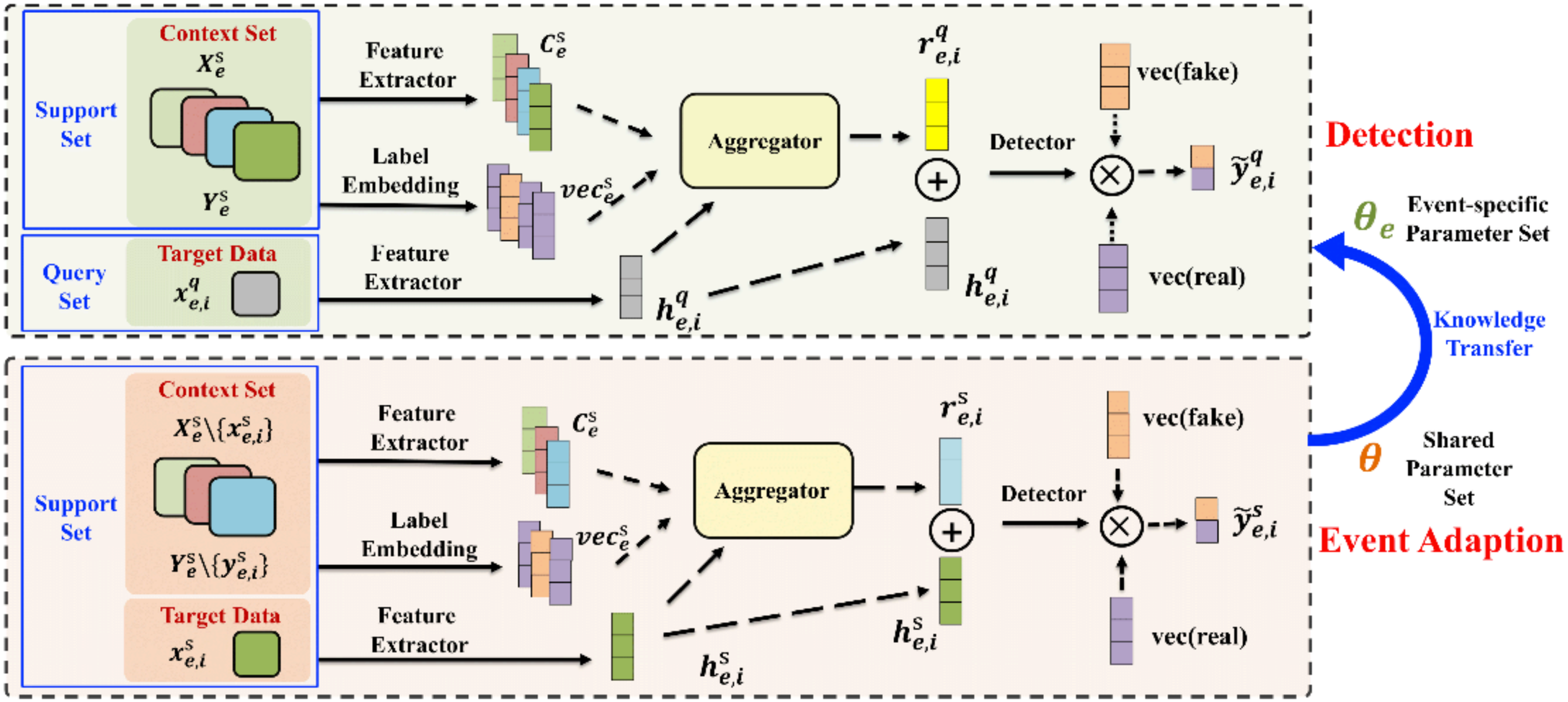


https://colab.research.google.com/github/deepmind/neural-processes/blob/master/conditional_neural_process.ipynb

- Under-fitting
- For different context data points, their importance is usually different in the prediction.
- However, the aggregator of CNP treat all the support data equally and can't achieve query-dependent context information.
- Moreover, the CNP simply concatenates the input features and numerical label values of post together as input, ignoring the categorical characteristic of labels.

Methodology

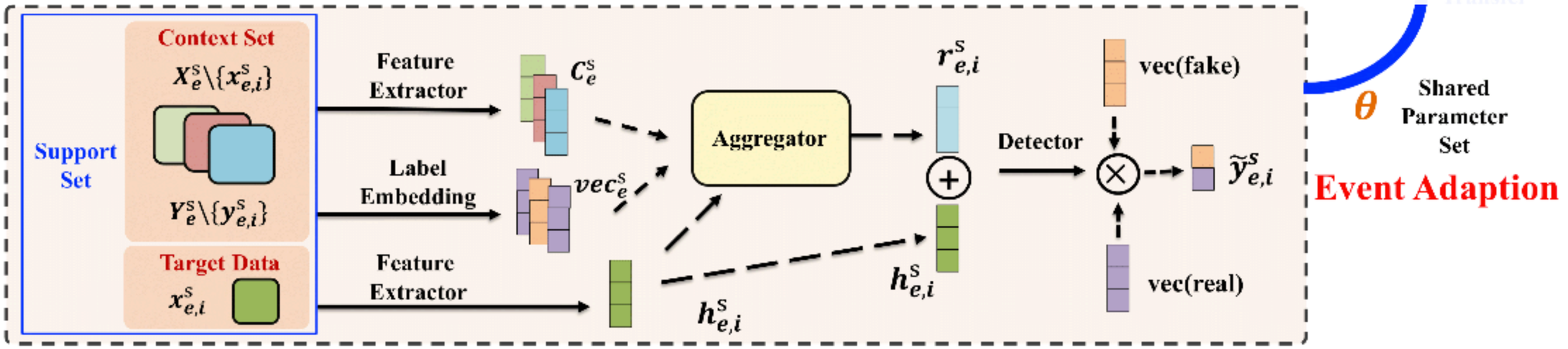
Meta-learning Neural Process Design



Methodology

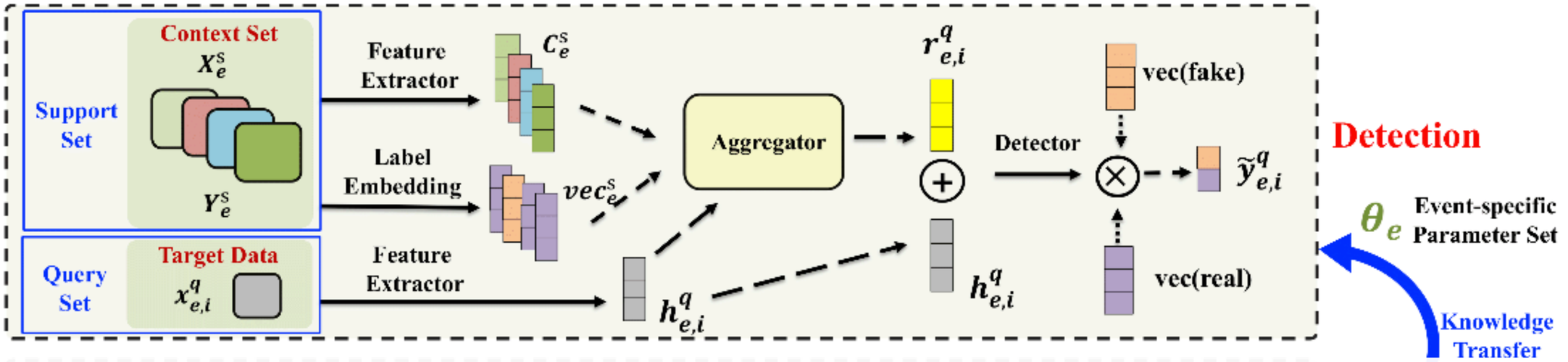
Event adaption stage

- Event adaption stage is to adapt the model parameters for specific event with the help of the support set.



