Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News

Reuben Tan
Boston University
rxtan@bu.edu

Bryan A. Plummer
Boston University
bplum@bu.edu

Kate Saenko
Boston University
MIT-IBM Watson AI Lab
saenko@bu.edu

EMNLP'20

211116 Chia-Chun Ho

Outline

Introduction

Related Work

Proposed Model

Experiments

Summary and Defense Directions

Comments

Introduction

Fake news generated by generative models

- Rapid progression of generative models in both computer vision and natural language processing
  - has led to the increasing likelihood of realistic-looking news articles generated by AI.
- By manipulating such technology, adversaries would be able to disseminate large amounts of online disinformation rapidly.
- It ignores the fact that news articles are often accompanied by images with captions.

Introduction
Against neural fake news

- In this paper, present the first line of defense against neural fake news with image and captions.
- To the best of authors' knowledge, first to address this challenging and realistic problem.
- Premised on the assumption that the adversarial text generator is unknown beforehand, propose to evaluate articles based on the semantic consistency between the linguistic and visual components.

Introduction

Visual-semantic consistency

- While SOTA approaches in bidirectional image-sentence retrieval have leveraged visual-semantic consistency to great success on standard datasets such as MSCOCO and Flickr30K.
- They are not able to reason effectively about objects in an image and named entities present in the caption or article body.
- This's due to discrepancies in the distribution of these datasets, as captions in the standard datasets usually contain general terms.
    - Like woman or dog as opposed to named entities such as Mrs Betram and a Golden Retriever, which are commonly contained in news article captions.

Introduction

Visual-semantic consistency

- Moreover, images are often not directly related to the articles they are associated with.
  - For example, the article contains mentions of the British Prime Minister.
  - Yet, it only contains an image of the United Kingdom Flag.

nytimes.com

What's Next for Britons after Brexit?

August 28, 2019 - Anne Smith

In September, voters overwhelming rejected a plan from Prime Minister Theresa May's team for the United Kingdom to stay in the European Union. On March 29, Britain will officially exit the union after years of campaigning and serious negotiations. The EU's chief Brexit negotiator, Michel Barnier, has warned that there could be no future trade deals with the United Kingdom if there is a "no deal." The transition period will allow the United Kingdom and the European Union to work out a new plan for their relationship. But we may not know ...

Parliament was scheduled to reconvene on Oct 9, but Mr. Johnson said he planned to extend its break.

Introduction
DIDAN

- A simple yet surprisingly effective approach which exploits possible semantic inconsistencies between the text and image/captions to detect machine-generated articles.
- For example, notice that the article and caption actually mention different Prime Ministers.

Introduction

DIDAN

- Besides evaluating the semantic relevance of images and captions to the article,
  - DIDAN also exploits the co-occurrences of named entities in the article and captions to determine the authenticity score.
  - The score can be thought of as the probability that an article is human-generated.
- Adopt a learning paradigm commonly used in image-sentence retrieval where models are trained to reason about dissimilarities between images and non-matching captions.

Introduction

Neural News dataset

- Construct dataset which contains both human and machine-generated articles.
- These articles contain a title, the main body as well as images and captions.
- The human-generated articles are sourced from the GoodNews dataset.
- Using the same titles and main article bodies as context, use GROVER to generate articles.
- Instead of using GAN-generated images which are easy to detect even without exposure to them during training time, consider the much harder setting where the articles are completed with the original images.

Introduction

Contribution

- Introduce the novel and challenging task of defending against full news article containing image-caption pairs.
    - First paper to address both the visual and linguistic aspects of defending against neural fake news.
- Introduce the NeuralNews dataset that contains both human and machine-generated articles with images and captions.
- Propose DIDAN, an effective named entity-based model that serves as a good baseline for defending against neural fake news.

