# Hierarchical Multi-modal Contextual Attention Network for Fake News Detection

ADVISOR: Jia-Ling Koh
PRESENTER: Xiao-Yuan Hung
SOURCE: SIGIR'21
DATE: 2022/09/27

# Outline

# Fake news detection

Input
(text)
[ image/user/source ]

⇨

Approaches

⇨

Output
(label)
[True / False]

# Approaches

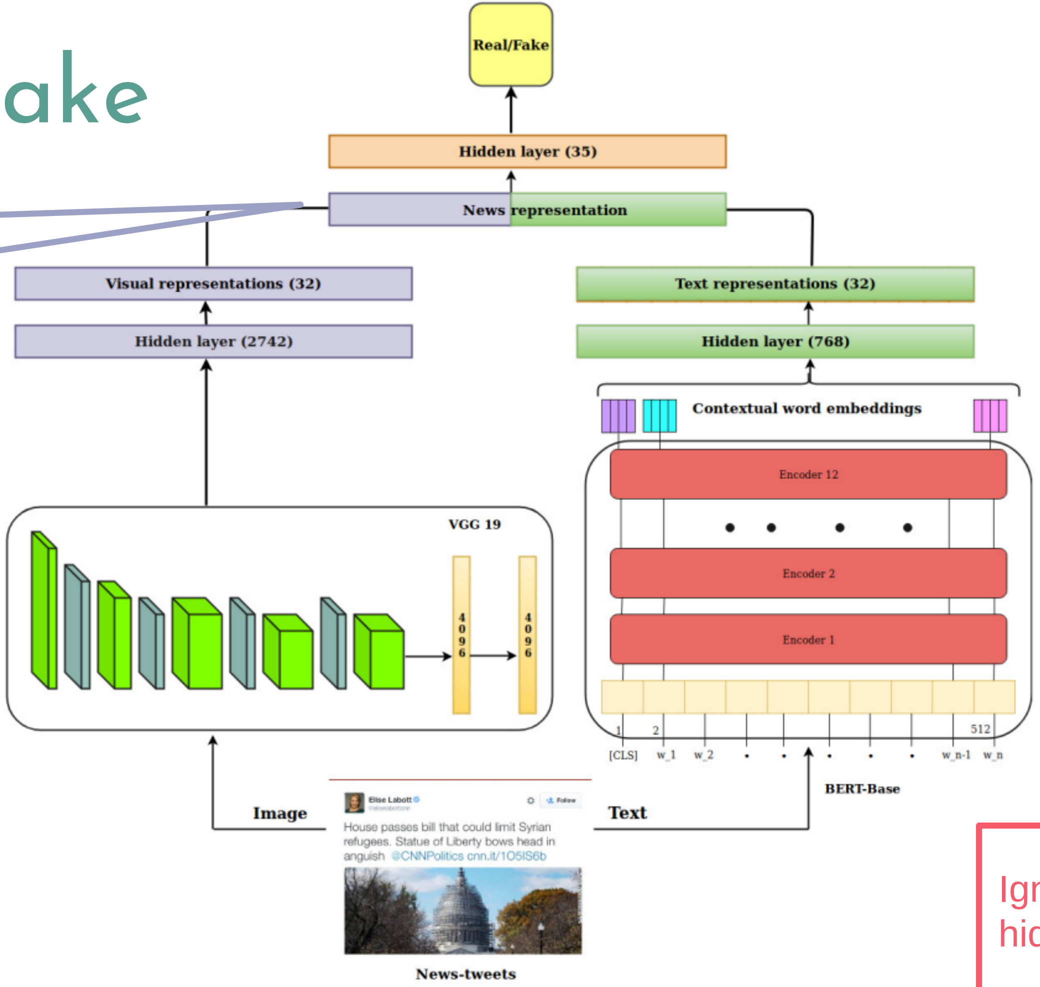
- **Text**
  - Traditional learning methods (hand-crafted features)
    - SVM
    - Decision Tree
  - Deep learning approaches
    - RNN
    - CNN
- **Multi-modal**
  - Text , Image
    - MVAE
    - SpotFake
  - Text,  News publishers, Users
    - SAME

# Problem

- **Components previously methods employ to capture multi-modal context are too simple**
- **Only utilize the output of the last layers of these hierarchical models, while ignoring the intermediate hidden states, which also capture rich linguistic information**

# SpotFake

Too simple




Ignoring the intermediate hidden states

Figure 4: A schematic diagram of the proposed SpotFake model. Value in () indicates number of neurons in a layer.