Methodology

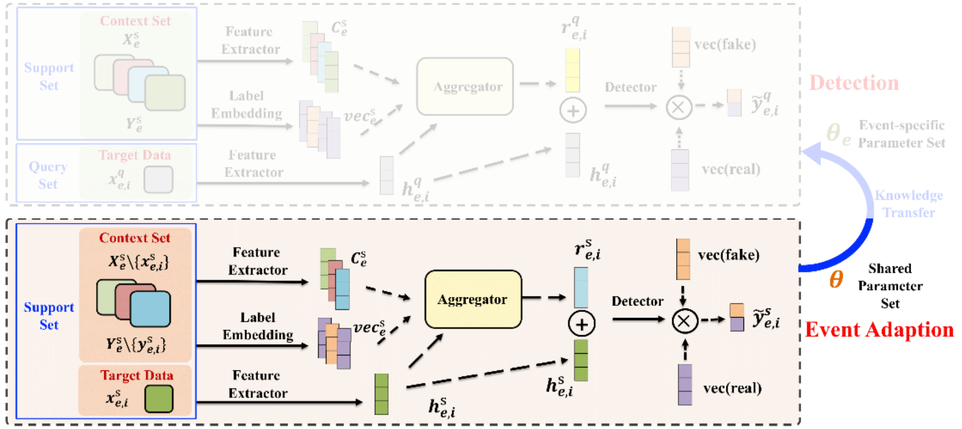
Detection stage



- The detection stage is to detect fake news on the given event with the help of the support set and the adapted parameter set.

Methodology

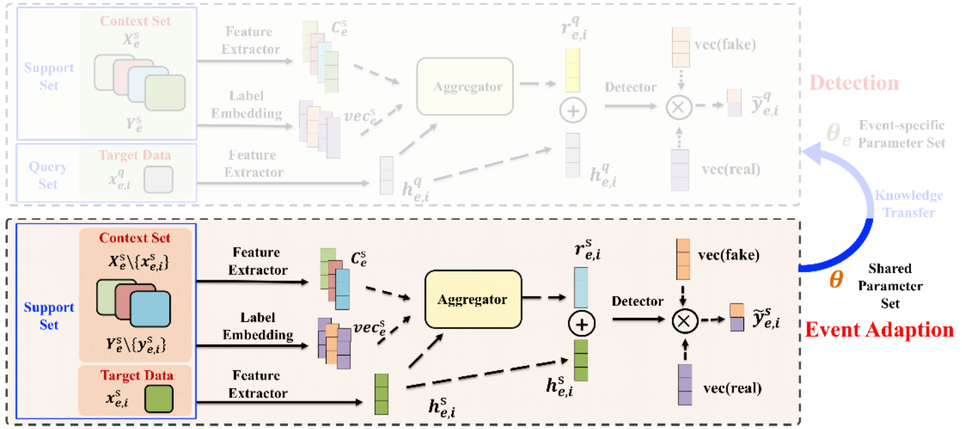
Event adaption stage



- Take i-th support data $\{x_{e,i}^s, y_{e,i}^s\}$ as an example.
- In the event adaption stage, $\{x_{e,i}^s, y_{e,i}^s\}$ is used as target data and rest of support set $\{\mathbf{X}_e^s, Y_e^s\} \backslash \{x_{e,i}^s, y_{e,i}^s\}$ are used as context set accordingly.
- The context set and target data $x_{e,i}^s$ are fed into the proposed model to output the prediction.

Methodology

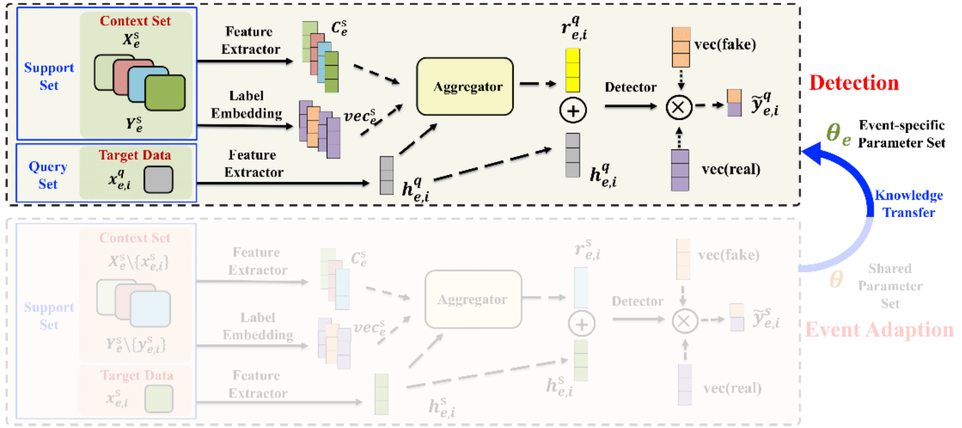
Event adaption stage



- The loss can be calculated between the prediction $\hat{y}^s_{e,i}$ and the corresponding label $y^s_{e,i}$.
- θ: all parameters included in the proposed model.
- The event adaption objective function on the support set can be represented as
  - $$\mathscr{L}^s_e = \sum_i \log p_\theta \left(y^s_{e,i} \mid \{\mathbf{X}^s_e, Y^s_e\} \backslash \left\{ x^s_{e,i}, y^s_{e,i} \right\}, x^s_{e,i} \right)$$
- Then update parameters θ one gradient descent updates on $\mathscr{L}^s_e$ for event e.
  - $\theta_e = \theta - \alpha \nabla_\theta \mathscr{L}^s_e$