Related Work
of fake news detection

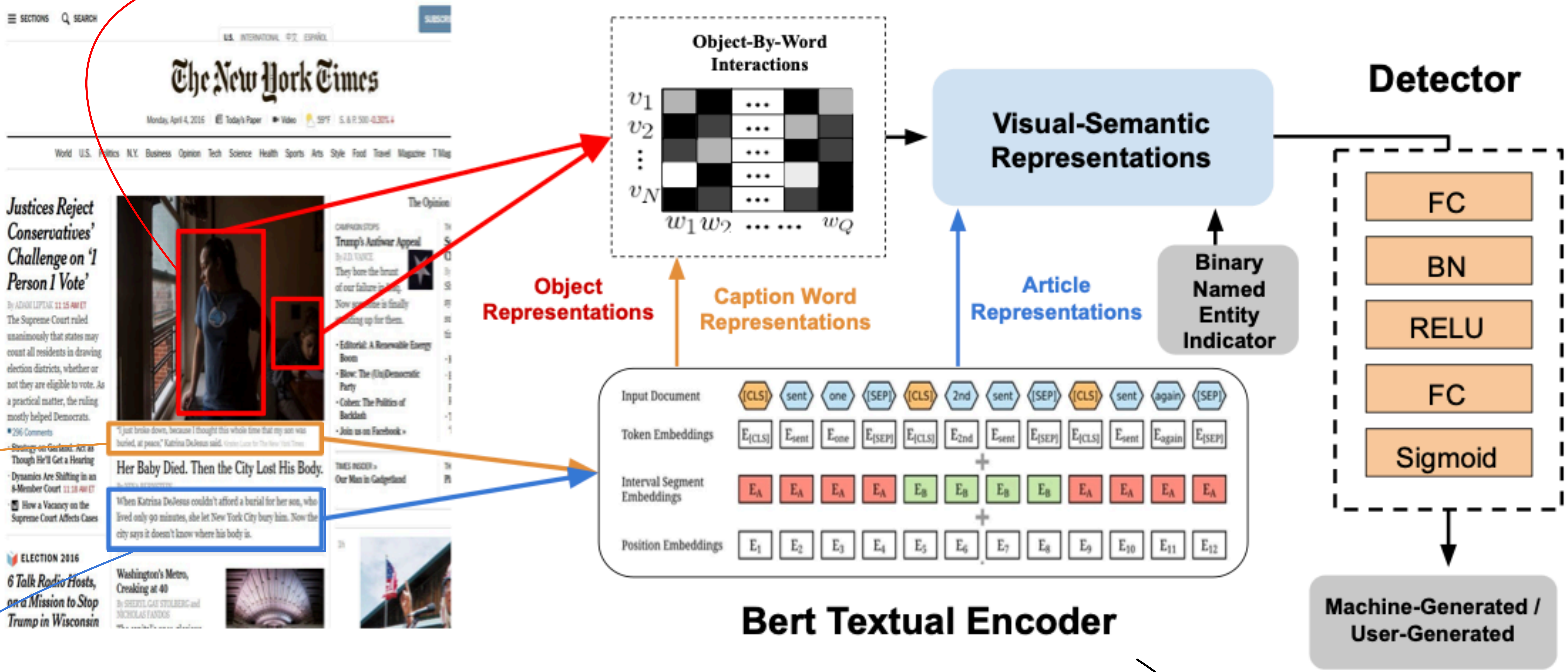
- Grover: A State-of-the-Art Defense against Neural Fake News (NeurIPS 2019)
  - Grover is a model for Neural Fake News -- both generation and detection.
  - However, it probably can also be used for other generation tasks.

Proposed Model

Framework overview

Each image I is represented by a set of regional object features $\{o_1, \cdots, o_I\}$

Each caption C contains a sequence of words $C = \{w_1, \cdots, w_I\}$



A consists of a set of sentences S where $S = \{S_1, \cdots, S_A\}$

Each sentence S_i contains a sequence of words $\{w_1, \cdots w_i\}$

Each sentence is tokenized and encoded with a BERT model that is pre-trained on BooksCorpus and English Wikipedia.

Proposed Model

Article Representations

- To extract relevant semantic context from the article, begin by computing sentence representations.
- For each sentence S^i in article A, the word representations are first projected into the article subspace as follows: $S^i = W^{art}V^i$
    - V^i: represent all word embeddings in S^i
- For a given sentence S^i, its representation S^i_f is computed as the average of all its word representations where f denotes the corresponding representation.
- The article representation A_f for an article A is computed as the average of all its sentence representation.