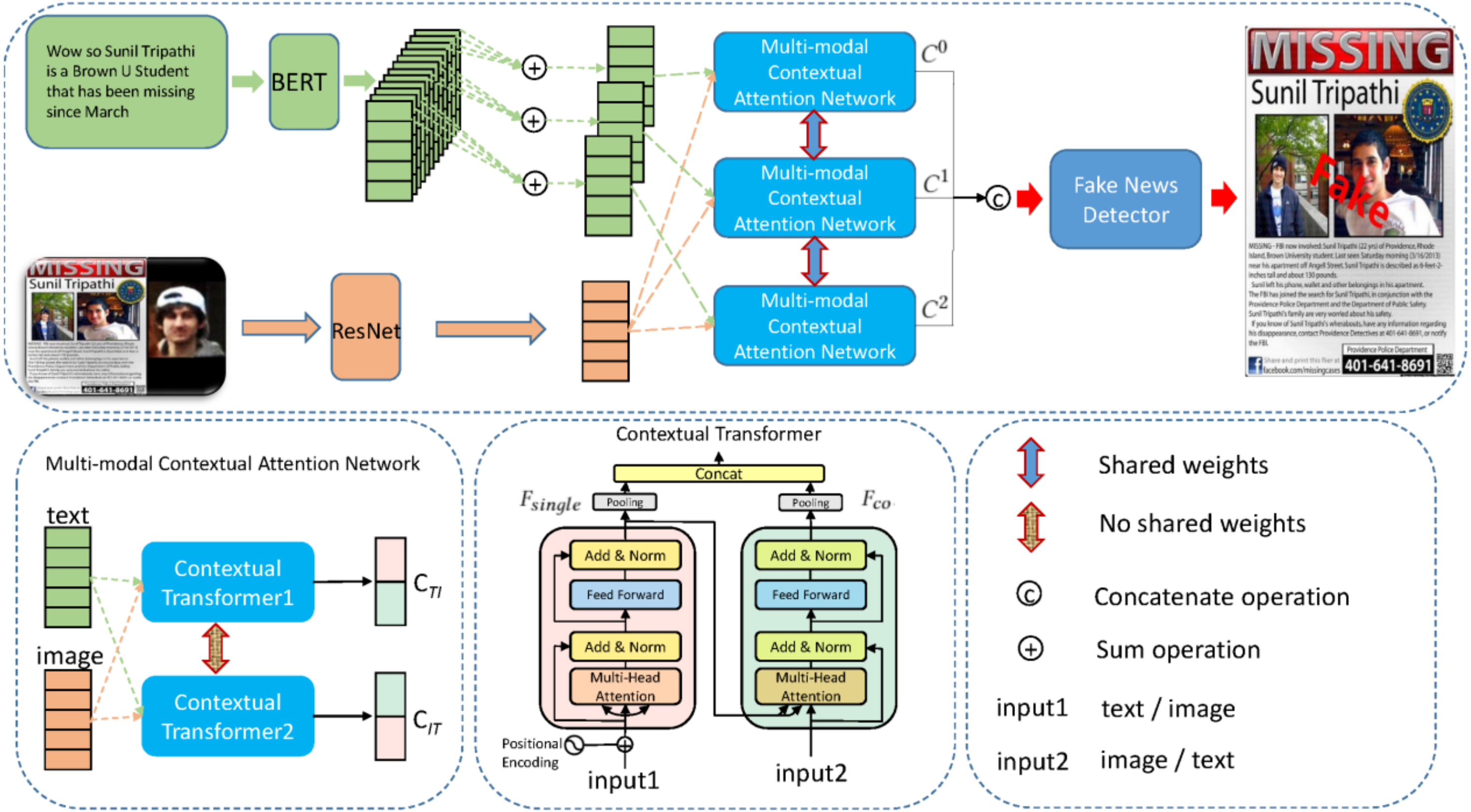
# Challenges

- **Challenge 1**
  - How to fully utilize the multi-modal context information and extract high-order complementary information from it to enhance the performance of fake news detection?
  - Propose a multi-modal contextual attention network to model the multi-modal context for each news posts
- **Challenge 2**
  - How to explore and capture the hierarchical semantics of text information to learn a better representation of multi-modal news?
  - Design a hierarchical encoding network to capture the rich hierarchical semantics for fake news detection