Methodology

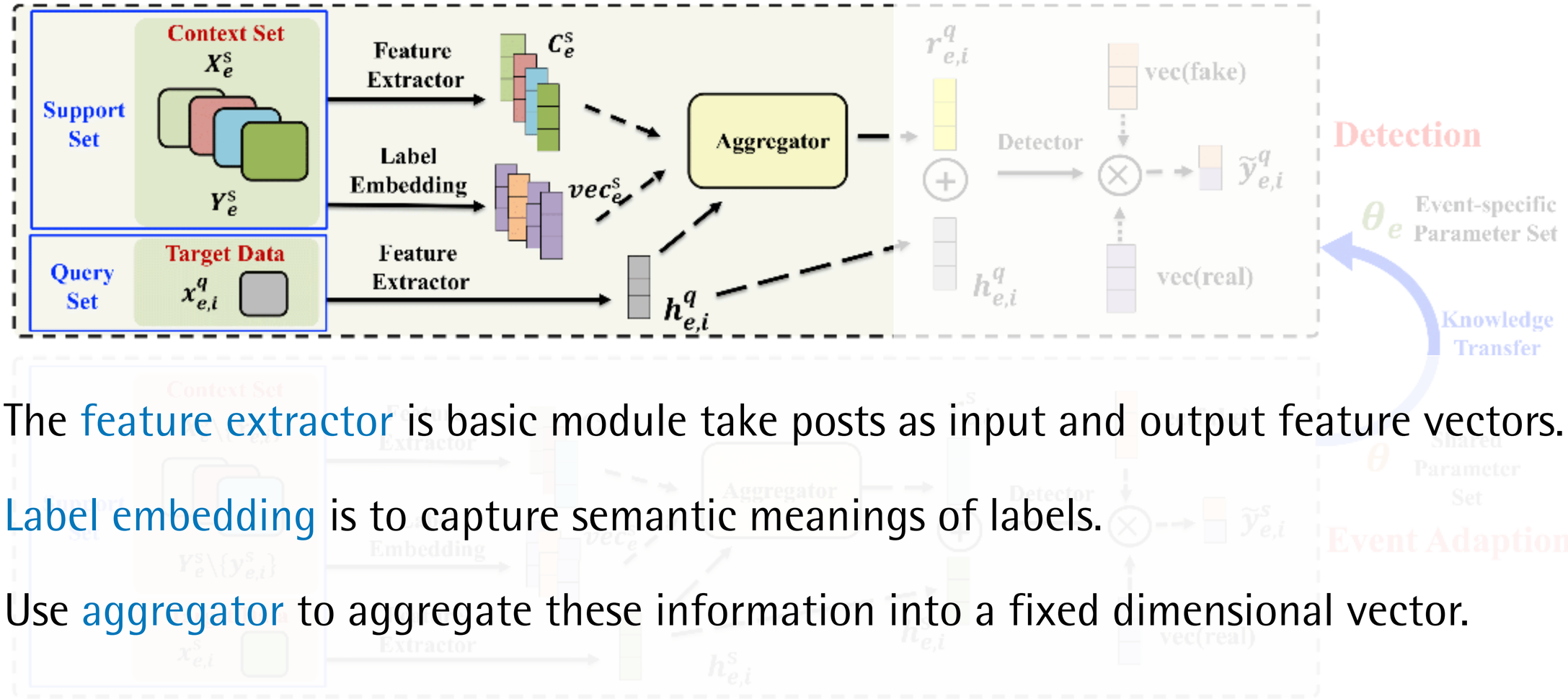
Detection stage



- The proposed model with event-specific parameter set θ_e takes query set $\mathbf{X}_e^q$ and entire support set $\{\mathbf{X}_e^s, \mathbf{Y}_e^s\}$ as input and outputs predictions $\hat{\mathbf{Y}}_e^q$ for query set $\mathbf{X}_e^q$.
- The loss function in the detection stage can be represented as
  - $\mathcal{L}_e^q = \log p_{\theta_e}\left(Y_e^q \mid X_e^s, Y_e^s, X_e^q\right)$
- Through this meta neural process, we can learn an initialization parameter set θ which can rapidly learn to use given context input-outputs as conditioning to detect fake news on newly arrived events.

Methodology

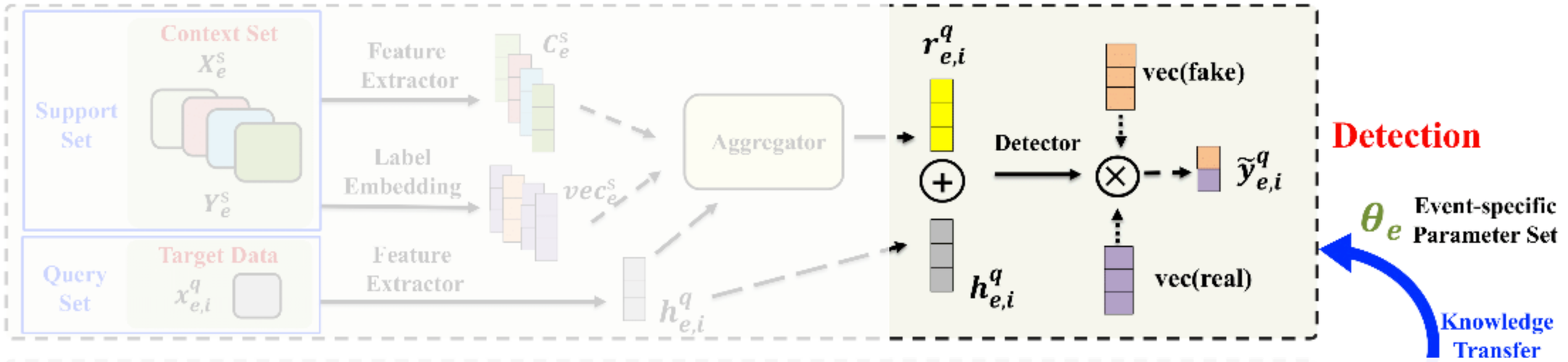
Neural Network Architecture



- The feature extractor is basic module take posts as input and output feature vectors.
- Label embedding is to capture semantic meanings of labels.
- Use aggregator to aggregate these information into a fixed dimensional vector.

Methodology

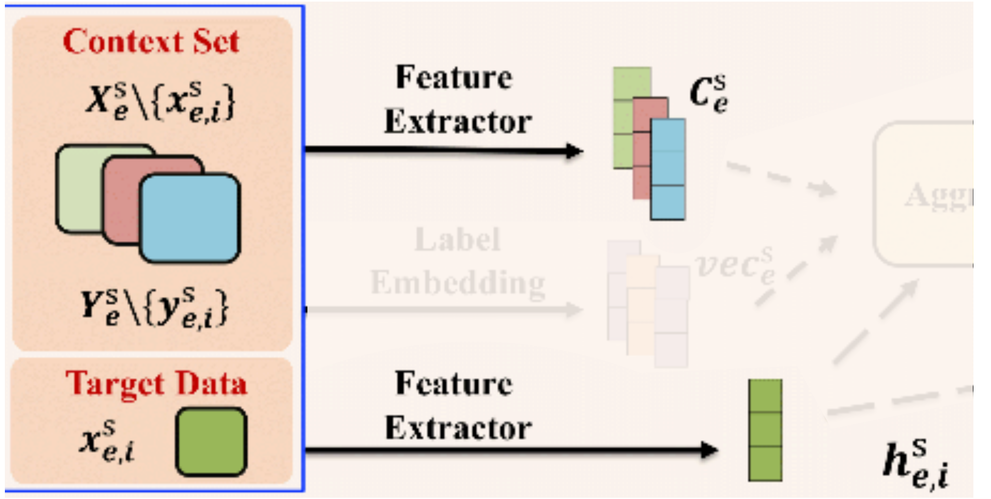
Neural Network Architecture



- Thereafter both 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.

