Multimodal CL and NLP: Combining Language and Vision for (Computational) Semantics

Carina Silberer

IMS, University of Stuttgart

CL Fall School 2024, Passau
16 - 20 September 2024

Outline

Outline

Multimodal Machine Translation (MMT)

[Shen et al., 2024]

Multimodal Machine Translation

Idea: Use visual context to improve machine translation (MT)

- Resolve ambiguities due to polysemy
  "Hold our course" $\rightarrow$ "Halten Sie unseren Kurs [?ab]" (course: Kurs; Lehrveranstaltung)
- Useful for subtitle translation
- Most common task: Scene–Image Translation

Multimodal Machine Translation (MMT)

[Shen et al., 2024]

Multimodal Machine Translation

Idea: Use visual context to improve machine translation (MT)

- Resolve ambiguities due to polysemy
  "Hold our course" → "Halten Sie unseren Kurs [?ab]" (course: Kurs; Lehrveranstaltung)
- Useful for subtitle translation
- Most common task: Scene–Image Translation

Other MMT tasks

- Text Image MT: translate the text contained in images or generate images with the text in the target language
- Video-guided MT (e.g., for subtitle translation in social media) and multimodal chat translation
  ⇒ ambiguous words and pronominal anaphora

Multimodal Machine Translation (MMT)

Multi30K: Multilingual English-German Image Descriptions [Elliott et al., 2016, Barrault et al., 2018]

Scene–Image MMT

- Given an image and a description in English (and 2 other languages)
- Generate a description in a target language
- Evaluation measures include "Lexical Translation Accuracy" focusing on a subset of ambiguous source language words

En: A boy dives into a pool near a water slide.
De: Ein Junge taucht in der Nähe einer Wasserrutsche in ein Schwimmbecken.
Fr: Un garçon plonge dans une piscine près d'un toboggan.
Cs: Chlapec skáče do bazénu poblíž skluzavky.

Multimodal Machine Translation (MMT)

Multi30K: Multilingual English-German Image Descriptions [Elliott et al., 2016, Barrault et al., 2018]

Multi30K Dataset

- Based on Flickr 30K [Young et al., 2014]: 31K images and their descriptions
- Includes German translated descriptions (without image) and newly collected descriptions
- French, Czech extensions [Barrault et al., 2018]

Multimodal Machine Translation (MMT)

[Shen et al., 2024]

Scene–Image MT: Approaches

1. Use image and text equally, e.g. applying cross-modal attention
2. Use image as supplement to text
   - for encoding the source text
   - for decoding, to assist the translation
   - for encoding and decoding
3. Text–to–image generation for model training
4. Text–to–image retrieval

Multimodal Machine Translation (MMT)

[Shen et al., 2024]

Scene–Image MT: Analysis

- Images can provide helpful contexts when textual context is ambiguous or underspecified
- Images are less effective if the textual context is sufficient
- Possible issues:
  - Dataset problematic to draw conclusions, if it contains too simple language (e.g., [Elliott et al., 2016])
    $\Leftarrow$ Incrementally degrading information from text shows that image context can be helpful [Caglayan et al., 2019]
  - Visual encoders not strong enough
    $\Leftarrow$ Stronger visual encoders more helpful for translation learning from images
- Image contexts helpful in inferring the gender pronoun [Li et al., 2021]

Multimodal Machine Translation

Further References

- CLIPTrans https://openaccess.thecvf.com/content/ICCV2023/papers/Gupta_CLIPTrans_Transferring_Visual_Knowledge_with_Pre-trained_Models_for_Multimodal_Machine_ICCV_2023_paper.pdf
- Multimodal Transformer https://aclanthology.org/2020.acl-main.400.pdf
- Survey on MMT, including tasks, methods, datasets and challenges [Shen et al., 2024]

Outline

Multimodal Affective Computing

Emotions, sentiment, personality, mood, attitudes

- Emotions influence human cognition, decision making, social interactions
- **Sentiment analysis:** Extracting the positive or negative orientation that a writer expresses
  - media: consumer reviews of books or movies, newspaper editorials, blogs, tweets, restaurants
- **Emotion recognition:** relevant in e-health, e-learning, robotics, and e-commerce
  - dialogue systems, e.g., tutoring systems: detect the student's (un)happiness, boredom, confidence etc.
  - customer reviews or responses, e.g., anger, dissatisfaction, trust
  - medical NLP, e.g., diagnose depression or suicidal intent
  - literary studies, e.g., how were different social groups viewed by society at different times?