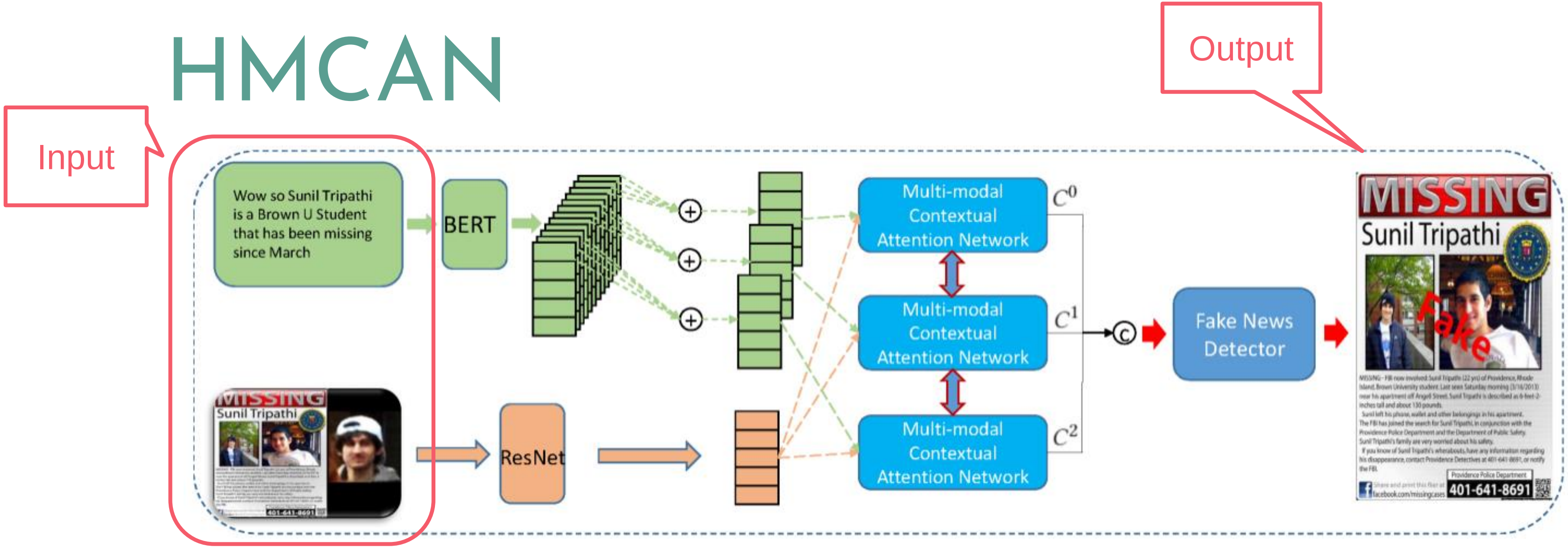
# Outline

1. Introduction
2. Method
3. Experiment
4. Conclusion

# HMCAN

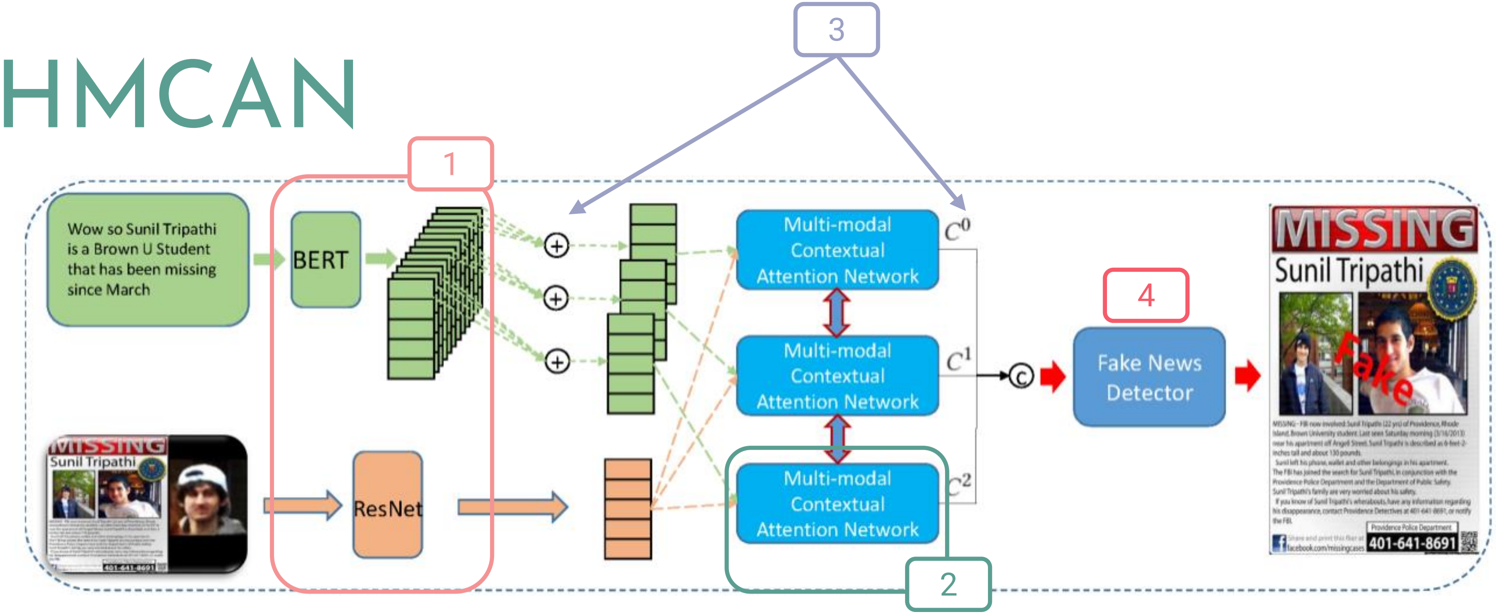

Wow so Sunil Tripathi is a Brown U Student that has been missing since March
BERT
ResNet
MISSING
Sunil Tripathi
Multi-modal Contextual Attention Network
$C^0$
$C^1$
$C^2$
Fake News Detector
Fake
401-641-8691
Multi-modal Contextual Attention Network
text
image
Contextual Transformer1
Contextual Transformer2
$C_{TI}$
$C_{IT}$
Contextual Transformer
Concat
$F_{single}$
$F_{co}$
Pooling
Add & Norm
Feed Forward
Multi-Head Attention
Positional Encoding
input1
input2
Shared weights
No shared weights
Concatenate operation
Sum operation
input1 text / image
input2 image / text

# HMCAN



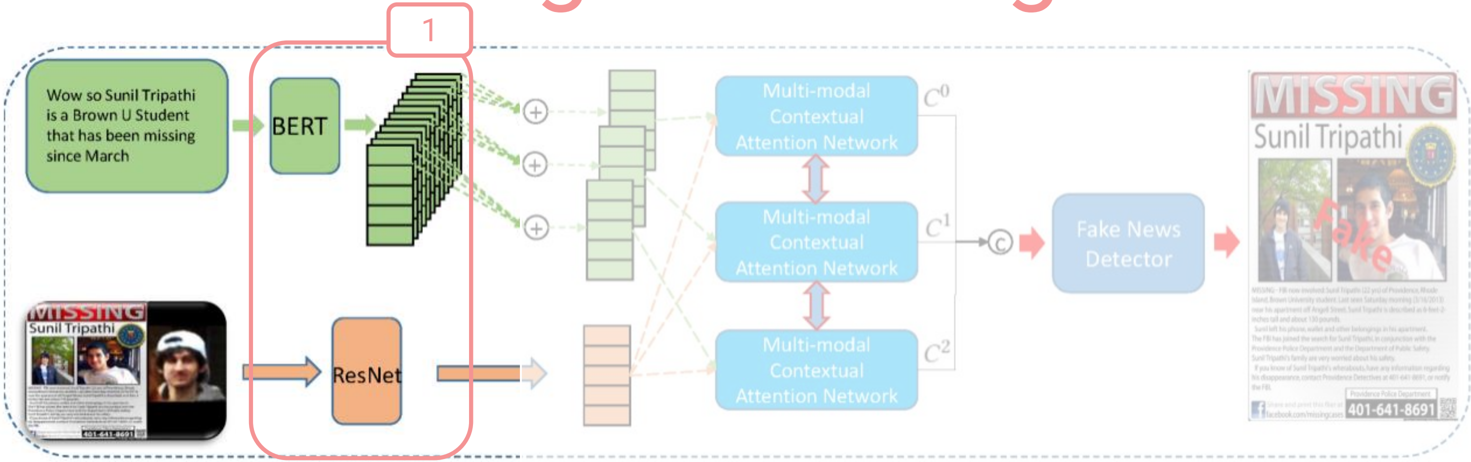

- **Input**
  - Multi-modal post $P$ from social media consisting of text messages and corresponding images
- **Output**
  - Label of the post $Y$ = {0, 1}
  - $Y$ = 0 is real news
  - $Y$ = 1 is fake news

# HMCAN




1. Text and Image Encoding Network
2. Multi-modal Contextual Attention Network
3. Hierarchical Encoding Network
4. Fake News Detector