Methodology

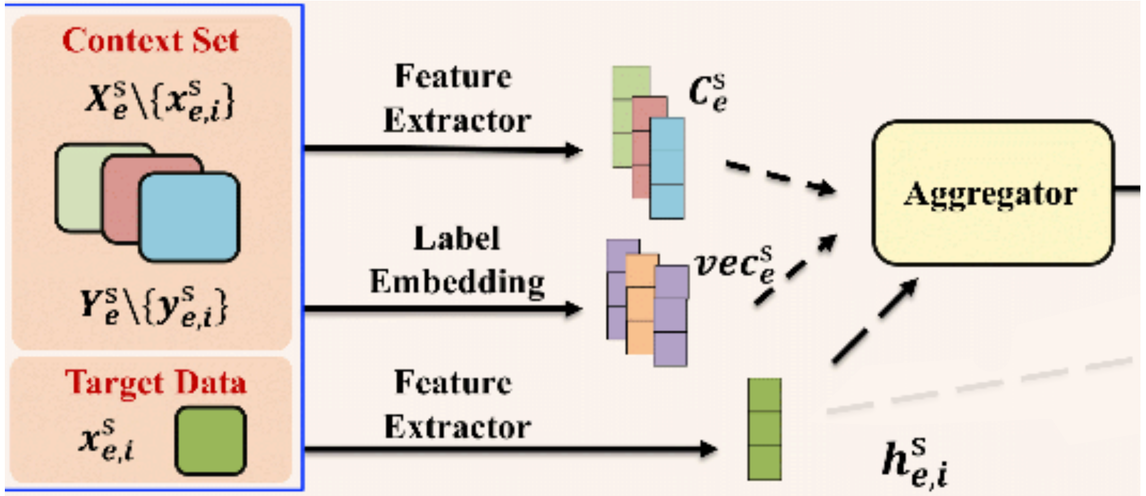
Feature Extractor



- **Textual feature extractor**
  - Text-CNN
  - Use 300-dimensional FastText pre-trained word-embedding
- **Visual feature extractor**
  - Pre-trained VGG-19
- On the top of extractors, both add a fully connected layer to adjust dimension to d_f.
- The output of two extractors are concatenated together to form a feature vector.

Methodology

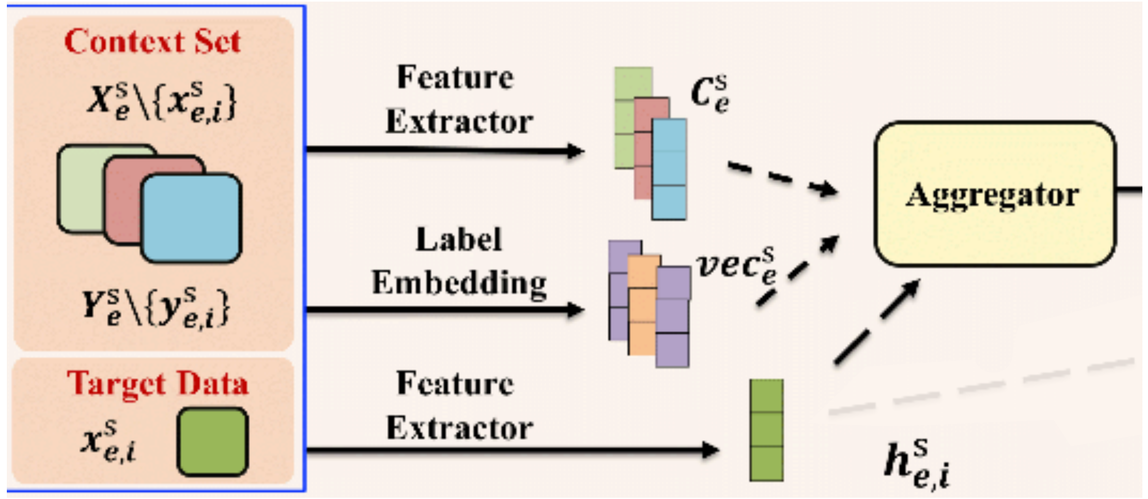
Aggregator



- Design aggregator satisfies two properties:
  - Permutation-invariant & Target-dependent
  - Adopt the attention mechanism
    - Compute weights of each observations in context set with respect to the target and aggregate the values according to their weights to form the new value.

Methodology

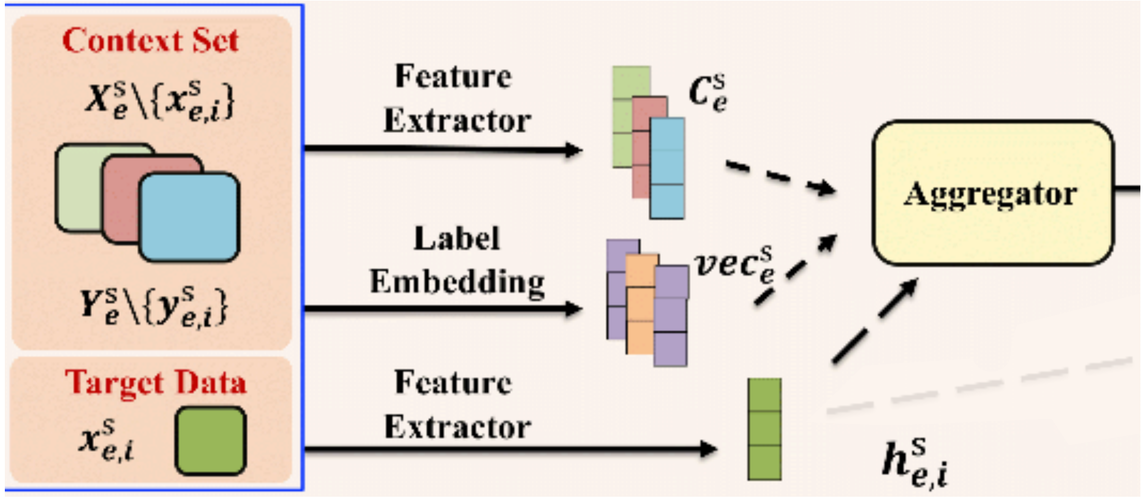
Aggregator: Attention



- Use scaled dot-product attention mechanism
- Mapping a query $\mathbf{Q}$ and a set of key $\mathbf{K}$ - value $\mathbf{V}$ pairs to an output
- $\mathbf{Q}_i = \mathbf{W}_q \mathbf{h}_{e,i}$, $\mathbf{K} = \mathbf{W}_k \mathbf{C}_e$, $\mathbf{V} = \mathbf{W}_v \left(\mathbf{C}_e \oplus \mathbf{vec}_e \right)$
- $\mathrm{a}_i = \mathrm{softmax} \left(\dfrac{\mathbf{Q}_i \, \mathrm{K}^T}{\sqrt{d}} \right)$
- Attention $\left(\mathbf{Q}_i, \mathbf{K}, \mathbf{V} \right) := a_i \mathbf{V}$

Methodology

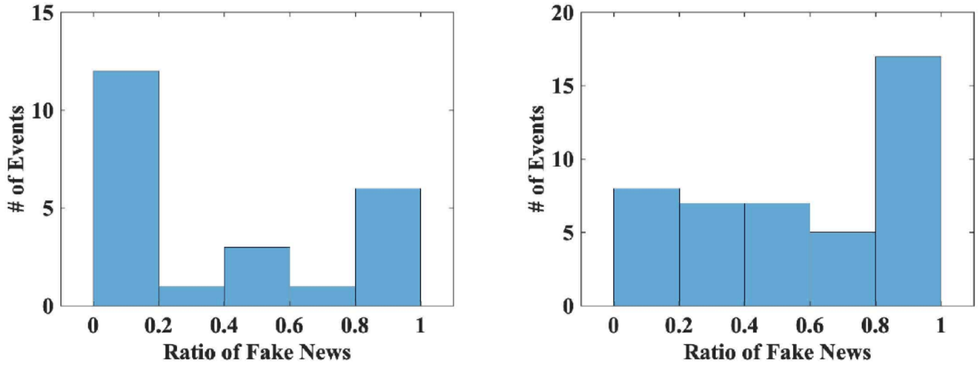
Aggregator: Limitation of Soft-attention



- $$\mathrm{a}_i = \mathrm{softmax}\left(\frac{\mathbf{Q}_i\,\mathrm{K}^T}{\sqrt{d}}\right)$$

- The attention mechanism with soft weight values is categorized into soft-attention.