Proposed Model

Visual-Semantic Representation

- Proposed approach leverages word-specific image representations learned from images and captions to determine their relevance to an article.
- A caption is represented by a feature matrix $V_f^{cap} \in \mathbb{R}^{n_c \times D^T}$ and an image is represented by a matrix of object features $V_f^{vis} \in \mathbb{R}^{n_o \times D^I}$.
- Word embeddings of a caption and image object features are projected into a common visual-semantic subspace using:
    - $C_f^{cap} = W^{cap} V_f^{cap}, I_f^{vis} = W^{vis} V_f^{vis}$

Proposed Model

Visual-Semantic Representation

- A key property of these visual-semantic representation is that they are built on fine-grained interaction between words in the caption and objects in the image.
- Semantic similarity score is computed for every possible pair of projected word and object features w_l, v_k, respectively.
- $s_{kl} = \dfrac{v_k^T w_l}{\| v_k \| \ \| w_l \|}$ where $k \in [1, n_o]$ and $l \in [1, n_c]$
- n_c, n_o: indicate the number of words and object in caption and image, respectively.

Proposed Model
Visual-Semantic Representation

- These similarity scores are normalized over the objects to determine the salience of each object with respect to a word in the caption.
- $a_{kl} = \frac{\exp(s_{kl})}{\sum_{i=1}^{n_o} \exp(s_{il})}$
- The word-specific image representations are computed as a weighted sum of the object features based on the normalized attention weights:
- $w_l^I = a_l^T I_f^{vis}$

Proposed Model

Detector

- Key contribution of approach is the utilization of a binary indicator feature, which indicates if the caption contains a reference to a named entity present in the main article body.

- The article representation and the average of the word-specific image representations are concatenated to create caption-specific article representations which are passed into the discriminator:

- $$A_f^c = \text{concat}\left\{\underbrace{A_f}_{\text{Article representation}}, \underbrace{\frac{1}{n_c}\sum_{l=1}^{n_c} w_l^I}_{\text{Average of the word-specific image representations}}, \underbrace{b_c}_{\text{Binary indicator feature}}\right\}$$

Proposed Model

Detector

- The final authenticity score of an article is determined across those of its images and captions.
- It can be thought of as the probability that an article is human-generated.
- The authenticity score is computed across the set of images and captions in an article:
  - $p_A = 1 - \prod_{images} \left(1 - p_A^I\right)$
  - p_A^I: authenticity score of image-caption pair I respect to article

Proposed Model

Detector

- Intuitively, if an image-caption pair is deemed to be relevant to the article body (scores close to 1), then the final authenticity score will be close to 1 as well.

- The entire model is optimized end-to-end with a binary cross-entropy loss.

- $$L = -\sum_{(A^+, I^+)} \sum_{I^-} y \log(p_A) + (1 - y)\log(1 - p_A)$$

Experiments

Dataset Statistics

# Sentences in Article	% of Articles		# Imgs	% of Articles
	Real	Generated		
$N \leq 10$	33.7	15.6	1	60.8
$10 < N \leq 40$	54.4	81.5	2	21.0
$N > 40$	11.9	2.9	3	18.2

- NeuralNews Dataset
- Most articles contain at most 40 sentences in their main body.
- In addition, even though most articles contain a single image and caption, a sizeable 18.2% have 3 images.
  - This setting will provide a challenging testbed for future work to investigate methods using varying number of images and captions.

Experiments

Setups

- Given a news article from dataset, goal is to automatically predict whether it is human or machine-generated.
- In experiments, only use Real/Generated Articles and Real Captions are used.
    - Due to the generated captions often contain repeated instances of named entities without any stop words, which is not challenging for humans to detect.

