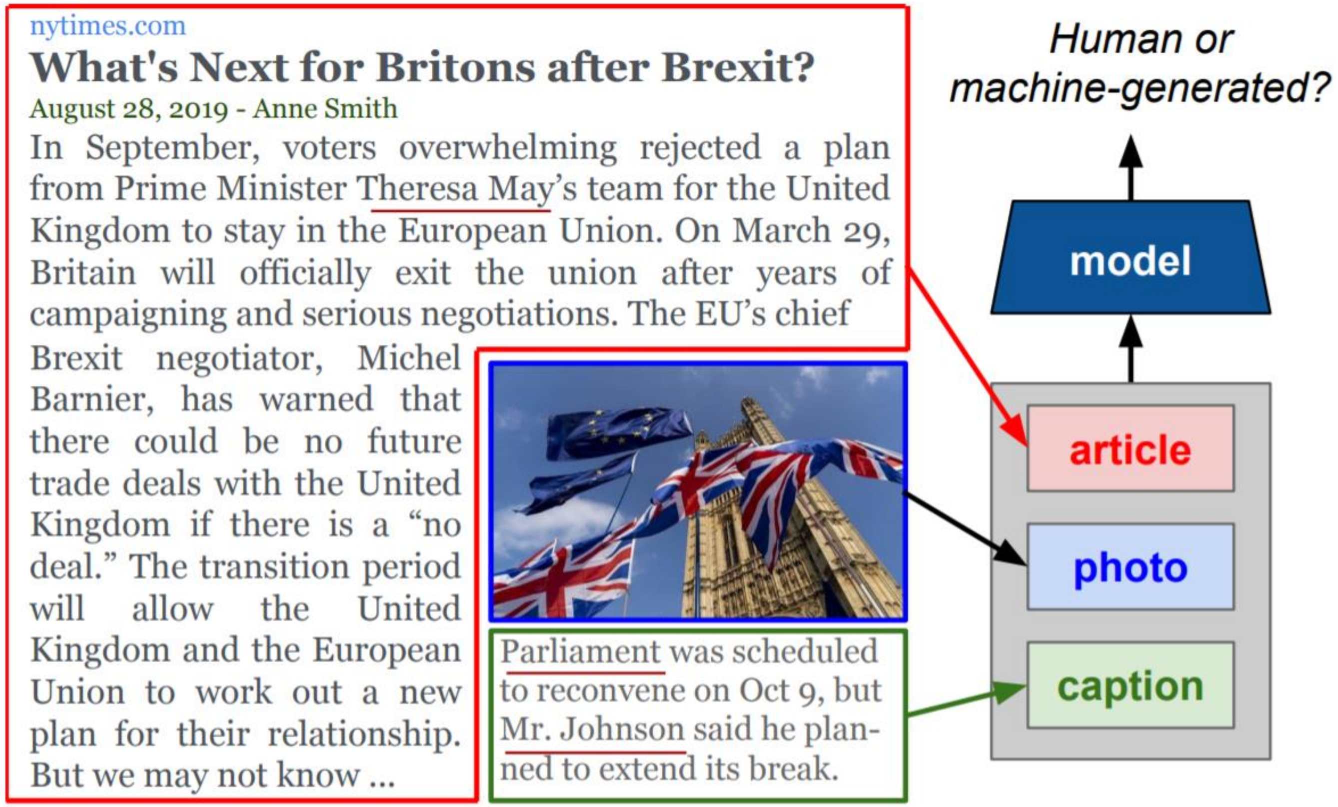
DIDAN

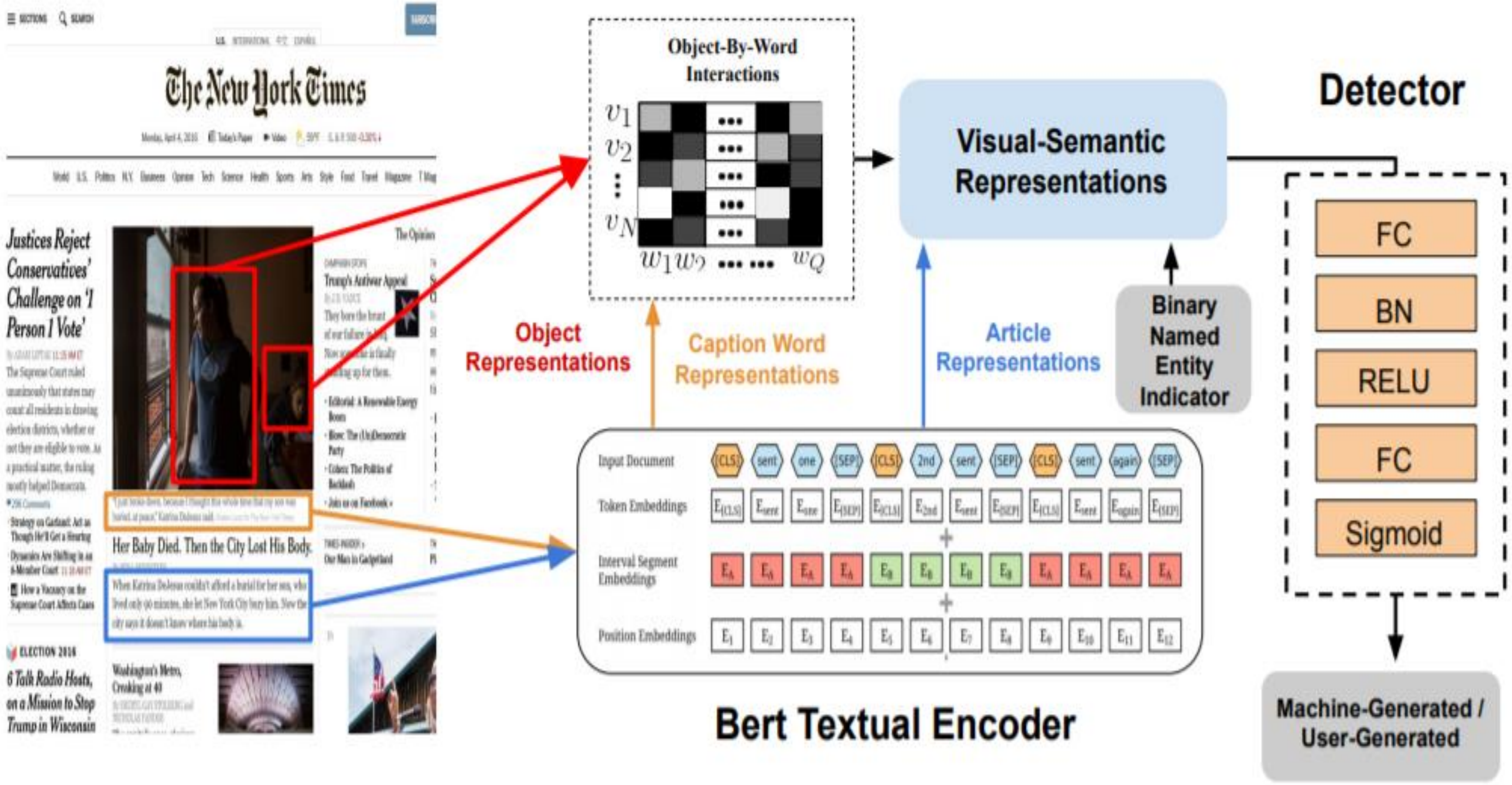
Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News

What is AI's biggest weakness?

- **Consistent Multimedia** -text matches with photo, audio etc
- Generating complex scenes
- Realistic person-specific style
- Factual consistency

Detecting Neural Fake News via Consistency





An overview of our proposed DIDAN model.

DIDAN

To reason about relationships between named entities present in the article and entities in an image, DIDAN integrates article context into the visual-semantic representation learned from fine-grained object-by-word interactions. The aforementioned visual-semantic representation is used to infer an authenticity score for the entire news article

Main Components of DIDAN

- Article Representations
- Visual-Semantic Representations
- Detector

Article Representations

To extract relevant semantic context from the article, we 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:

$$\boldsymbol{S^i = W^{art} V^i}$$

Where V^i represent all word embeddings in S^i . For a given sentence S^i , its representation S_f^i is computed as the average of all its word representations where the subscript f denotes the corresponding representation. In turn, the article representation A_f for an article A is computed as the average of all its sentence representations.

Visual-Semantic Representations

Our 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 and an image is represented by a matrix of object features.

The word embeddings of a caption and image object features are projected into a common visual-semantic subspace.

Detector

Key contribution of the 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.

Discriminator is a simple neural network that is comprised of a series of Fully-Connected (FC), Rectified Linear Unit (ReLU), Batch Normalization (BN) and sigmoid layers. It outputs an **authenticity score** for every image-caption pair.