Joint Image-Text Classification Using a Transformer-Based Architecture

Patrick Y. Wu [1] Walter R. Mebane, Jr. [2]

[1]University of Michigan, pywu@umich.edu [2]University of Michigan, wmebane@umich.edu

Multimodality in Social Media Research

Say we were interested in classifying tweets depicting or discussing long lines at polling places and we come across this tweet.

"Early voting lines in Palm Beach County, Florida #iReport #vote #Florida @CNN"

The tweet depicts a long line at a polling place: the image shows a long line and the text indicates that it is a polling place. The text and image have to be jointly considered together to make the proper classification. A classifier that separately considers images and text may incorrectly classify this tweet.

Without considering images, much of the nuance of discussions on Twitter is lost. We can use *multimodal models* to create joint representations of image and text.

Related Work

Prior approaches have focused on using separate classifiers for images and text. More recent approaches have created joint representations of images and text using pretrained language models and pretrained image models (see, e.g., VisualBERT (Li et al. 2019) and ViLBERT (Lu et al. 2019)). They typically involve inputting both images and text into transformers and are usually pretrained on image captioning data, such as MS COCO or Conceptual Captions. These models, however, require all observations to have both image and text and require substantial computational resources. As a result, they may not be suitable for social media research applications, where posts are often missing either image or text.

Summary of Contributions

We introduce a multimodal framework that produces joint representations of images and text called **MARMOT**: **M**ultimod**a**l **R**epresentations using **Mo**dality **T**ranslation. The model uses BERT, a pretrained language model (Devlin et al. 2018), and ResNet-152 (He et al. 2015), a pretrained image model. It has two key methodological contributions. First, instead of using image features directly with the transformer, **we first "translate" the image features using a pretrained transformer decoder** before using the image features with text features. Second, **the model does not require that every observation have both image and text**. Training and inference require modest computational resources. The model **outperforms the current state-of-the-art multimodal classifiers**.

MARMOT Architecture

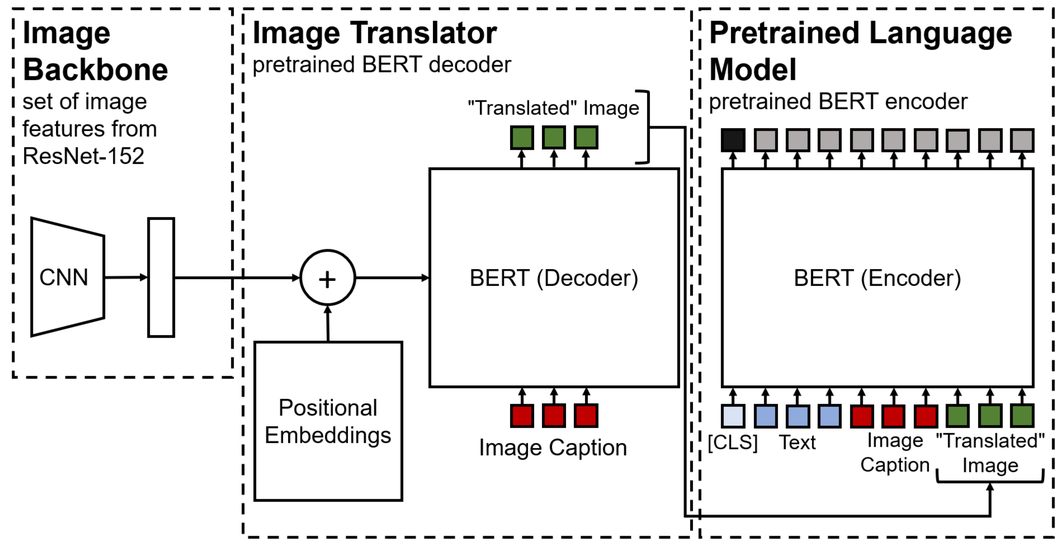

Image Backbone. We use ResNet-152 to extract lower-resolution activation maps of our images.

Image Translator. We use self-critical sequence training to generate image captions for each image (Rennie et al. 2016). We input the image captions into a transformer decoder initialized with pretrained BERT weights. It receives as initial input the image captions and receives, at the encoder-decoder attention layer, the image feature maps. This step, inspired by the use of transformers for machine translation tasks, produces "translated" image feature maps.

Pretrained Language Model. Text, image captions, and "translated" image feature maps are all inputted into a transformer architecture initialized with pretrained BERT weights. The transformer outputs the joint image-text representation.

Observations Missing Modalities. We can simply mask out the missing modality using attention masks. Observations can be missing text or image.

Future Extensions. The goal is to eventually generalize the model such that it can produce joint representations of an arbitrary number and types of modalities.

Application: Tweets of Election Incidents

We use MARMOT with the tweet data from Mebane et al. 2018, which looked at tweets reporting election incidents during the 2016 U.S. general election. Tweets labeled election incidents were further labeled with a category and an adjective under that category. Results are from a test set created from 20% of the approximately 4,000 labeled tweets. The reported results are F1 scores over the positive class.

	Ensemble	BERT	MARMOT
Not an Incident	**0.66**	0.61	0.65
Line Length	0.91	**0.93**	0.92
(a) No crowd or no line	0.61	0.64	**0.70**
(b) Small crowd or short line	0.21	0.21	**0.33**
(c) Large crowd or long line	0.82	0.86	**0.87**
Polling Place Event	0.78	**0.81**	0.80
(a) Did not function as expected	0.08	0.32	**0.42**
(b) Neutral observation	0.47	0.65	**0.67**
(c) Functioned properly	0.63	**0.65**	0.65
Electoral System	0.63	0.63	**0.64**
(a) Neutral observation	0.05	0.24	**0.25**
(b) No comment on function	0.65	0.67	**0.68**
Absentee / Mail-in Voting Issue	**0.87**	0.83	0.85
(a) Did not function properly	0.34	**0.52**	0.52
(b) Neutral observation	0.60	0.65	**0.70**
(c) Functioned properly	0.70	0.73	**0.74**
Registration	0.85	0.86	**0.86**
(a) Not able to register	0.17	0.13	**0.50**
(b) Neutral observation	0.84	0.81	**0.87**
(c) Able to register	0.50	0.60	**0.80**

Application: Hateful Memes

We use MARMOT with the recently released hateful memes dataset (Kiela et al. 2020). The goal is to classify memes as hateful or not hateful.

Model	Accuracy	AUC
Image - Grid	0.5200	0.5263
Image - Region	0.5213	0.5592
Text BERT	0.5920	0.6508
Late Fusion	0.5966	0.6475
Concat BERT	0.5913	0.6579
MMBT - Grid	0.6006	0.6792
MMBT - Region	0.6023	0.7073
ViLBERT	0.6230	0.7045
VisualBERT	0.6320	0.7133
ViLBERT CC	0.6110	0.7003
VisualBERT COCO	0.6473	0.7141
MARMOT	**0.6620**	**0.7489**