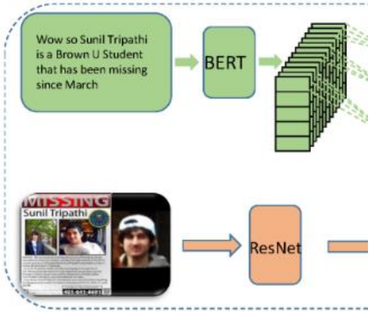
# Text and Image Encoding Network



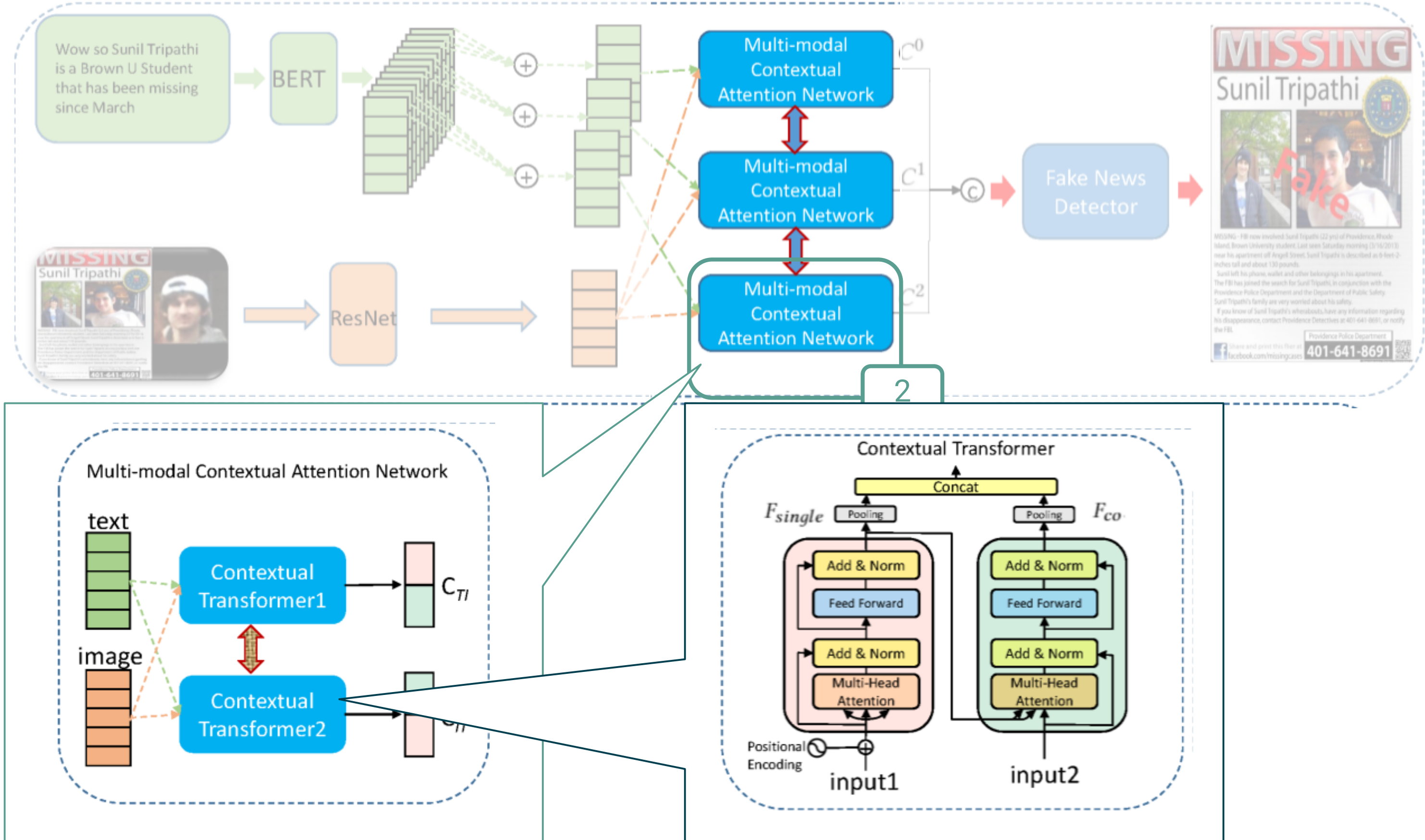

- Text Encoding Network
- Image Encoding Network

# Text and Image Encoding Network

- **Multi-modal post $P$ = {$W$ , $R$}**
- **Text Encoding Network**
  - Input : $W$ as a sequence of words $W = \{w_1, w_2, \cdots, w_m\}$
  - Output : word representation $S = \{s_1, \cdots, s_m\}$
  - Model : pre-trained BERT
- **Image Encoding Network**
  - Input : visual content *R*
  - Output : a set of region features $O = \{o_1, \cdots, o_n\}$
  - Model : ResNet50

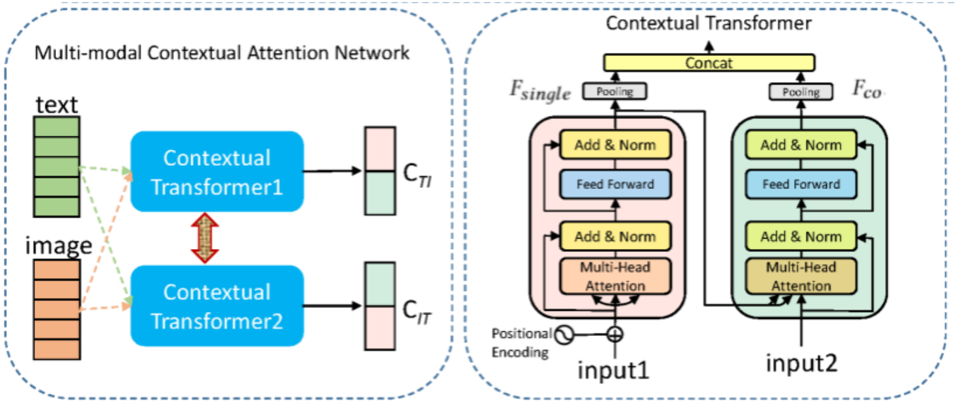


# Multi-modal Contextual Attention Network



# Multi-modal Contextual Attention Network

- **Purpose: Extract high-order complementary information**
- **Input1: Text/Image**
- **Input2: Image/Text**



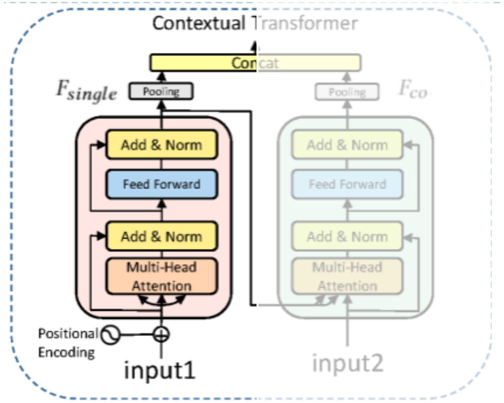
# Self-attention network $F_{single}$

- **Purpose: learn the representation of input1**$(text)$
- **Intra-modality affinity matrix** $As$
- **Representation of text** $Hs$

$$A_s = softmax\left(\frac{FC_s^Q(input1) \cdot FC_s^K(input1)^\intercal}{\sqrt{d}}\right) \quad (3)$$

$$H_s' = layer\_norm(input1 + A_s \cdot FC_s^V(input1)) \quad (4)$$

$$H_s = layer\_norm(H_s' + FC_s^{ff}(H_s')) \quad (5)$$




* FC = full-connected layers

* $FC^{ff}$ = two-layer full-connected network

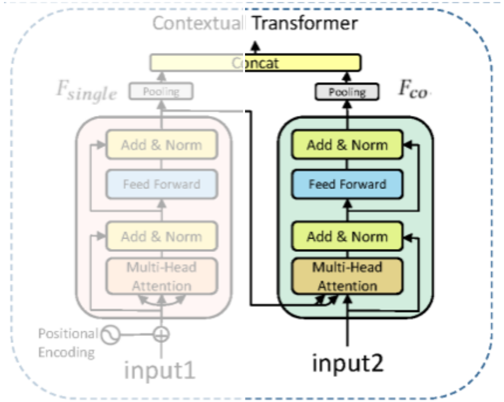
# Inter-modality attention network $F_{co}$