Multimodal Affective Computing

Emotions, sentiment, personality, mood, attitudes

- Stance detection in human-human conversations
- User personality detection for conversational agents to mimic user's personality, e.g., extroversion, introversion, openness
- Affect generation and recognition for conversational agents in various domains, including literacy tutors such as children's storybooks, computer games

Emotion Recognition

Theories of emotion

- **Discrete/categorical model** with basic emotions: surprise, happiness, anger, fear, disgust, sadness (Ekman 1999); joy–sadness, anger–fear, trust–disgust, anticipation–surprise (Plutchik, 1980)

Emotion Recognition

Theories of emotion

- **Discrete/categorical model** with basic emotions: surprise, happiness, anger, fear, disgust, sadness (Ekman 1999); joy–sadness, anger–fear, trust–disgust, anticipation–surprise (Plutchik, 1980)
- **Dimensional models**: interconnected emotions varying in intensity Parameters (Russell and Mehrabian, 1977):
  - valence (positive/negative; pleasantness of the stimulus),
  - arousal (high/low; level of excitement, alertness, activeness, or energy elicited by the stimulus),
  - dominance (High/low; level of influence/control/dominance exerted by the stimulus or the emotion)

Emotion Recognition

Theories of emotion

- **Discrete/categorical model** with basic emotions: surprise, happiness, anger, fear, disgust, sadness (Ekman 1999); joy–sadness, anger–fear, trust–disgust, anticipation–surprise (Plutchik, 1980)
- **Dimensional models**: interconnected emotions varying in intensity
- **Appraisal theory (componential model)**: emotions are influenced by person's evaluation/appraisal of event

Emotion Recognition

Theories of emotion

- **Discrete/categorical model** with basic emotions: surprise, happiness, anger, fear, disgust, sadness (Ekman 1999); joy–sadness, anger–fear, trust–disgust, anticipation–surprise (Plutchik, 1980)

Emotion Recognition

[Ezzameli and Mahersia, 2023]

Can Automatic Emotion Recognition Work?

Attributes of emotions (Ekman, 1999)

- Originate from innate instincts
- Various individuals manifest the same emotion in response to the same circumstances
- People tend to express basic emotions similarly
- The physiological patterns of different people are consistent when experiencing basic emotions

Mutimodal Emotion Recognition

Why Images and Text (or other modalities)? [Ezzameli and Mahersia, 2023]

- Different modalities express complementary information voice, facial expressions, etc.
- The combination of different modalities may express a different emotion than the individual parts
- Social media revolves around images and text

Multimodal Emotion Recognition

Aspects in ER

- Stimulus: Which object/event/word caused the emotion?
- Author vs. reader perspective
- Intent
- Modalities in emotion research: physiological signals, speech, text, facial expressions and body movements

Mutimodal Emotion Recognition

Approaches in VL-ER [Ezzameli and Mahersia, 2023]

- Modality fusion
  - Feature-level fusion + classifier on fused representations
    Emotional characteristics of various modalities not considered
  - Decision-level fusion
    Interplay between features not considered
  - Model-level fusion
    Learns multimodal interaction within the model and builds a shared representation space

Emotion/Sentiment/Intent: Multimodal Document Intent

[Kruk et al., 2019]

Meaning Multiplication (Bateman, 2014)

- Complex relation between text and image
- Image and text jointly may create a new meaning beyond the literal meaning

Image I	Image II
	
Nothing like a toke after a long shift at work. Nicotine to the rescue!	Propaganda. No one who smokes for a long while ever looks this good.

