NewsCLIPpings

Automatic Generation of Out-of-Context Multimodal Media

Grace Luo, Trevor Darrell, Anna Rohrbach

University of California, Berkeley
Motivation: Out-of-Context Images
Motivation: Out-of-Context Images

UNELECTED *House of Lords* told to stop FALLING ASLEEP.

Legitimate old photograph
Motivation: Out-of-Context Images

UNELECTED House of Lords told to stop FALLING ASLEEP.

$85/hr is too much to pay our senators to literally do nothing.

Legitimate old photograph

Evidence of recent events
Motivation: Out-of-Context Images

UNELECTED House of Lords told to stop FALLING ASLEEP.

$85/hr is too much to pay our senators to literally do nothing.

Evidence of recent events
Motivation: Out-of-Context Images

UNELECTED House of Lords told to stop FALLING ASLEEP.

$85/hr is too much to pay our senators to literally do nothing.

Alert! Automation of cheapfakes.
## Prior Out-of-Context Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAIM</strong> (Jaiswal et al, 2017)</td>
<td>Random</td>
<td>Little text-image alignment</td>
</tr>
<tr>
<td>TamperedNews (Muller-Budack et al, 2020)</td>
<td>Text entity manipulation</td>
<td>Linguistic biases</td>
</tr>
<tr>
<td><strong>COSMOS</strong> (Aneja et al, 2021)</td>
<td>Human generated</td>
<td>Small number of labelled samples</td>
</tr>
<tr>
<td>NewsCLIPpings (Ours)</td>
<td>Automatic retrieval</td>
<td>Convincing to humans, No linguistic biases, Almost 1M labelled samples</td>
</tr>
</tbody>
</table>
## Prior Out-of-Context Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAIM (Jaiswal et al, 2017)</td>
<td>Random</td>
<td>Little text-image alignment</td>
</tr>
<tr>
<td>TamperedNews (Muller-Budack et al, 2020)</td>
<td>Text entity manipulation</td>
<td>Linguistic biases</td>
</tr>
<tr>
<td>COSMOS (Aneja et al, 2021)</td>
<td>Human generated</td>
<td>Small number of labelled samples</td>
</tr>
<tr>
<td>NewsCLIPpings (Ours)</td>
<td>Automatic retrieval</td>
<td>Convincing to humans, No linguistic biases, Almost 1M labelled samples</td>
</tr>
</tbody>
</table>
## Prior Out-of-Context Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAIM (Jaiswal et al, 2017)</td>
<td>Random</td>
<td>Little text-image alignment</td>
</tr>
<tr>
<td>TamperedNews (Muller-Budack et al, 2020)</td>
<td>Text entity manipulation</td>
<td>Linguistic biases</td>
</tr>
<tr>
<td>COSMOS (Aneja et al, 2021)</td>
<td>Human generated</td>
<td>Small number of labelled samples</td>
</tr>
<tr>
<td>NewsCLIPpings (Ours)</td>
<td>Automatic retrieval</td>
<td>Convincing to humans</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No linguistic biases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Almost 1M labelled samples</td>
</tr>
</tbody>
</table>
## Prior Out-of-Context Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAIM (Jaiswal et al, 2017)</td>
<td>Random</td>
<td>Little text-image alignment</td>
</tr>
<tr>
<td>TamperedNews (Muller-Budack et al, 2020)</td>
<td>Text entity manipulation</td>
<td>Linguistic biases</td>
</tr>
<tr>
<td>COSMOS (Aneja et al, 2021)</td>
<td>Human generated</td>
<td>Small number of labelled samples</td>
</tr>
<tr>
<td>NewsCLIPpings (Ours)</td>
<td>Automatic retrieval</td>
<td>Convincing to humans</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No linguistic biases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Almost 1M labelled samples</td>
</tr>
</tbody>
</table>
NewsCLIPpings Dataset Overview

- Dataset of ~1M samples of news image repurposing.
- Diagnostic benchmark for machine generated misinformation.
- General automatic framework for generating challenging mismatches.

Approach

“Angela Merkel speaks”

Pool of candidate images

Scoring

Filtering

Angela Merkel speaks
Approach

✓ Semantics
✓ Person
✓ Scene
Approach: In-the-Wild

Photo shows U.S. Sen. Tom Cotton of Arkansas sitting on a bed of gold bars...
Approach: In-the-Wild

Photo shows U.S. Sen. Tom Cotton of Arkansas sitting on a bed of gold bars...
Photo shows U.S. Sen. **Tom Cotton** of Arkansas sitting on a bed of gold bars...
Angela Merkel speaks to the German parliament...
Approach: NewsCLIPpings

Angela Merkel speaks to the German parliament...
Angela Merkel speaks to the German parliament...
Approach: NewsCLIPpings
Semantics CLIP Text-Image
Approach: NewsCLIPpings
Semantics CLIP Text-Image

Angela Merkel speaks
to the German parliament...
Approach: NewsCLIPpings

Semantics CLIP Text-Image

CTI Sim: 0.2940

CTI Sim: 0.2799

CTI Sim: 0.2114

Angela Merkel speaks to the German parliament...
Approach: NewsCLIPpings
Semantics CLIP Text-Image

Angela Merkel speaks to the German parliament...