Experiments

Baselines

- In addition to ablations of proposed model, also compare to a baseline using Canonical Correlation Analysis (CCA), which learns a shared semantic space between two sets of paired features, as well as the GROVER Discriminator.
- In CCA implementation, images are represented as the average of its object region features and the caption is represented by average of its word features.
- Apply CCA between the article features, and the concatenation of the image and caption features.
- The projection matrices in CCA are learned from positive samples constituting articles and their corresponding images and captions.

Experiments

Results:
Training on
Real News Only

Approach	Trained With Mismatch	Named Entity Indicator	Generated Articles in Training (%)	GROVER-Mega Accuracy (%)	GROVER-Large Accuracy (%)
CCA	-	-	None	52.1	-
DIDAN	✓	-	None	54.5	-
	✓	✓	None	**64.1**	-
Grover Discriminator	-	-	50	56.0	-
DIDAN	-	-	25	51.2	49.9
	-	-	50	56.4	53.7
	-	✓	25	64.9	64.6
	-	✓	50	68.8	66.3
	✓	-	25	61.0	65.0
	✓	-	50	70.3	57.4
	✓	✓	25	80.9	69.8
	✓	✓	50	**85.6**	**77.6**

- Proposed approach significantly improves detection accuracy when trained without any generated examples (i.e. with mismatch real news as negatives) compared to CCA.

- Named entity indicator (NEI) features provide a large improvement in this most difficult setting.

Experiments

Results:
Training with Generated Samples

Approach	Trained With Mismatch	Named Entity Indicator	Generated Articles in Training (%)	GROVER-Mega Accuracy (%)	GROVER-Large Accuracy (%)
CCA	-	-	None	52.1	-
DIDAN	✓	-	None	54.5	-
	✓	✓	None	**64.1**	-
Grover Discriminator	-	-	50	56.0	-
DIDAN	-	-	25	51.2	49.9
	-	-	50	56.4	53.7
	-	✓	25	64.9	64.6
	-	✓	50	68.8	66.3
	✓	-	25	61.0	65.0
	✓	-	50	70.3	57.4
	✓	✓	25	80.9	69.8
	✓	✓	50	**85.6**	**77.6**

- Grover Discriminator (like text-only variant) is substantially worse than the result reported in original paper.
    - Because train it with BERT representations as opposed to leveraging Grover learned representations to detect its own generated articles.
- Based on the consistent trend of the results, training on generated articles from the same generator as appears in test data improves the capability of a neural network to detect them.

Experiments

Results: Training with Generated Samples

Approach	Trained With Mismatch	Named Entity Indicator	Generated Articles in Training (%)	GROVER-Mega Accuracy (%)	GROVER-Large Accuracy (%)
CCA	-	-	None	52.1	-
DIDAN	✓	-	None	54.5	-
	✓	✓	None	**64.1**	-
Grover Discriminator	-	-	50	56.0	-
DIDAN	-	-	25	51.2	49.9
	-	-	50	56.4	53.7
	-	✓	25	64.9	64.6
	-	✓	50	68.8	66.3
	✓	-	25	61.0	65.0
	✓	-	50	70.3	57.4
	✓	✓	25	80.9	69.8
	✓	✓	50	**85.6**	**77.6**

- The binary NEI features also prove to be very beneficial to increasing the detection accuracy of DIDAN.
- Interestingly, even when have access to generated articles during training, the large improvement in detection accuracy going from 68.8% to 85.6%.
    - When also training on mismatched real images and captions suggests that visual-semantic consistency has an important role to play in defending against neural fake news.

Experiments

Results:

Unseen Generator

Approach	Trained With Mismatch	Named Entity Indicator	Generated Articles in Training (%)	GROVER-Mega Accuracy (%)	GROVER-Large Accuracy (%)
CCA	-	-	None	52.1	-
DIDAN	✓	-	None	54.5	-
	✓	✓	None	**64.1**	-
Grover Discriminator	-	-	50	56.0	-
DIDAN	-	-	25	51.2	49.9
	-	-	50	56.4	53.7
	-	✓	25	64.9	64.6
	-	✓	50	68.8	66.3
	✓	-	25	61.0	65.0
	✓	-	50	70.3	57.4
	✓	✓	25	80.9	69.8
	✓	✓	50	**85.6**	**77.6**

- To evaluate DIDAN's capability to generalize to articles created by generators unseen during training.
  - Train on GROVER-Large generated articles and evaluate on GROVER-Mega articles.
- While overall accuracy drops, observe the same trend where our proposed training with mismatched real data helps increase the detection accuracy from 66.3% to 77.6, and removing NEI lowers accuracy.