Emotion/Sentiment: Multimodal Document Intent

[Kruk et al., 2019]

Multimodal Document Intent Dataset (MDID

- 1299 Instagram posts of image–caption pairs
- Three taxonomies:
  1. the **author's intent** behind the image–caption pair
  2. the **contextual relationship** between the literal meanings of the image and caption
  3. the **semiotic relationship** between the signified meanings of the image and caption

Emotion/Sentiment: Multimodal Document Intent

[Kruk et al., 2019]

MDID: Intent Taxonomy

- 8 classes derived by clustering and inspecting instagram posts

- advocative: advocate for a figure, idea, movement etc.
- promotive: promote events, products, organisations etc.
- expressive: express emotion, attachment, admiration towards an external entity
- ...

Advocative

Please, pick up after yourself when out hiking and camping. We want to perserve these beautiful sights! Do not litter!

Provocative

These punks are coming around our neighborhood, tagging everything and leaving their trash. STAY OUT!

Expressive

Had such a beautiful morning with my family. Love my little nephew, every moment is a joyful one.

Promotive

The Moulin Rouge broadway show is spectacular! You must see it! Get your tickets right away for the best show ever.

Emotion/Sentiment: Multimodal Document Intent

[Kruk et al., 2019]

MDID: Contextual Taxonomy

- What is the relationship between the literal meanings of the image and the text?
  - *Minimal* literal overlap between image and text
  - *Close* literal overlap considerably
    e.g., a selfie of a person at a crowded waterfall, "Selfie at Hemlock falls on a crowded sunny day"
  - *Transcendent*: literal meaning in one modality picks up and extends literal meaning of the other

Emotion/Sentiment: Multimodal Document Intent

[Kruk et al., 2019]

MDID: Semiotic Taxonomy

- What is the relationship between the signified meanings of the image and the text?
- *Divergent* vs. *parallel* vs. *additive*
- Do the meanings of the image and the text pull in an *opposite* vs. the *same* direction, or do they *amplify/modify* each other?

Additive	Parallel	Divergent
		
An old favorite that feels relevant these days. Art is a reflection of our society. The beauty of it as well as ugliness. To censor it is an attempt to stifle awareness, education, and critical thinking.	An aerial view of the flowers left outside Buckingham palace after the death of Princess Diana, 1997.	Good Morning.

Emotion/Sentiment: Multimodal Document Intent

[Kruk et al., 2019]

Classification Approaches

- Baseline:
  - ResNet visual encoder (V),
    ELMo contextualised caption embeddings (T)
  - Modality fusion by mapping to same space, then vector addition (V+T);
    classifier on top
  - Accuracy:
    Intent: V+T 56.7
    Contextual: T 65.4
    Semiotic: T vs. V+T 61.7 vs. 61.8