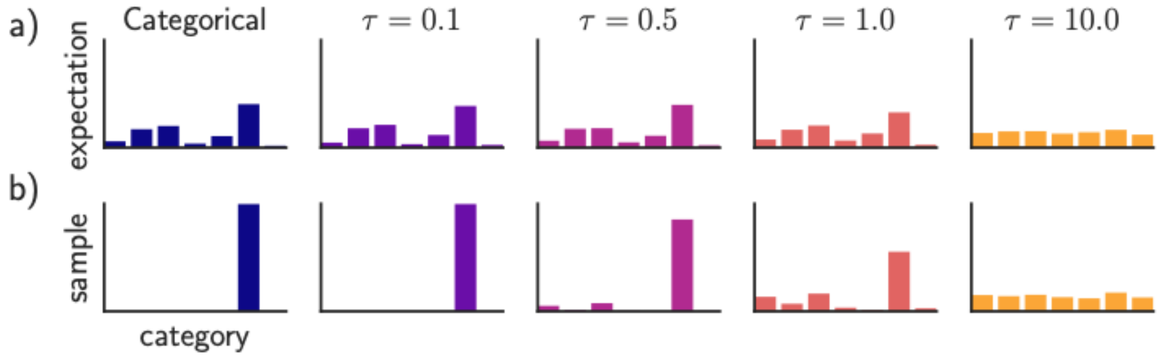
- However, soft-attention cannot effectively trim irrelevant data especially when have a context set with an imbalanced class distribution as mentioned before.



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

Methodology

Aggregator: Hard-Attention

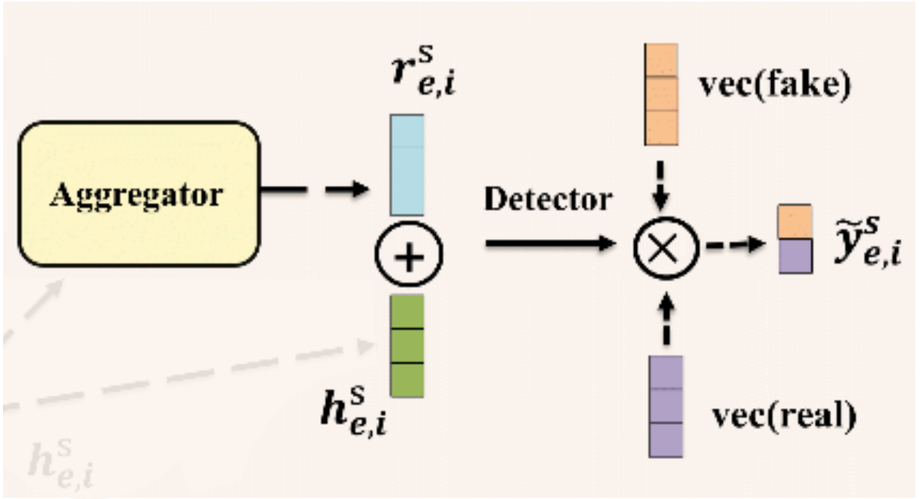


https://arxiv.org/abs/1611.01144

- To overcome this limitation, propose to select the most related context data point instead of weighted average.
- To enable argmax operation to be differentiable, use Straight-Through (ST) Gumbel SoftMax (ICLR'17) for discretely sampling the context information given target data.
- Through gumbel-softmax, the hard-attention is able to trim the irrelevant data and draw the most informative sample for given target sample $x_{e,i}$.
- The selected data point $\mathbf{c}_{e,k} \oplus \mathbf{v}_{e,k}$ is fed into fully connected layer that top of the aggregator to adjust dimension and output context embedding $\mathbf{r}_{e,i}$.

Methodology

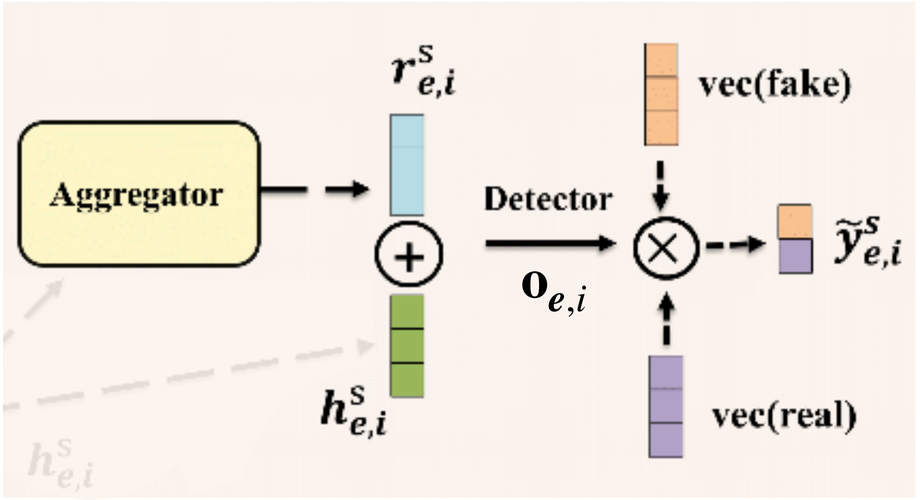
Detector based on Label Embedding



- Existing works like CNP and ANP usually simply concat the input feature and numerical label values together as input.
- These works discard the fact that label variables are categorical, and underestimate the importance of labels as dimension of input features is usually significantly larger than single dimensional numerical value.
- Propose to embed labels into fixed dimension vector inspired by word embedding.

Methodology

Detector based on Label Embedding



- Define two embeddings **vec(fake)** and **vec(real)**.
- To ensure that the label embedding can capture the semantic meanings of corresponding labels, propose to use embeddings **vec(fake)** and **vec(real)** in the detector as metrics and output prediction are determined based on metric matching.
- The detector is fully connected layer output vector $\mathbf{o}_{e,i}$.
- similarity $\left(\mathbf{o}_{e,i}, \text{vec(fake)}\right) = \|\mathbf{o}_{e,i} \circ \text{vec(fake)}\|$,
  similarity $\left(\mathbf{o}_{e,i}, \text{vec(real)}\right) = \|\mathbf{o}_{e,i} \circ \text{vec(real)}\|$
- The two scores then mapped into [0,1] as probabilities via softmax.

Experiments

Datasets

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

- Twitter, Weibo datasets
- The news events are included in the Twitter dataset, obtain events on Weibo dataset via single-pass clustering method.
- Only keep the events which associated with more than 20 posts and randomly split the posts on same event into support and query data.
- Training and testing set do not contain any common event.