Falsified

CTI Sim. 0.2940

CTI Sim. 0.2799

CTI Sim. 0.2114

Angela Merkel speaks to the German parliament...
Approach: NewsCLIPpings
Semantics CLIP Text-Text

*Angela Merkel speaks to the German parliament...*
Angela Merkel speaks to the German parliament...
Ingeborg Berggreen Merkel delivers the task force's report in Berlin.

Half of Tessa Jowell's settlement will go direct to a charity in her south London constituency.

Tennessee Gov Bill Haslam delivers his State of the State address to a joint session of the General Assembly on Monday in Nashville.

Angela Merkel speaks to the German parliament...
Approach: In-the-Wild

Photo shows Tampa Mayor Jane Castor maskless at Super Bowl.
Approach: In-the-Wild

Photo shows Tampa Mayor Jane Castor maskless at Super Bowl.
Photo shows Tampa Mayor Jane Castor maskless at Super Bowl.
Prince William and Duchess Kate introduce their son to the world.
Prince William and Duchess Kate introduce their son to the world.
Prince William and Duchess Kate introduce their son to the world.
Approach: NewsCLIPpings
Person SBERT-WK Text-Text
Prince William and Duchess Kate introduce their son to the world.
Prince William and Duchess Kate introduce their son to the world.
Prince William and Duchess Kate introduce their son to the world.

Prince William the Duke of Cambridge carries his newborn baby boy to the car after leaving the hospital.

STT Sim. 0.8753

William Catherine and Harry attend the forum.

STT Sim. 0.8480

Prince William missed the net this summer at the Somba K'e Civic Plaza Yellowknife.

STT Sim. 0.7854
Approach: In-the-Wild

... trash that was left behind by ... refugees traveling ... to the United States
Approach: In-the-Wild

...trash that was left behind by ... refugees traveling ... to the United States
... trash that was left behind by ... refugees traveling ... to the United States
Approach: NewsCLIPpings

Fukushima Daiichi nuclear power plant after Japan’s earthquake and tsunami...
Approach: NewsCLIPpings

Fukushima Daiichi nuclear power plant after Japan’s earthquake and tsunami...

✓ Semantics
X NE

✓ Person
X Semantics

✓ Scene
X NE
Fukushima Daiichi nuclear power plant after Japan’s earthquake and tsunami...
Approach: NewsCLIPpings
Scene ResNet Place
Approach: NewsCLIPpings
Scene ResNet Place
Approach: NewsCLIPpings
Scene ResNet Place

RP Sim. 0.8272

RP Sim. 0.7829

RP Sim. 0.7823
Approach: NewsCLIPpings
Scene ResNet Place

Falsified

Fukushima Daiichi nuclear power plant after Japan's earthquake and tsunami in March.
Approach

- **Semantics**
  CLIP Text-Image, CLIP Text-Text

- **Person**
  SBERT-WK Text-Text

- **Scene**
  ResNet Place
Approach

✓ Semantics
CLIP Text-Image, CLIP Text-Text

✓ Person
SBERT-WK Text-Text

✓ Scene
ResNet Place
Approach

✓ Semantics
CLIP Text-Image, CLIP Text-Text

✓ Person
SBERT-WK Text-Text

✓ Scene
ResNet Place
Approach

✓ Semantics
CLIP Text-Image, CLIP Text-Text

✓ Person
SBERT-WK Text-Text

✓ Scene
ResNet Place
Note that we exclude samples in the NewsCLIPping val / test sets across all splits for the VisualNews pretraining dataset.

Experimental Setup

• Binary classification: pristine or falsified

Pristine or Falsified?

The fire at the Darul Uloom school...
Note that we exclude samples in the NewsCLIPping val / test sets across all splits for the VisualNews pretraining dataset.

Experimental Setup

- Binary classification: pristine or falsified
- Balanced wrt captions, labels

Pristine or Falsified?

The fire at the Darul Uloom school...
Experimental Setup

- Binary classification: pristine or falsified
- Balanced wrt captions, labels

- **CLIP** (Radford et al., 2021): pretrained on 400M web corpus
- **VisualBERT** (Li et al., 2019): pretrained on 3M Conceptual Captions (Sharma et al., 2018) or 1M VisualNews (Liu et al., 2020)

Note that we exclude samples in the NewsCLIPping val / test sets across all splits for the VisualNews pretraining dataset.

Pristine or Falsified?