- **Purpose:**
  - extract information that is relevant to the image from the learned text representation, which can complement the visual information
- **Inter-modality affinity matrix** $Aco$
- **Multi-modal context-aware text representation** $Hco$

$$A_{co} = softmax\left(\frac{FC_{co}^{Q}(input2) \cdot FC_{co}^{K}(H_s)^{\top}}{\sqrt{d}}\right) \quad (6)$$

$$H'_{co} = layer\_norm(input2 + A_{co} \cdot FC_{co}^{V}(H_s)) \quad (7)$$

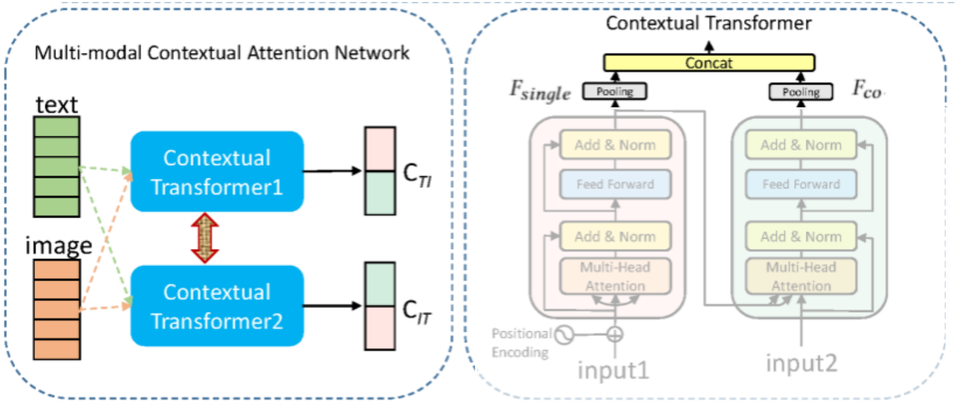
$$H_{co} = layer\_norm(H'_{co} + FC_{co}^{ff}(H'_{co})) \quad (8)$$



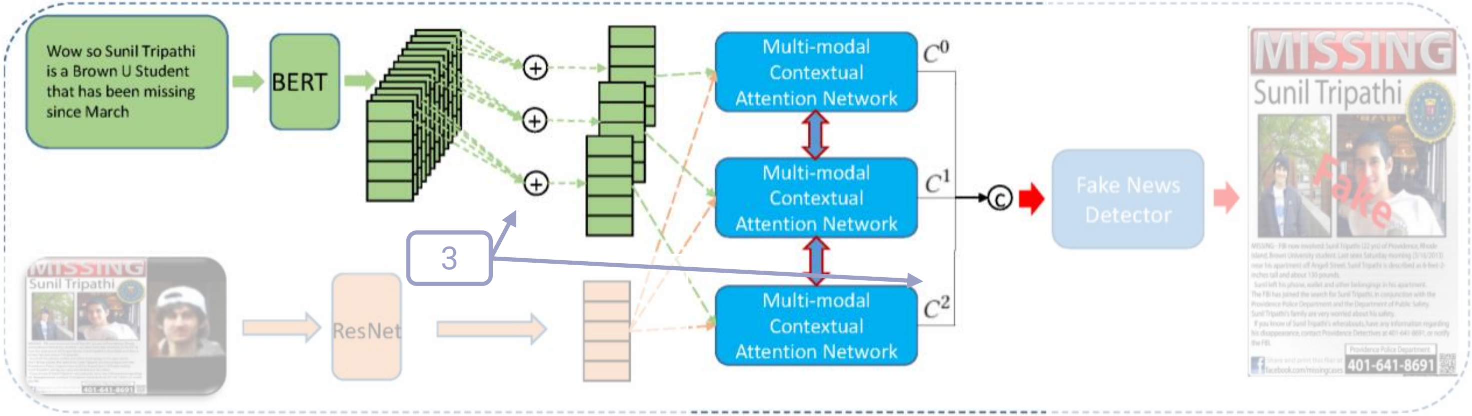
# Contextual Transformer 1 and 2

- **Pooled into two feature vectors, and concatenated into a feature vector**
- **Contextual Transformer1 output** $C_{TI}$
- **Contextual Transformer2 output** $C_{IT}$

$$C = \alpha C_{TI} + \beta C_{IT}, \text{ where } \alpha + \beta = 1.$$



# Hierarchical Encoding Network




- **BERT can provide hierarchical semantics for text**
- **Through different multi-modal contextual attention network units, get different $C$ values**

# Hierarchical Encoding Network

- **Purpose: to capture the rich hierarchical semantics**
- **Group 12 layer outputs into $g$ groups ($g$ = 3)**
- $fB\,(W\,)_{j,\mathrm{i}}$
  - $j$-th layer BERT for the $i$-th word in text W
- $s_{\mathrm{i}}^{\mathrm{k}}$
  - Initial representation of the $k$-th group of the $i$-th word
- $d_{\mathrm{W}}$
  - Dimension of the word embedding

$$\mathbf{s}_{\mathbf{i}}^{\mathbf{0}} = \sum_{j=1}^{4} f_B(W)_{j,i},\ \mathbf{s}_{\mathbf{i}}^{\mathbf{1}} = \sum_{j=5}^{8} f_B(W)_{j,i},\ \mathbf{s}_{\mathbf{i}}^{\mathbf{2}} = \sum_{j=9}^{12} f_B(W)_{j,i} \qquad (9)$$

$$C = concat(C^0, C^1, C^2) \qquad (10)$$

# Fake News Detector

- **Input**
  - Multi-modal representation $C$
- **Output**
  - label Y
- $P_n$ hat
  - Predicted probabilities of the $n$-th post
- $C_n$
  - feature representation of the $n$-th post
- **Activation function**
  - softmax


$$\hat{P}_n = \sigma(W_f C_n + b) \tag{11}$$

# Loss function

- **Cross entropy**

$$\mathcal{L}(\Theta) = \sum_{n=1}^{N} -[\mathrm{Y}_n \log(\hat{P}_n) + (1 - Y_n) \log(1 - \hat{P}_n)] \qquad (12)$$

# Outline

# Dataset

- **WEIBO**
  - each post contains three elements (i.e., id, text and image)
- **TWITTER**
  - textual information, visual information and social context information
- **PHEME**
  - 5 breaking news, each containing a set of posts

| News | WEIBO | TWITTER | PHEME |
|---|---|---|---|
| # of Fake News | 4749 | 7898 | 1972 |
| # of Real News | 4779 | 6026 | 3830 |
| # of Images | 9528 | 514 | 3670 |