Experiments

Baselines

- **Fine-tune models** (fine-tune by support set of testing set for fair comparison)
  - VQA(Visual Question Answering, ICCV'15): aims to answer the questions based on the given images and is used as a baseline for multimodal fake news.
  - att-RNN(MM'17): uses attention mechanism to fuse the textual, visual and social context features (remove social part).
  - EANN(KDD'18): captures shared features across different events of news to improve generalization ability.

Experiments

Baselines

- **Few-shot learning models**
  - CNP(Conditional neural process, ICML'18): combines neural network and gaussian process by using a small set of input-output pairs as context to output prediction.
  - ANP(Attentive neural process, ICLR'19): outputs prediction based on concatenation of learned distribution of context, context features and given input.
  - MAML(ICML'17): 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(arXiv'17): beside MAML, also learns step sizes and update direction during the training procedure.

Experiments

Performance Comparison: Twitter

Method	Twitter				Weibo			
	5-Shot		10-Shot		5-Shot		10-Shot	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
VQA	73.62 ± 1.83	76.69 ± 1.23	73.49 ± 2.61	74.69 ± 2.97	76.93 ± 0.71	75.88 ± 0.45	77.80 ± 1.43	76.36 ± 1.77
attRNN	63.04 ± 2.09	60.25 ± 4.63	63.14 ± 2.00	56.60 ± 5.25	76.07 ± 1.63	74.36 ± 2.96	78.09 ± 0.58	77.69 ± 0.35
EANN	70.01 ± 3.58	72.95 ± 2.86	70.56 ± 1.00	67.77 ± 0.80	76.43 ± 0.84	74.51 ± 0.56	77.49 ± 1.95	76.56 ± 1.28
CNP	71.42 ± 2.58	72.58 ± 3.57	72.47 ± 3.61	72.11 ± 5.74	77.47 ± 5.19	77.01 ± 4.66	78.81 ± 1.57	78.07 ± 1.98
ANP	77.08 ± 2.92	79.65 ± 3.81	74.25 ± 0.76	75.16 ± 1.27	77.85 ± 1.67	76.00 ± 3.61	76.52 ± 1.84	73.73 ± 2.78
MAML	82.24 ± 1.54	82.97 ± 1.76	85.22 ± 0.64	84.98 ± 1.70	74.68 ± 0.75	74.16 ± 0.33	75.87 ± 0.33	73.41 ± 0.86
Meta-SGD	74.13 ± 2.31	75.35 ± 2.56	74.63 ± 2.46	74.57 ± 2.74	71.73 ± 1.81	69.51 ± 2.28	73.34 ± 2.35	71.42 ± 2.80
MetaFEND (Improvement)	**86.45 ± 1.83 (↑5.12%)**	**86.21 ± 1.32 (↑3.91%)**	**88.79 ± 1.27 (↑4.19%)**	**88.66 ± 1.09 (↑4.33%)**	**81.28 ± 0.75 (↑4.41%)**	**80.19 ± 1.27 (↑4.13%)**	**82.92 ± 0.13 (↑5.22%)**	**82.37 ± 0.28 (↑5.51%)**

- In 5-shot setting, compared with CNP, ANP incorporates the attention mechanism and hence achieve more informative context information.
- In 10-shot setting, as the size of give support data increases, the soft-attention ANP unavoidably incorporates the irrelevant data points.

Experiments

Performance Comparison: Twitter

Method	Twitter				Weibo			
	5-Shot		10-Shot		5-Shot		10-Shot	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
VQA	73.62 ± 1.83	76.69 ± 1.23	73.49 ± 2.61	74.69 ± 2.97	76.93 ± 0.71	75.88 ± 0.45	77.80 ± 1.43	76.36 ± 1.77
attRNN	63.04 ± 2.09	60.25 ± 4.63	63.14 ± 2.00	56.60 ± 5.25	76.07 ± 1.63	74.36 ± 2.96	78.09 ± 0.58	77.69 ± 0.35
EANN	70.01 ± 3.58	72.95 ± 2.86	70.56 ± 1.00	67.77 ± 0.80	76.43 ± 0.84	74.51 ± 0.56	77.49 ± 1.95	76.56 ± 1.28
CNP	71.42 ± 2.58	72.58 ± 3.57	72.47 ± 3.61	72.11 ± 5.74	77.47 ± 5.19	77.01 ± 4.66	78.81 ± 1.57	78.07 ± 1.98
ANP	77.08 ± 2.92	79.65 ± 3.81	74.25 ± 0.76	75.16 ± 1.27	77.85 ± 1.67	76.00 ± 3.61	76.52 ± 1.84	73.73 ± 2.78
MAML	82.24 ± 1.54	82.97 ± 1.76	85.22 ± 0.64	84.98 ± 1.70	74.68 ± 0.75	74.16 ± 0.33	75.87 ± 0.33	73.41 ± 0.86
Meta-SGD	74.13 ± 2.31	75.35 ± 2.56	74.63 ± 2.46	74.57 ± 2.74	71.73 ± 1.81	69.51 ± 2.28	73.34 ± 2.35	71.42 ± 2.80
MetaFEND (Improvement)	**86.45 ± 1.83 (↑5.12%)**	**86.21 ± 1.32 (↑3.91%)**	**88.79 ± 1.27 (↑4.19%)**	**88.66 ± 1.09 (↑4.33%)**	**81.28 ± 0.75 (↑4.41%)**	**80.19 ± 1.27 (↑4.13%)**	**82.92 ± 0.13 (↑5.22%)**	**82.37 ± 0.28 (↑5.51%)**

- Due to the event heterogeneity, it's not easy for Meta-SGD to learn a shareable learning directions and step size across all events.
- Thus, performance of Meta-SGD is lower than MAML.

Experiments

Performance Comparison: Twitter

Method	Twitter				Weibo			
	5-Shot		10-Shot		5-Shot		10-Shot	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
VQA	73.62 ± 1.83	76.69 ± 1.23	73.49 ± 2.61	74.69 ± 2.97	76.93 ± 0.71	75.88 ± 0.45	77.80 ± 1.43	76.36 ± 1.77
attRNN	63.04 ± 2.09	60.25 ± 4.63	63.14 ± 2.00	56.60 ± 5.25	76.07 ± 1.63	74.36 ± 2.96	78.09 ± 0.58	77.69 ± 0.35
EANN	70.01 ± 3.58	72.95 ± 2.86	70.56 ± 1.00	67.77 ± 0.80	76.43 ± 0.84	74.51 ± 0.56	77.49 ± 1.95	76.56 ± 1.28
CNP	71.42 ± 2.58	72.58 ± 3.57	72.47 ± 3.61	72.11 ± 5.74	77.47 ± 5.19	77.01 ± 4.66	78.81 ± 1.57	78.07 ± 1.98
ANP	77.08 ± 2.92	79.65 ± 3.81	74.25 ± 0.76	75.16 ± 1.27	77.85 ± 1.67	76.00 ± 3.61	76.52 ± 1.84	73.73 ± 2.78
MAML	82.24 ± 1.54	82.97 ± 1.76	85.22 ± 0.64	84.98 ± 1.70	74.68 ± 0.75	74.16 ± 0.33	75.87 ± 0.33	73.41 ± 0.86
Meta-SGD	74.13 ± 2.31	75.35 ± 2.56	74.63 ± 2.46	74.57 ± 2.74	71.73 ± 1.81	69.51 ± 2.28	73.34 ± 2.35	71.42 ± 2.80
MetaFEND	**86.45 ± 1.83**	**86.21 ± 1.32**	**88.79 ± 1.27**	**88.66 ± 1.09**	**81.28 ± 0.75**	**80.19 ± 1.27**	**82.92 ± 0.13**	**82.37 ± 0.28**
(Improvement)	**(↑5.12%)**	**(↑3.91%)**	**(↑4.19%)**	**(↑4.33%)**	**(↑4.41%)**	**(↑4.13%)**	**(↑5.22%)**	**(↑5.51%)**

- MetaFEND inherits the advantage of MAML to learn a set of parameters which can rapidly learn to detect with small support set.
- MetaFEND can use support data as conditioning set explicitly to better capture the uncertainty of events.