The fire at the Darul Uloom school...
Experiments: Unimodal vs Multimodal

CLIP Binary Classification Performance by Modality

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP Text-Only</td>
<td>0.5</td>
</tr>
<tr>
<td>CLIP Image-Only</td>
<td>0.6</td>
</tr>
<tr>
<td>CLIP</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Dataset Split
- Semantics / CLIP Text-Text
- Semantics / CLIP Text-Image
- Person / SBERT-WK Text-Text
- Scene / ResNet Place
Experiments: Unimodal vs Multimodal

CLIP Binary Classification Performance by Modality

Dataset Split
- Semantics / CLIP Text-Text
- Semantics / CLIP Text-Image
- Person / SBERT-WK Text-Text
- Scene / ResNet Place
Experiments: Unimodal vs Multimodal

CLIP Binary Classification Performance by Modality

Accuracy

Dataset Split
- Semantics / CLIP Text-Text
- Semantics / CLIP Text-Image
- Person / SBERT-WK Text-Text
- Scene / ResNet Place

Model
- CLIP Text-Only
- CLIP Image-Only
- CLIP
Experiments: Unimodal vs Multimodal

CLIP Binary Classification Performance by Modality

- CLIP Text-Only
- CLIP Image-Only
- CLIP

Dataset Split
- **Semantics** / CLIP Text-Text
- **Semantics** / CLIP Text-Image
- **Person** / SBERT-WK Text-Text
- **Scene** / ResNet Place
Experiments: Unimodal vs Multimodal

CLIP Binary Classification Performance by Modality

Accuracy

CLIP Text-Only  CLIP Image-Only  CLIP

Model

Dataset Split
- Semantics / CLIP Text-Text
- Semantics / CLIP Text-Image
- Person / SBERT-WK Text-Text
- Scene / ResNet Place
Experiments: Multimodal

CLIP, VisualBERT Binary Classification Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP</td>
<td>0.7</td>
</tr>
<tr>
<td>VisualBERT-CC</td>
<td>0.6</td>
</tr>
<tr>
<td>VisualBERT-VN</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Dataset Split
- Semantics / CLIP Text-Text
- Semantics / CLIP Text-Image
- Person / SBERT-WK Text-Text
- Scene / ResNet Place
Experiments: Multimodal

CLIP, VisualBERT Binary Classification Performance

Accuracy

Model

Dataset Split
- **Semantics / CLIP Text-Text**
- **Semantics / CLIP Text-Image**
- **Person / SBERT-WK Text-Text**
- **Scene / ResNet Place**
Experiments: Multimodal

CLIP, VisualBERT Binary Classification Performance

Accuracy

Model

Dataset Split
- **Semantics** / CLIP Text-Text
- **Semantics** / CLIP Text-Image
- **Person** / SBERT-WK Text-Text
- **Scene** / ResNet Place

CLIP
VisualBERT-CC
VisualBERT-VN
Experiments: Multimodal

CLIP, VisualBERT Binary Classification Performance

Dataset Split
- Semantics / CLIP Text-Text
- Semantics / CLIP Text-Image
- Person / SBERT-WK Text-Text
- Scene / ResNet Place
Experiments: Multimodal

CLIP, VisualBERT Binary Classification Performance

Accuracy

- **CLIP**
- **VisualBERT-CC**
- **VisualBERT-VN**

**Dataset Split**
- Semantics / CLIP Text-Text
- Semantics / CLIP Text-Image
- Person / SBERT-WK Text-Text
- Scene / ResNet Place
Experiments: Multimodal

CLIP, VisualBERT Binary Classification Performance

Accuracy

Model

Dataset Split

- **Semantics / CLIP Text-Text**
- **Semantics / CLIP Text-Image**
- **Person / SBERT-WK Text-Text**
- **Scene / ResNet Place**
Experiments: Multimodal

CLIP, VisualBERT Binary Classification Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset Split</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP</td>
<td>Semantics / CLIP Text-Text, Semantics / CLIP Text-Image, Person / SBERT-WK Text-Text, Scene / ResNet Place</td>
</tr>
<tr>
<td>VisualBERT-CC</td>
<td></td>
</tr>
<tr>
<td>VisualBERT-VN</td>
<td></td>
</tr>
</tbody>
</table>
Human Evaluation

• Same binary classification task

Does the image belong to the given caption?

The fire at the Darul Uloom school...
Human Evaluation

- Same binary classification task
- 100 sample subset
  - ~25% from each split
  - ~50% pristine, falsified
Human Evaluation

- Same binary classification task
- 100 sample subset
  - ~25% from each split
  - ~50% pristine, falsified
- Before and after internet search

Does the image belong to the given caption?

The fire at the Darul Uloom school...
Human Evaluation

Human Binary Classification Performance

Accuracy

- Human - Before Search
- Human - After Search
- CLIP
Human Evaluation

Human Binary Classification Performance

Accuracy

- Human - Before Search
- Human - After Search
- CLIP
Future Directions

North Carolina State
Wolfpack forward Beejay Anya...

Guinea's David Alvarez celebrates after scoring the winner against Senegal.

... rally against labour reforms in Athens...
Guinea’s David Alvarez celebrates after scoring the winner against Senegal.
Future Directions

North Carolina State Wolfpack forward Beejay Anya...

Guinea's David Alvarez celebrates after scoring the winner against Senegal.

... rally against labour reforms in Athens ...
Takeaways

- The automation of cheapfakes is now a realistic threat.

- Humans are susceptible to fakes from NewsCLIPpings.

- Models require improvements in understanding of symbols, facial expressions, and locations.
Thank you!

https://github.com/g-luo/news_clippings
http://arxiv.org/abs/2104.05893