Prompt-based Alignment of Headlines and Images Using OpenCLIP

Lucien Heitz, Yuin Kwan Chan, Hongji Li, Kerui Zeng,
Abraham Bernstein, Luca Rossetto
University of Zurich, Switzerland

MediaEval Workshop 1-2 February 2024, Amsterdam, Netherlands and Online

Takeaway Message

- *Out-of-the-box* CLIP models perform well for aligning images and descriptive captions
- Creating ‘caption-like’ article text from headline and lead instead of headlines *underperforms*
- *Decreased performance* with the inclusion of AI-generated pictures when compared with editorially selected images

Motivation & Related Work

- Using LAION-5B OpenClip model, pretrained on web-images
- Caption-focus motivated by CLIP being trained with captions…
- …but headlines make use of specific grammatical constructs
- Bridging this gap by…
    - rewriting of headlines and leads
    - inclusion of additional article information
    - inclusion of named entity information

Approach

- We used OpenCLIP for processing all three datasets
- Formulation of ‘caption-like’ text using:
  - Raw title (baseline for comparison)
  - Pre-processed title (adjusted title)
  - Raw tags (additional article information)
  - T5 (completely rewrite text)
  - NER-TextRank 10 (additional named entity information)

Results (Hits@100)

	GDELT1 Train	GDELT1 Test	GDELT2 Train	GDELT2 Test	RT Train	RT Test
Raw title	.818	.943	.778	.915	.481	.635
Pre-processed title	.790	.923	.747	.902	.456	.628
Raw Tags	.778	.925	.727	.892	.429	.545
T5	.788	.927	.751	.906	.356	.491
NER	.713	.871	.677	.848	.368	.507

Lessons Learned: Insight

- Best performance of raw title indicates…
  - editors *predominantly* focus on titles for image selection
  - additional information (tags & NER) *do not help*
- Focus on *language-specific* rewriting of prompts
- *Separation* of editorial and AI-generated content

Lessons Learned: Outlook

- Conduct interviews to ask editors…
  - what the basis is for the image selection
  - what the tools are at their disposal
  - ask about user preferences (for generation model)
- Reversing the pipeline by creating captions for images
- Train model to generate better caption-like prompts

Questions?

Lucien Heitz, heitz@ifi.uzh.ch

Luca Rossetto, rossetto@ifi.uzh.ch