Experiments

Performance Comparison: Weibo

Method	Twitter				Weibo			
	5-Shot		10-Shot		5-Shot		10-Shot	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
VQA	73.62 ± 1.83	76.69 ± 1.23	73.49 ± 2.61	74.69 ± 2.97	76.93 ± 0.71	75.88 ± 0.45	77.80 ± 1.43	76.36 ± 1.77
attRNN	63.04 ± 2.09	60.25 ± 4.63	63.14 ± 2.00	56.60 ± 5.25	76.07 ± 1.63	74.36 ± 2.96	78.09 ± 0.58	77.69 ± 0.35
EANN	70.01 ± 3.58	72.95 ± 2.86	70.56 ± 1.00	67.77 ± 0.80	76.43 ± 0.84	74.51 ± 0.56	77.49 ± 1.95	76.56 ± 1.28
CNP	71.42 ± 2.58	72.58 ± 3.57	72.47 ± 3.61	72.11 ± 5.74	77.47 ± 5.19	77.01 ± 4.66	78.81 ± 1.57	78.07 ± 1.98
ANP	77.08 ± 2.92	79.65 ± 3.81	74.25 ± 0.76	75.16 ± 1.27	77.85 ± 1.67	76.00 ± 3.61	76.52 ± 1.84	73.73 ± 2.78
MAML	82.24 ± 1.54	82.97 ± 1.76	85.22 ± 0.64	84.98 ± 1.70	74.68 ± 0.75	74.16 ± 0.33	75.87 ± 0.33	73.41 ± 0.86
Meta-SGD	74.13 ± 2.31	75.35 ± 2.56	74.63 ± 2.46	74.57 ± 2.74	71.73 ± 1.81	69.51 ± 2.28	73.34 ± 2.35	71.42 ± 2.80
MetaFEND (Improvement)	**86.45 ± 1.83 (↑5.12%)**	**86.21 ± 1.32 (↑3.91%)**	**88.79 ± 1.27 (↑4.19%)**	**88.66 ± 1.09 (↑4.33%)**	**81.28 ± 0.75 (↑4.41%)**	**80.19 ± 1.27 (↑4.13%)**	**82.92 ± 0.13 (↑5.22%)**	**82.37 ± 0.28 (↑5.51%)**

- On the Weibo dataset, most of posts are associated with different images.
- Thus, can evaluate the performance of models under the circumstance where support sets don't include direct clues with query set.

Experiments

Performance Comparison: Weibo

Method	Twitter				Weibo			
	5-Shot		10-Shot		5-Shot		10-Shot	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
VQA	73.62 ± 1.83	76.69 ± 1.23	73.49 ± 2.61	74.69 ± 2.97	76.93 ± 0.71	75.88 ± 0.45	77.80 ± 1.43	76.36 ± 1.77
attRNN	63.04 ± 2.09	60.25 ± 4.63	63.14 ± 2.00	56.60 ± 5.25	76.07 ± 1.63	74.36 ± 2.96	78.09 ± 0.58	77.69 ± 0.35
EANN	70.01 ± 3.58	72.95 ± 2.86	70.56 ± 1.00	67.77 ± 0.80	76.43 ± 0.84	74.51 ± 0.56	77.49 ± 1.95	76.56 ± 1.28
CNP	71.42 ± 2.58	72.58 ± 3.57	72.47 ± 3.61	72.11 ± 5.74	77.47 ± 5.19	77.01 ± 4.66	78.81 ± 1.57	78.07 ± 1.98
ANP	77.08 ± 2.92	79.65 ± 3.81	74.25 ± 0.76	75.16 ± 1.27	77.85 ± 1.67	76.00 ± 3.61	76.52 ± 1.84	73.73 ± 2.78
MAML	82.24 ± 1.54	82.97 ± 1.76	85.22 ± 0.64	84.98 ± 1.70	74.68 ± 0.75	74.16 ± 0.33	75.87 ± 0.33	73.41 ± 0.86
Meta-SGD	74.13 ± 2.31	75.35 ± 2.56	74.63 ± 2.46	74.57 ± 2.74	71.73 ± 1.81	69.51 ± 2.28	73.34 ± 2.35	71.42 ± 2.80
MetaFEND	**86.45 ± 1.83**	**86.21 ± 1.32**	**88.79 ± 1.27**	**88.66 ± 1.09**	**81.28 ± 0.75**	**80.19 ± 1.27**	**82.92 ± 0.13**	**82.37 ± 0.28**
(Improvement)	**(↑5.12%)**	**(↑3.91%)**	**(↑4.19%)**	**(↑4.33%)**	**(↑4.41%)**	**(↑4.13%)**	**(↑5.22%)**	**(↑5.51%)**

- ANP & CNP achieves better performance compared with gradient-based parameter meta-learning methods MAML & Meta-SGD.
- This's because parameter adaptation may not be effective when support set and query set do not share the same patterns.

