Late Fusion with Triplet Margin Objective for Multimodal Ideology Prediction and Analysis

Changyuan Qiu*, Winston Wu*, Xinliang Frederick Zhang, Lu Wang
Computer Science and Engineering, University of Michigan
{peterqiu, wuws, xlfzhang, wangluxy}@umich.edu

Introduction

Media bias manifests in many different ways and can influence readers' perception of events. Previous work on identifying media bias has focused solely on *text*. We introduce the new task of **multimodal ideology prediction**, which seeks to identify 5-way political ideology given both text and a corresponding image. For this task, we collect five new large-scale datasets with images and text across the political spectrum from 11 US media sources, Reddit, and Twitter. We conduct in-depth analyses of news articles and reveal differences in image content and usage across the political spectrum. Furthermore, we perform extensive experiments and ablation studies, demonstrating the effectiveness of targeted pretraining objectives on different model components, including our new ideology-targeted triplet margin objective.

Why multimodal?

Images can contain information not expressed in the text. Images are often chosen to reinforce viewpoints expressed in the text. Thus, images should help with predicting the ideology of an article. For example,

Topic: *Federal Judge Pauses Justice Department Effort to Dismiss Michael Flynn Case*

New York Times

- many people
- positive expression

Fox News

- single person
- negative expression

Data

We collect five new datasets of articles and images. First, we build upon BigNewsBln [Liu+ 2022], a collection of 1.9M English news articles from 11 news sources.

- **BN-ImgCap**: 1.2M image-caption pairs that occur anywhere within a news article
- **BNA-Img**: images from a subset of articles in BigNewsBln containing stories associated with a story cluster

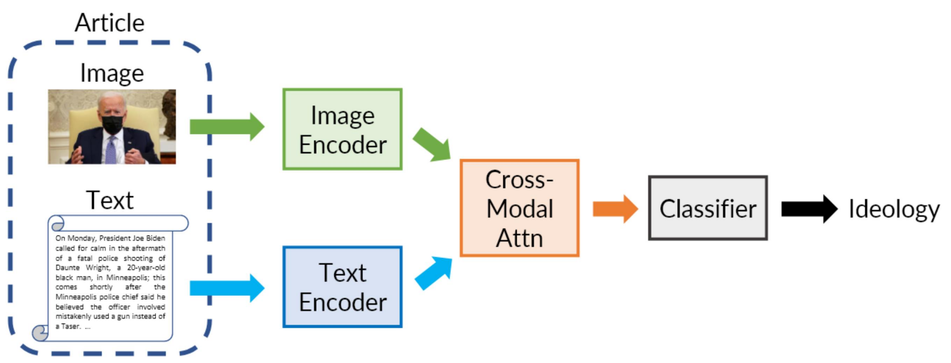
AllSides: 12K images and articles from AllSides, a website that associates stories about a particular event with articles from different sources writing about the same event

Reddit: 357K Reddit posts with images from left-leaning (r/Liberal, r/democrats, r/progressive) and right-leaning (r/Conservative, and r/Republican) subreddits

Twitter: 57K tweets with images from 1.4K politicians

Please see the paper for in-depth data analysis!

Multimodal Ideology Prediction



5-way political ideology (Left, Lean-Left, Center, Lean-Right, Right)

Multimodal Models

Early-Fusion
- VisualBERT [Li+ 2019]
- ViLT [Kim+ 2021]

Late-Fusion
- Text Encoder
  - RoBERTA [Liu+ 2019]
  - POLITICS [Liu+ 2022]
- Image Encoder
  - Swin Transformer [Hu+ 2019]
- Joining Method
  - Concatenation
  - Hadamard product
  - Gated fusion [Wu+ 2021]
  - Cross-modal attention (LXMERT) [Tan and Bansal, 2019]

Continued Pretraining

Text Encoder: POLITICS [Liu+ 2022]
Image Encoder: InfoNCE loss [Radford+ 2021]
Joint
- Bidirectional caption loss (VirTex) [Desai and Johnson, 2021]
- Ideology-driven Triplet Margin Loss [this work]

$$\sum_{\mathbf{t}\in T}\left[||\mathbf{t}^{(a)}-\mathbf{t}^{(p)}||_2-||\mathbf{t}^{(a)}-\mathbf{t}^{(n)}||_2+\alpha\right]_+$$

Experimental Results

- Late-fusion with cross-modal attention works the best
- Continued pretraining with our triplet margin loss improves performance overall, and especially on Right articles

Category	Model	Acc.	Macro F_1
Early Fusion	VisualBERT	78.45 ± 0.69	75.34 ± 0.67
	ViLT	78.39 ± 1.24	76.22 ± 1.43
Late Fusion	RoBERTa+Swin-S		
	Concat.	82.39 ± 0.59	79.82 ± 1.02
	Hadamard Prod.	85.14 ± 0.74	82.62 ± 1.15
	Gated Fusion	82.77 ± 1.24	80.71 ± 1.20
	Cross-modal Attn.	**86.88 ± 0.38**	**85.47 ± 0.41**

Pre-training Component & Objective			Acc.						Macro F1
Text Enc.	Image Enc.	Cross-modal Attn.	Overall Acc.	Left	Lean Left	Center	Lean Right	Right	Overall F1
✗	✗	✗	85.47 ± 0.41	69.10 ± 3.34	88.03 ± 0.85	91.41 ± 0.83	88.15 ± 1.31	76.40 ± 2.38	82.62 ± 1.15
✓(POLITICS)	✗	✗	86.80 ± 0.72	**96.42 ± 0.96**	**92.42 ± 1.95**	90.36 ± 0.93	81.80 ± 1.45	72.58 ± 3.43	86.39 ± 0.72
✗	✓(InfoNCE)	✗	86.40 ± 0.74	88.35 ± 1.24	90.09 ± 0.50	89.96 ± 0.88	82.37 ± 1.04	81.24 ± 1.84	86.12 ± 0.73
✗	✓(VirTex-style)	✗	87.60 ± 0.47	88.72 ± 1.89	89.77 ± 0.73	89.87 ± 0.91	85.58 ± 0.74	84.06 ± 0.84	87.84 ± 0.45
✗	✗	✓(Triplet Margin)	87.86 ± 0.93	87.99 ± 0.62	90.88 ± 0.76	90.67 ± 1.68	86.19 ± 0.61	83.59 ± 2.16	87.45 ± 0.87
✓(POLITICS)	✓(VirTex-style)	✗	88.49 ± 0.58	94.03 ± 1.30	88.87 ± 2.65	**92.30 ± 2.73**	**88.62 ± 2.17**	78.52 ± 2.22	88.26 ± 0.88
✗	✓(VirTex-style)	✓(Triplet Margin)	88.18 ± 0.56	87.80 ± 0.99	90.32 ± 0.92	91.83 ± 1.11	85.98 ± 0.76	**84.97 ± 0.94**	88.16 ± 0.59
✓(POLITICS)	✓(VirTex-style)	✓(Triplet Margin)	**88.98 ± 0.65**	91.04 ± 1.50	91.08 ± 0.59	91.36 ± 0.84	85.82 ± 0.55	84.90 ± 0.57	**88.64 ± 0.68**

More ablations and results on Reddit and Twitter in paper!