| Dataset | Methods | Accuracy | Fake news | | | Real news | | |
|---|---|---|---|---|---|---|---|---|
| | | | Precision | Recall | F1 | Precision | Recall | F1 |
| WEIBO | # SVM-TS | 0.640 | 0.741 | 0.573 | 0.646 | 0.651 | 0.798 | 0.711 |
| | # GRU | 0.702 | 0.671 | 0.794 | 0.727 | 0.747 | 0.609 | 0.671 |
| | # CNN | 0.740 | 0.736 | 0.756 | 0.744 | 0.747 | 0.723 | 0.735 |
| | SAFE | 0.763 | 0.833 | 0.659 | 0.736 | 0.717 | 0.868 | 0.785 |
| | att_RNN | 0.772 | 0.854 | 0.656 | 0.742 | 0.720 | 0.889 | 0.795 |
| | EANN | 0.782 | 0.827 | 0.697 | 0.756 | 0.752 | 0.863 | 0.804 |
| | # TextGCN | 0.787 | 0.975 | 0.573 | 0.727 | 0.712 | 0.985 | 0.827 |
| | MVAE | 0.824 | 0.854 | 0.769 | 0.809 | 0.802 | 0.875 | 0.837 |
| | SpotFake | 0.869 | 0.877 | 0.859 | 0.868 | 0.861 | 0.879 | 0.870 |
| | SpotFake* | **0.892** | 0.902 | 0.964 | **0.932** | 0.847 | 0.656 | 0.739 |
| | SpotFake+ | 0.870 | 0.887 | 0.849 | 0.868 | 0.855 | 0.892 | 0.873 |
| | *HMCAN* | 0.885 | 0.920 | 0.845 | 0.881 | 0.856 | 0.926 | **0.890** |

# Single-modal Models
SpotFake : reproduction by author

| Dataset | Methods | Accuracy | Fake news | | | Real news | | |
|---|---|---|---|---|---|---|---|---|
| | | | Precision | Recall | F1 | Precision | Recall | F1 |
| TWITTER | # SVM-TS | 0.529 | 0.488 | 0.497 | 0.496 | 0.565 | 0.556 | 0.561 |
| | # GRU | 0.634 | 0.581 | 0.812 | 0.677 | 0.758 | 0.502 | 0.604 |
| | # CNN | 0.549 | 0.508 | 0.597 | 0.549 | 0.598 | 0.509 | 0.550 |
| | SAFE | 0.766 | 0.777 | 0.795 | 0.786 | 0.752 | 0.731 | 0.742 |
| | att_RNN | 0.664 | 0.749 | 0.615 | 0.676 | 0.589 | 0.728 | 0.651 |
| | EANN | 0.648 | 0.810 | 0.498 | 0.617 | 0.584 | 0.759 | 0.660 |
| | # TextGCN | 0.703 | 0.808 | 0.365 | 0.503 | 0.680 | 0.939 | 0.779 |
| | MVAE | 0.745 | 0.801 | 0.719 | 0.758 | 0.689 | 0.777 | 0.730 |
| | SpotFake | 0.771 | 0.784 | 0.744 | 0.764 | 0.769 | 0.807 | 0.787 |
| | SpotFake* | 0.777 | 0.751 | 0.900 | 0.820 | 0.832 | 0.606 | 0.701 |
| | SpotFake+ | 0.790 | 0.793 | 0.827 | 0.810 | 0.786 | 0.747 | 0.766 |
| | *HMCAN* | **0.897** | 0.971 | 0.801 | **0.878** | 0.853 | 0.979 | **0.912** |

| Dataset | Methods | Accuracy | Fake news | | | Real news | | |
|---|---|---|---|---|---|---|---|---|
| | | | Precision | Recall | F1 | Precision | Recall | F1 |
| PHEME | # SVM-TS | 0.639 | 0.546 | 0.576 | 0.560 | 0.729 | 0.705 | 0.717 |
| | # GRU | 0.832 | 0.782 | 0.712 | 0.745 | 0.855 | 0.896 | 0.865 |
| | # CNN | 0.779 | 0.732 | 0.606 | 0.663 | 0.799 | 0.875 | 0.835 |
| | SAFE | 0.811 | 0.827 | 0.559 | 0.667 | 0.806 | 0.940 | 0.866 |
| | att_RNN | 0.850 | 0.791 | 0.749 | 0.770 | 0.876 | 0.899 | 0.888 |
| | EANN | 0.681 | 0.685 | 0.664 | 0.694 | 0.701 | 0.750 | 0.747 |
| | # TextGCN | 0.828 | 0.775 | 0.735 | 0.737 | 0.827 | 0.828 | 0.828 |
| | MVAE | 0.852 | 0.806 | 0.719 | 0.760 | 0.871 | 0.917 | 0.893 |
| | SpotFake | 0.823 | 0.743 | 0.745 | 0.744 | 0.864 | 0.863 | 0.863 |
| | SpotFake+ | 0.800 | 0.730 | 0.668 | 0.697 | 0.832 | 0.869 | 0.850 |
| | *HMCAN* | **0.881** | 0.830 | 0.838 | **0.834** | 0.910 | 0.905 | **0.907** |

# Confusion matrix

| | | Predicted Failure | |
|---|---|---|---|
| | | True | False |
| Actual Failure | True | TP := True Positive | FN := False Negative |
| | False | FP := False Positive | TN := True Negative |

*Recall* := TP / (TP+FN)

*Precision* := TP / (TP+FP)

*F1-Score* := 2*Precision*Recall / (Precision+Recall)

https://www.researchgate.net/figure/Confusion-matrix-illustrating-the-calculation-of-precision-recall-and-F1-score_fig1_327982563

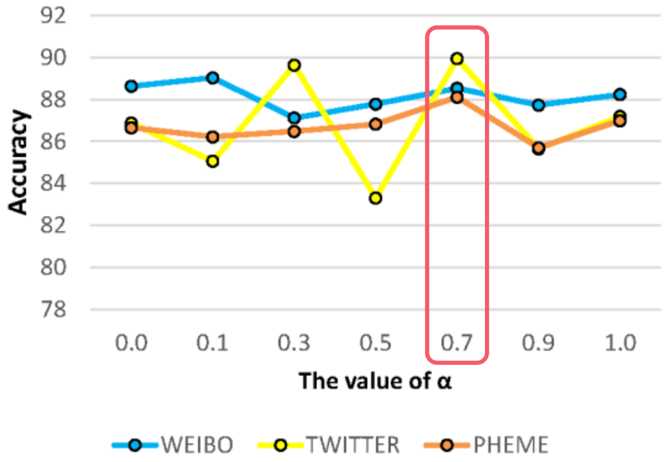
# Ablation

**HMCAN-$V$ : remove visual information**
**HMCAN-$C$ : remove multi-modal contextual attention network removed**
**HMCAN-$H$ : remove hierarchical information of words**