Exercise

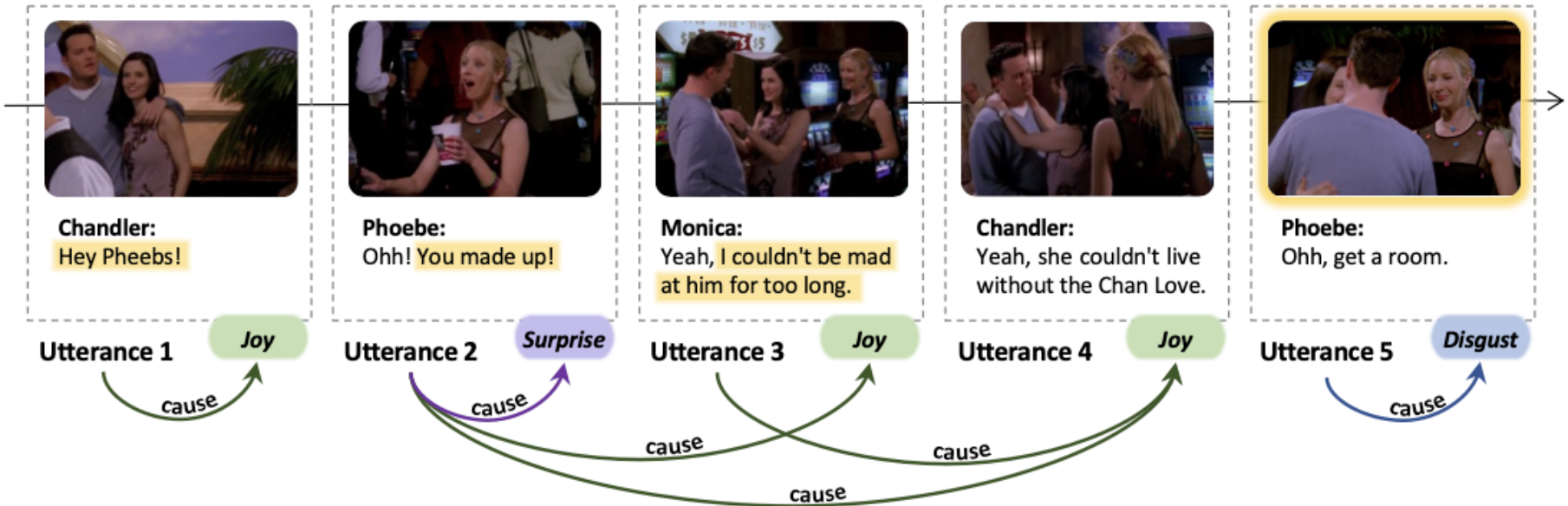
- Apply the contextual and semiotic taxonomy onto multimodal recipes, see also `https://aclanthology.org/N19-1056.pdf`
- Apply the intent taxonomy to advertisements, see also [Hussain et al., 2017] and [Zhang et al., 2018]

Multimodal Emotion Stimulus/Cause Detection

[Wang et al., 2024] https://nustm.github.io/SemEval-2024_ECAC/

Multimodal Emotion-Cause Pair Extraction in Conversations (MECPE)

- Which objective event or the subjective argument elicits (causes) which emotion?
- Task: Given a multimodal conversation (text/speech and video), extract all emotion–cause utterance pairs
- ECF 2.0 *Friends* dataset with 12K emotion–cause pairs



Multimodal Emotion Stimulus/Cause Detection

[Wang et al., 2024] https://nustm.github.io/SemEval-2024_ECAC/

MECPE: Approaches

- Majority adopted a two-set framework of emotion detection and cause detection and many used large language models (LLMs)
- Winner: instruction-based LLM; LLaVA for vision (openSMILE for audio), also face module
- Second: chat LLM, ImageBind for visual component
- Some indication that LLMs perform poorly in zero-shot and few-shot settings on emotion and cause recognition tasks
- Audio or visual modalities resulted in minimal performance improvements or a decrease in some system's

Advertisements: Image–Text Relations

[Zhang et al., 2018]

Equal But Not The Same: Understanding the Implicit Relationship Between Persuasive Images and Text

- Dataset of advertisements annotated with parallel vs. non-parallel relation between image and slogan

Advertisements: Image–Text Relations

[Zhang et al., 2018]

Equal But Not The Same: Understanding the Implicit Relationship Between Persuasive Images and Text

- Dataset of advertisements annotated with parallel vs. non-parallel relation between image and slogan
- Do the image and the text convey *the same message*?
  1. equivalent parallel vs. non-equivalent parallel (different level of detail)
  2. one ambiguous (non-parallel)
  3. opposite (non-parallel)
  4. decorative (non-parallel)

Advertisements: Image–Text Relations

[Zhang et al., 2018]

Equal But Not The Same: Understanding the Implicit Relationship Between Persuasive Images and Text

- Dataset of advertisements annotated with parallel vs. non-parallel relation between image and slogan
- Do the image and the text convey *the same message*?
  1. equivalent parallel vs. non-equivalent parallel (different level of detail)
  2. one ambiguous (non-parallel)
  3. opposite (non-parallel)
  4. decorative (non-parallel)
- Applications: advertisement interpretation/generation; fake/poor journalism detection