Articles	Images	Captions	DIDAN Accuracy (%)	CCA Accuracy (%)
✓	✓	✓	85.6	51.4
✓	-	✓	81.9	50.1
✓	✓	-	56.9	52.1

Experiments

Results:

Images vs Captions

- Observe an improvement of 2% in accuracy achieved by CCA variants with images.
    - This suggests that visual cues from images can provide contextual information vital to detecting generated articles.
- This is also corroborated by the ablation result obtained by DIDAN.
- While the contribution of the captions is the most significant, note that the visual cues provided by images are integral to achieving the best accuracy.

Summary and Defense Directions

- Provide an effective initial defense mechanism against article with images and captions.

- Based on the findings from user evaluation, adversaries are easily exploit this fact to create misleading disinformation by generating fake articles and combining them with manually sourced images and captions.

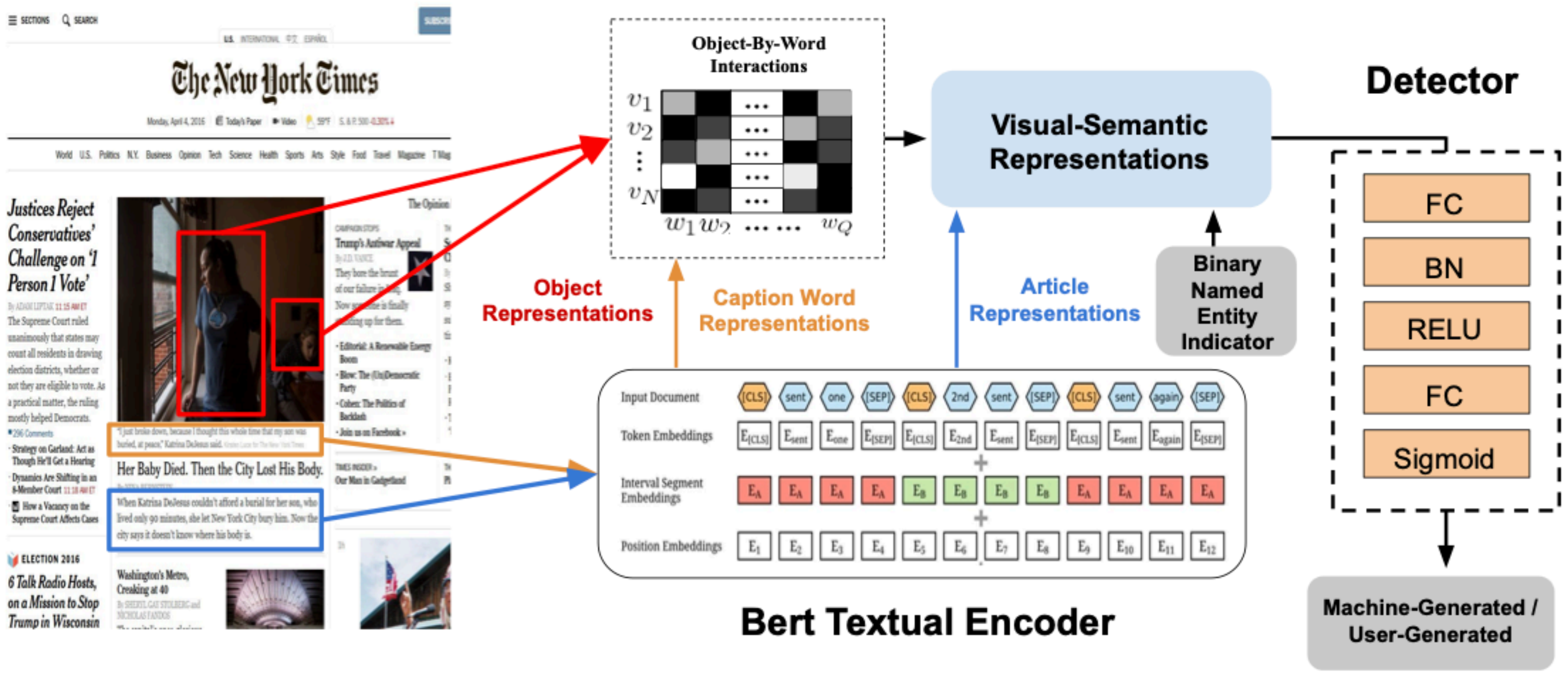
- Experimental results suggest that visual-semantic consistency is an important and promising research area in our defense against neural news.