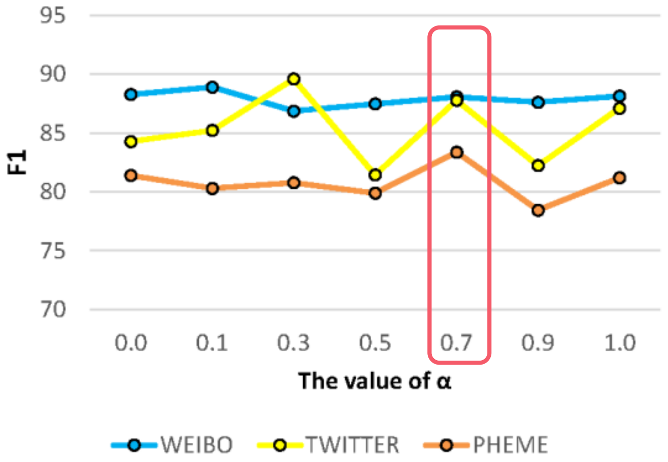
| Dataset | Methods | Accuracy | Fake news | | | Real news | | |
|---|---|---|---|---|---|---|---|---|
| | | | Precision | Recall | F1 | Precision | Recall | F1 |
| WEIBO | HMCAN¬$V$ | 0.809 | 0.832 | 0.774 | 0.802 | 0.788 | 0.843 | 0.815 |
| | HMCAN¬$C$ | 0.872 | 0.902 | 0.836 | 0.868 | 0.847 | 0.909 | 0.877 |
| | HMCAN¬$H$ | 0.877 | 0.871 | 0.885 | 0.878 | 0.883 | 0.869 | 0.876 |
| | HMCAN | **0.885** | 0.920 | 0.845 | **0.881** | 0.856 | 0.926 | **0.890** |
| TWITTER | HMCAN¬$V$ | 0.755 | 0.828 | 0.590 | 0.689 | 0.719 | 0.896 | 0.798 |
| | HMCAN¬$C$ | 0.790 | 0.886 | 0.622 | 0.731 | 0.743 | 0.932 | 0.827 |
| | HMCAN¬$H$ | 0.879 | 0.884 | 0.849 | 0.866 | 0.875 | 0.906 | 0.890 |
| | HMCAN | **0.897** | 0.971 | 0.801 | **0.878** | 0.853 | 0.979 | **0.912** |
| PHEME | HMCAN¬$V$ | 0.854 | 0.814 | 0.763 | 0.788 | 0.873 | 0.904 | 0.888 |
| | HMCAN¬$C$ | 0.858 | 0.788 | 0.821 | 0.804 | 0.899 | 0.878 | 0.888 |
| | HMCAN¬$H$ | 0.871 | 0.808 | 0.828 | 0.818 | 0.906 | 0.894 | 0.900 |
| | HMCAN | **0.881** | 0.830 | 0.838 | **0.834** | 0.910 | 0.905 | **0.907** |