Current Work: ITEM

Christopher Bagdon, Aidan Combs, Roman Klinger, Carina Silberer

- Author's and reader's perspective: intended vs. perceived emotion
- Stimulus detection in image and text
- Image–text relation classification
- Preliminary work: [Khlyzova et al., 2022]

References

- Multimodal Intent Recognition – popular dataset: `https://github.com/thuiar/MIntRec`
- The MUTE Multimodal Hateful Memes dataset [Hossain et al., 2022]
- MemoSen: A Multimodal Dataset for Sentiment Analysis of Memes `https://aclanthology.org/2022.lrec-1.165.pdf`
- The Hateful Memes Detection Challenge [Kiela et al., 2020, Kiela et al., 2021]
  - Team page (3rd place): `https://github.com/rizavelioglu/hateful_memes-hate_detectron` [Velioglu and Rose, 2020]
- Surveys: [Das and Singh, 2023], [Ezzameli and Mahersia, 2023], [Lai et al., 2023]
- Multimodal EC with emojis [Illendula and Sheth, 2019]
- Mapping different emotion categorisations [Gong et al., 2024]
- Stimulus detection: ?
- Reader vs. author perspective: ?

Outline

Discourse Relations

[Prasad et al., 2008]

Discourse Relations in Text

- Discourse relations capture the logical flow of events, states, and propositions in a coherent text – how are segments of discourse logically/structurally connected?
- Description of how two segments of discourse are logically and/or structurally connected to one another
- Called also coherence relation or rhetorical relation
- *Discourse connectives* explicitly express the relations

"I never gamble too far; in particular, I quit after one try." (Expansion)
"Make sure you cool the cookies before icing, to prevent your design melting" (Contingency)
"Stir in vanilla before pouring chocolate mixture over the cereal" (Temporal)

Discourse Relations

Image–Text Relations

Different modalities can convey different meanings or information

- Required to understand non-literal meaning, e.g., metaphors, affect (sentiment, emotions)
- To fully understand situations
- To elaborate and explain in more detail
- ?

Cross-Modal Discourse Relations in Instructions

[Alikhani et al., 2019] https://github.com/malihealikhani/CITE

CITE: A Corpus of Image–Text Discourse Relations

How do visual and linguistic information contribute to a coherent and effective communication?

- Image–text coherence relations in cooking recipes
- Assumption: Images and text integrate via coherence relations
- Requires spatio-temporal inferences and inferences about the objects' place in the narrative progression
- Dataset of 2K image–text pairs (based on RecipeQA)
- Crowdsourcing (Amazon Mechanical Turk) annotations with 10 relations
  annotation procedure: in the form of binary questions

Cross-Modal Discourse Relations in Instructions

[Alikhani et al., 2019]

CITE: Examples

Roll dough into small balls and place on cookie sheet.

- *Image gives visual information about the described step*
- *Image shows the described action in progress*
- *Text describes several actions, image depicts only one*
- ***One would have to repeat the action shown in the image many times in order to complete the step***

Exemplification

In a large bowl, dice up 4 cups of apples of your choice

- *Image gives visual information about the described step*
- *Text provides specific quantities (e.g., measurements) that you would not know by just looking at the image*
- *The image shows the result of the described action*

Elaboration

Cross-Modal Discourse Relations in Instructions

[Alikhani et al., 2019]

CITE: Examples

Cut lemons in half and squeeze juice.

- *Text describes several actions, image depicts only one*

The gravy should look like this after mashing.

- *Image shows a tool used in the step but not mentioned in the text*

- *Image gives visual information about the described step*
- *Image shows the described action in progress*
- *Image shows a tool used in the step but not mentioned in the text*

Elaboration

Exemplification

- *Image gives visual information about the described step*
- *The image shows the result of the described action*
- ***You need to see the image in order to be able to carry out the step properly***

Blend the mango sorbet and the orange juice together.

Pour the mixture into a paper cup and insert a popsicle stick into it. (i used a plastic spoon in the picture.

Cross-Modal Discourse Relations in Instructions

[Alikhani et al., 2019]

CITE: What do the images contribute?