Summary and Defense Directions

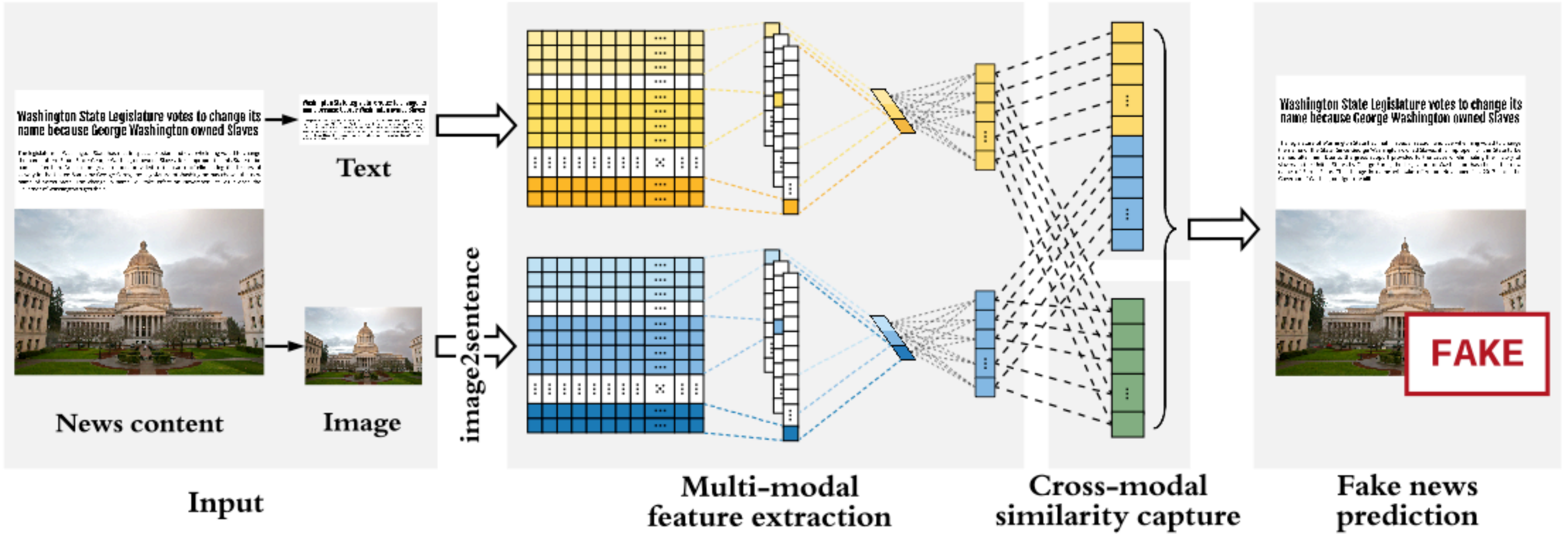
- Other interesting avenues for future research is to understand the importance of metadata in this multimodal setting.
- DIDAN and NeuralNews may leveraged to supplement fact verification in detecting human-written misinformation in general by evaluating visual-semantic consistency.

Comments
of DIDAN

- Focus on machine-generate neural fake news detection.
- Provide important rule
    - Article & Image consistency
    - Object in image - entity in caption
    - Like concept propose by SAFE (PAKDD'20)



DIDAN



SAFE