# Impact of the value of α

$$C = \alpha C_{TI} + \beta C_{IT}, \text{ where } \alpha + \beta = 1.$$



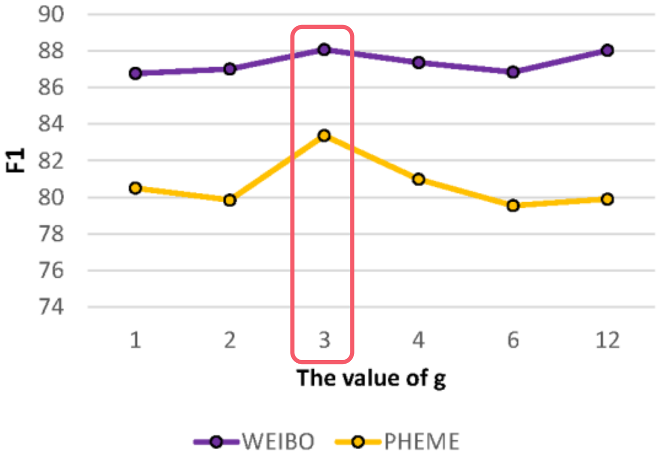

(a) Accuracy




(b) F1 score of fake news

# Impact of the number of group $g$


(a) Accuracy

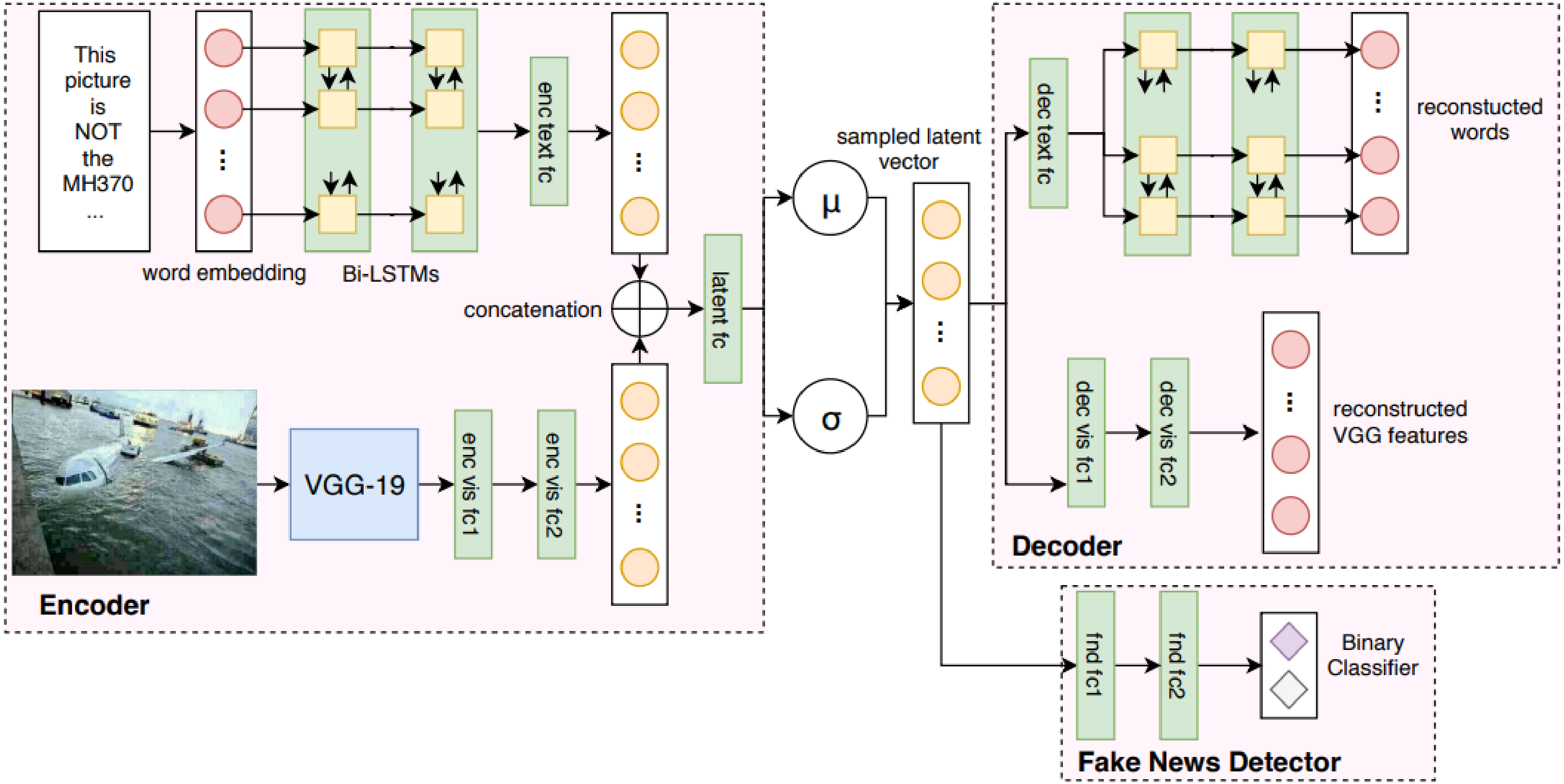



(b) F1 score of fake news

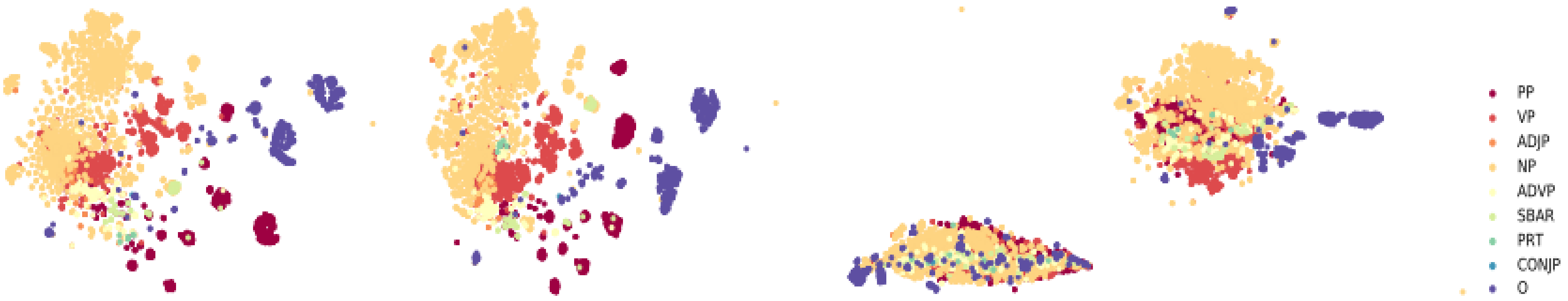
# Conclusion

- Propose a novel hierarchical multi-modal contextual attention network (HMCAN) for fake news detection task
- A multi-modal contextual attention network is proposed to fuse both inter-modality and intra-modality relationships
- Design a hierarchical encoding network to capture the rich hierarchical semantics

# (Add Info)MVAE



| Layer | SentLen (Surface) | WC (Surface) | TreeDepth (Syntactic) | TopConst (Syntactic) | BShift (Syntactic) | Tense (Semantic) | SubjNum (Semantic) | ObjNum (Semantic) | SOMO (Semantic) | CoordInv (Semantic) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 93.9 (2.0) | 24.9 (24.8) | 35.9 (6.1) | 63.6 (9.0) | 50.3 (0.3) | 82.2 (18.4) | 77.6 (10.2) | 76.7 (26.3) | 49.9 (-0.1) | 53.9 (3.9) |
| 2 | 95.9 (3.4) | 65.0 (64.8) | 40.6 (11.3) | 71.3 (16.1) | 55.8 (5.8) | 85.9 (23.5) | 82.5 (15.3) | 80.6 (17.1) | 53.8 (4.4) | 58.5 (8.5) |
| 3 | **96.2 (3.9)** | 66.5 (66.0) | 39.7 (10.4) | 71.5 (18.5) | 64.9 (14.9) | 86.6 (23.8) | 82.0 (14.6) | 80.3 (16.6) | 55.8 (5.9) | 59.3 (9.3) |
| 4 | 94.2 (2.3) | **69.8 (69.6)** | 39.4 (10.8) | 71.3 (18.3) | 74.4 (24.5) | 87.6 (25.2) | 81.9 (15.0) | 81.4 (19.1) | 59.0 (8.5) | 58.1 (8.1) |
| 5 | 92.0 (0.5) | 69.2 (69.0) | 40.6 (11.8) | 81.3 (30.8) | 81.4 (31.4) | 89.5 (26.7) | 85.8 (19.4) | 81.2 (18.6) | 60.2 (10.3) | 64.1 (14.1) |
| 6 | 88.4 (-3.0) | 63.5 (63.4) | **41.3 (13.0)** | 83.3 (36.6) | 82.9 (32.9) | 89.8 (27.6) | **88.1 (21.9)** | 82.0 (20.1) | 60.7 (10.2) | 71.1 (21.2) |
| 7 | 83.7 (-7.7) | 56.9 (56.7) | 40.1 (12.0) | **84.1 (39.5)** | 83.0 (32.9) | 89.9 (27.5) | 87.4 (22.2) | **82.2 (21.1)** | 61.6 (11.7) | 74.8 (24.9) |
| 8 | 82.9 (-8.1) | 51.1 (51.0) | 39.2 (10.3) | 84.0 (39.5) | 83.9 (33.9) | 89.9 (27.6) | 87.5 (22.2) | 81.2 (19.7) | 62.1 (12.2) | 76.4 (26.4) |
| 9 | 80.1 (-11.1) | 47.9 (47.8) | 38.5 (10.8) | 83.1 (39.8) | **87.0 (37.1)** | **90.0 (28.0)** | 87.6 (22.9) | 81.8 (20.5) | 63.4 (13.4) | **78.7 (28.9)** |
| 10 | 77.0 (-14.0) | 43.4 (43.2) | 38.1 (9.9) | 81.7 (39.8) | 86.7 (36.7) | 89.7 (27.6) | 87.1 (22.6) | 80.5 (19.9) | 63.3 (12.7) | 78.4 (28.1) |
| 11 | 73.9 (-17.0) | 42.8 (42.7) | 36.3 (7.9) | 80.3 (39.1) | 86.8 (36.8) | 89.9 (27.8) | 85.7 (21.9) | 78.9 (18.6) | 64.4 (14.5) | 77.6 (27.9) |
| 12 | 69.5 (-21.4) | 49.1 (49.0) | 34.7 (6.9) | 76.5 (37.2) | 86.4 (36.4) | 89.5 (27.7) | 84.0 (20.2) | 78.7 (18.4) | **65.2 (15.3)** | 74.9 (25.4) |




(a) Layer 1 (b) Layer 2 (c) Layer 11 (d) Layer 12

# (Add Info) TextGCN