- 6% of pairs: Image required to know how to carry out the step properly
- 13%: Image shows tool not mentioned in the text
- 49%: **Image shows the result of the action**
- 21%: Image depicts an action in progress described in the text
- 21%: Text provides specific quantities one would not know just by looking at the image

References I

Alikhani, M., Nag Chowdhury, S., de Melo, G., and Stone, M. (2019).

CITE: A corpus of image-text discourse relations.

In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 570–575, Minneapolis, Minnesota. Association for Computational Linguistics.

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. (2018).

Findings of the third shared task on multimodal machine translation.

In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. (2019).

Probing the need for visual context in multimodal machine translation.

In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Das, R. and Singh, T. D. (2023).

Multimodal sentiment analysis: A survey of methods, trends, and challenges.

ACM Comput. Surv., 55(13s).

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016).

Multi30k: Multilingual english-german image descriptions.

In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

References II

Ezzameli, K. and Mahersia, H. (2023).

Emotion recognition from unimodal to multimodal analysis: A review.
Information Fusion, 99:101847.

Gong, Z., Yao, M., Hu, X., Zhu, X., and Hirschberg, J. (2024).

A mapping on current classifying categories of emotions used in multimodal models for emotion recognition.
In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 19–28.

Hossain, E., Sharif, O., and Hoque, M. M. (2022).

MUTE: A multimodal dataset for detecting hateful memes.
In Hanqi, Y., Zonghan, Y., Ruder, S., and Xiaojun, W., editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Illendula, A. and Sheth, A. (2019).

Multimodal emotion classification.
In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 439–449, New York, NY, USA. Association for Computing Machinery.

Khlyzova, A., Silberer, C., and Klinger, R. (2022).

On the complementarity of images and text for the expression of emotions in social media.
In Barnes, J., De Clercq, O., Barriere, V., Tafreshi, S., Alqahtani, S., Sedoc, J., Klinger, R., and Balahur, A., editors, *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 1–15, Dublin, Ireland. Association for Computational Linguistics.

References III

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Fitzpatrick, C. A., Bull, P., Lipstein, G., Nelli, T., Zhu, R., Muennighoff, N., Velioglu, R., Rose, J., Lippe, P., Holla, N., Chandra, S., Rajamanickam, S., Antoniou, G., Shutova, E., Yannakoudakis, H., Sandulescu, V., Ozertem, U., Pantel, P., Specia, L., and Parikh, D. (2021).
The hateful memes challenge: Competition report.
In Escalante, H. J. and Hofmann, K., editors, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 344–360. PMLR.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. (2020).
The hateful memes challenge: detecting hate speech in multimodal memes.
In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Kruk, J., Lubin, J., Sikka, K., Lin, X., Jurafsky, D., and Divakaran, A. (2019).
Integrating text and image: Determining multimodal document intent in Instagram posts.
In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4622–4632, Hong Kong, China. Association for Computational Linguistics.

Lai, S., Hu, X., Xu, H., Ren, Z., and Liu, Z. (2023).
Multimodal sentiment analysis: A survey.
Displays, page 102563.

Li, J., Ataman, D., and Sennrich, R. (2021).
Vision matters when it should: Sanity checking multimodal machine translation models.
In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

References IV

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008).

The Penn Discourse TreeBank 2.0.

In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Shen, H., Shao, L., Li, W., Lan, Z., Liu, Z., and Su, J. (2024).

A survey on multi-modal machine translation: Tasks, methods and challenges.

arXiv preprint arXiv:2405.12669.

Velioglu, R. and Rose, J. (2020).

Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge.

Wang, F., Ma, H., Xia, R., Yu, J., and Cambria, E. (2024).

SemEval-2024 task 3: Multimodal emotion cause analysis in conversations.

In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2039–2050, Mexico City, Mexico. Association for Computational Linguistics.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014).

From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions.

TACL, 2:67–78.

Zhang, M., Hwa, R., and Kovashka, A. (2018).

Equal but not the same: Understanding the implicit relationship between persuasive images and text.

In *British Machine Vision Conference (BMVC)*.