Experiments

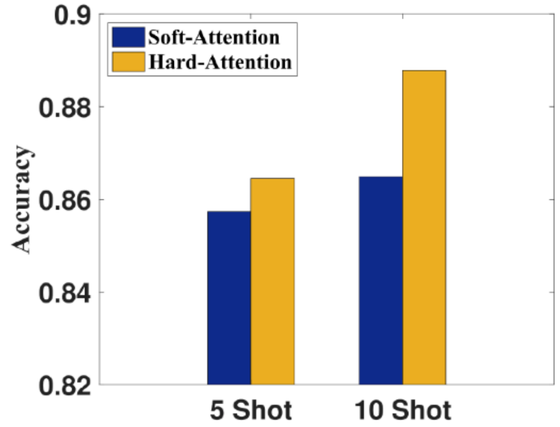
Performance Comparison: Weibo

Method	Twitter				Weibo			
	5-Shot		10-Shot		5-Shot		10-Shot	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
VQA	73.62 ± 1.83	76.69 ± 1.23	73.49 ± 2.61	74.69 ± 2.97	76.93 ± 0.71	75.88 ± 0.45	77.80 ± 1.43	76.36 ± 1.77
attRNN	63.04 ± 2.09	60.25 ± 4.63	63.14 ± 2.00	56.60 ± 5.25	76.07 ± 1.63	74.36 ± 2.96	78.09 ± 0.58	77.69 ± 0.35
EANN	70.01 ± 3.58	72.95 ± 2.86	70.56 ± 1.00	67.77 ± 0.80	76.43 ± 0.84	74.51 ± 0.56	77.49 ± 1.95	76.56 ± 1.28
CNP	71.42 ± 2.58	72.58 ± 3.57	72.47 ± 3.61	72.11 ± 5.74	77.47 ± 5.19	77.01 ± 4.66	78.81 ± 1.57	78.07 ± 1.98
ANP	77.08 ± 2.92	79.65 ± 3.81	74.25 ± 0.76	75.16 ± 1.27	77.85 ± 1.67	76.00 ± 3.61	76.52 ± 1.84	73.73 ± 2.78
MAML	82.24 ± 1.54	82.97 ± 1.76	85.22 ± 0.64	84.98 ± 1.70	74.68 ± 0.75	74.16 ± 0.33	75.87 ± 0.33	73.41 ± 0.86
Meta-SGD	74.13 ± 2.31	75.35 ± 2.56	74.63 ± 2.46	74.57 ± 2.74	71.73 ± 1.81	69.51 ± 2.28	73.34 ± 2.35	71.42 ± 2.80
MetaFEND	**86.45 ± 1.83**	**86.21 ± 1.32**	**88.79 ± 1.27**	**88.66 ± 1.09**	**81.28 ± 0.75**	**80.19 ± 1.27**	**82.92 ± 0.13**	**82.37 ± 0.28**
(Improvement)	**(↑5.12%)**	**(↑3.91%)**	**(↑4.19%)**	**(↑4.33%)**	**(↑4.41%)**	**(↑4.13%)**	**(↑5.22%)**	**(↑5.51%)**

- MetaFEND can learn a base parameter which can rapidly learn use a few example as reference information for fake news detection.
- MetaFEND enjoys the benefits of neural process and meta-learning model families.

Experiments

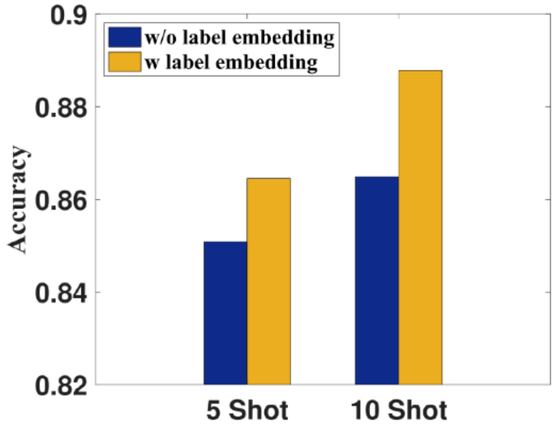
Ablation Study: Soft-Attention v.s. Hard-Attention



- Hard-attention in 5 or 10-shot settings are greater than those of Soft-attention.
- As the number of support set increases, hard-attention mechanism doesn't have limitation of soft-attention mechanism which unavoidably incorporates unrelated data points.
- Conclude that hard-attention can take effectively advantage of support set, and the superiority is more significant when enlarge size of support set.

Experiments

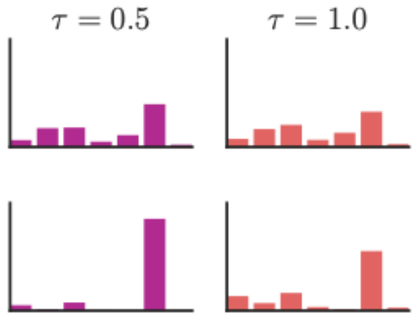
Ablation Study: w/o v.s. w/ Label Embedding



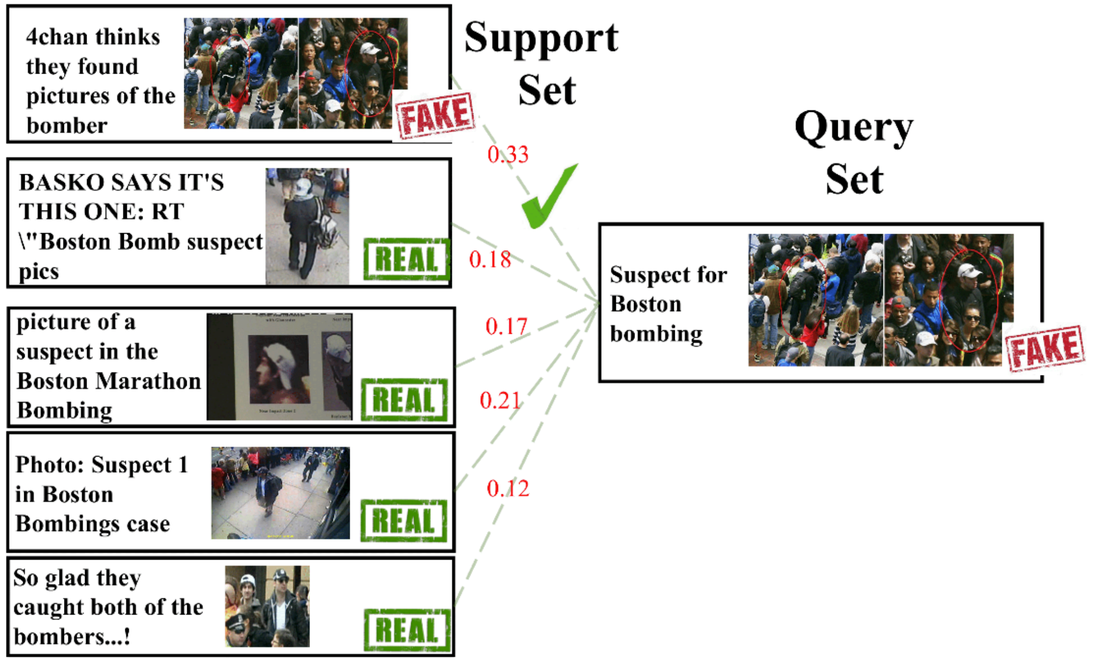
- Reduced model by replacing label embedding with label value 0 or 1.
- Change the multiplication between output with label embedding to a binary-class fully connected layer to directly output the probabilities.
- Observe that the accuracy score of w/ label embedding is greater than w/o label embedding in two shot settings.
  - Demonstrating the effectiveness of label embedding.

Case Study

Effective of Hard-Attention



https://arxiv.org/abs/1611.01144



- 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.
- Due to imbalanced class condition in the support set, it's difficult for Soft-Attention to provide correct prediction for news of interest in the query set.
- Hard-Attention can achieve correct result by focusing on the most similar sample in the support set.

Conclusions

- Study the problem of fake news detection on emergent events.
- Propose a novel fake news detection framework MetaFEND, which can rapidly learn to detect fake news for emergent events with a few labeled examples.
- MetaFEND can enjoy the benefits of meta-learning and neural process model families without suffering their own limitations.

Comments
of MetaFEND

- Fusion the meta-learning and neural process and overcome their limitations (event heterogeneity, under-fitting).
- Use hard-attention (Gumbel SoftMax) to trim irrelevant data and select the most relative samples.
- Use label embedding to consider the categorical characteristics.
- Need used a set of posts which associated with specific event in real application.
  - When real news in the wrong event set. (Episodic classification)