11-877 Advanced Multimodal Machine Learning **Spring 2024**

Week 3: Modality Connections

Instructors: Paul Liang and Daniel Fried *Synopsis Leads: Simran Khanuja*

Edited by Paul Liang *Scribes: Simran Khanuja*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2024/

Summary: Multimodal machine learning is the study of computer algorithms that learn and improve through the use and experience of multimodal data. It brings unique challenges for both computational and theoretical research given the heterogeneity of various data sources.

In week 3's discussion session, the class aims to dive deep into modality connections. We first discuss different dimensions in which modalities could be connected, before attempting to operationalize these definitions via metrics to discover connections in data and trained models. The following was a list of provided research probes that the class has discussed:

1. What are the reasons why modalities can be connected with each other? Come up with a taxonomy of various dimensions. Think along both statistical, data-driven dimensions and semantic, hypothesis or knowledge driven dimensions. How can we define estimators, where we can accurately quantify the presence of each type of connection given a dataset?
2. Are connections always strong and one-to-one? Reflect on what could make some cross-modal connections stronger or weaker, including many-to-many connections, ambiguity, noises, or adversarial attacks. How can we adapt our learning methods to account for these imperfections?
3. Given trained multimodal models, how can we understand or visualize the nature of connections captured by the model? What benchmarks should we design to probe the quality of learned connections?
4. How can we better learn connections that happen at a very fine-grained and compositional level? Are there new inductive biases we might need to build into vision-language connection models?

As background, students read the following papers:

1. (Required) When and why vision-language models behave like bags-of-words, and what to do about it? [Yuksekgonul et al., 2022]: This paper studies the compositionality of fine-grained connections between images and text, and proposes new ways of contrastive learning to better learn fine-grained connections.
2. (Required) Characterization and classification of semantic image-text relations [Otto et al., 2020]: This paper takes a semantic view on connections and discusses potential metrics for classifying them.
3. (Suggested) What Makes for Good Views for Contrastive Learning? [Tian et al., 2020]: This paper helps define the notion of connections from a statistical point of view, and has implications towards contrastive representation learning.
4. (Suggested) Relaxing contrastiveness in multimodal representation learning [Lin et al., 2023]: This paper studies how to learn weaker connections, when negative pairs may not be strictly negative in contrastive learning.
5. (Suggested) Non-Contrastive Learning Meets Language-Image Pre-Training [Zhou et al., 2023]: This paper explores how we can use non-contrastive methods to learn vision-language connections, to alleviate some issues of contrastive methods.
6. (Suggested) A taxonomy of relationships between images and text [Marsh and Domas White, 2003]: This paper may not be from ML but gives some very important categorizations of semantic connections between images and text.
7. (Suggested) Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data

[Hager et al., 2023]: Contrastive learning to learn connections between tabular and visual data.

8. (Suggested) CleanCLIP: Mitigating Data Poisoning Attacks in Multimodal Contrastive Learning [Bansal et al., 2023]: Improving robustness of multimodal contrastive learning.
9. (Suggested) IMAGE-MUSIC-TEXT [Barthes, 1977]: Image Music and Text is a classic book on text linguistics and hugely popular among musicians, writers, and linguists. It gives several semantic perspectives on the relationships between these 3 modalities. Worth a quick read.
10. (Suggested) A system for image–text relations in new (and old) media [Martinec and Salway, 2005]: Paper from multimedia research studying various relationships between image and text.

We summarize several main takeaway messages from group discussions below:

1 A semantic view on modality connections based on automatic metrics

Oftentimes, the connection between modalities is not apparent and might even be at odds with each other, in contexts of sarcasm or humor. Consider this example from the ACL 2023 best paper [Hessel et al., 2022], which collects a corpus of jokes from the New Yorker Cartoon Caption Contest. Models are tasked with matching a joke to a cartoon, identifying a winning caption, and explaining why a winning caption is funny. However, all state-of-the-art LLMs fall 30 points beyond human accuracy in these tasks. [Otto et al., 2020] make this observation as well, and propose the need to go beyond pre-existing measures of cross-mutual information (CMI) and semantic correlation, to capture "Status", which describes the hierarchical relation between an image and text with respect to their relative importance. Using these three quantitative metrics, they define a semantic taxonomy of multimodal connections as shown in Table 1.

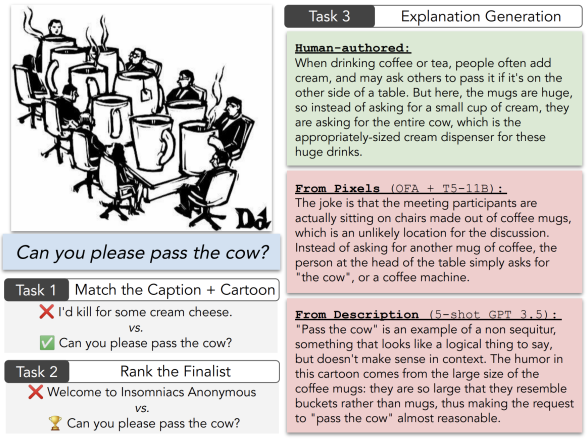

Figure 1: Taken from [Hessel et al., 2022]. They create a dataset from the New Yorker Caption Contest, to evaluate LLMs on their humor and sarcasm understanding capabilities, where modalities are mostly NOT directly connected to each other.

Another point of discussion was whether statistical analysis of modalities and their connections precedes or follows modeling methods? Modern ML techniques typically rely on training on large amounts of data and less on hand-designing features. However, for newer tasks and resource-lean scenarios, studying modality connections does gain precedence. However, as discussed in the previous week as well, statistical knowledge of the multimodal connections in data can help provide for strong inductive biases in the models, which eventually aids learning.

2 Are connections always strong and one-to-one?

As per the taxonomy defined in Table 1, modalities can be classified into only one of the categories. However, one point of discussion in class was around how language and vision can be ambiguous and the connections can be task-dependant. Therefore, it might be a good idea to extend the taxonomy such that a multi-class categorization is possible. This will help capture weak connections across modalities, which might become strong conditioned upon the end-task at hand.

Ambiguity also ties in with the discussion on whether connections are always one-to-one. We observe that indeed one-to-many or many-to-many mappings are also possible. For example, in VQA, when asked for the color of an umbrella in a picture with many umbrellas, which umbrella would one pick? This is largely due to the under-specification in the question, which is often how natural language instructions are in the real world.

Table 1: Semantic Taxonomy of Modality Connections

Semantic Class	Metric Values	Description	Use-cases
Uncorrelated	CMI=0, SC=0, STAT=0	No shared concepts or semantic backgrounds	Filter for retrieval tasks; Adblocker
Interdependent	CMI=0, SC=1, STAT=0	No shared concepts, but joint message on a higher semantic level	Adblocker; Marketing retrieval tasks
Complementary	CMI=1, SC=1, STAT=0	Modalities complement each other	Recommender systems; Cross-modal retrieval; Web search
Illustration	CMI=1, SC=1, STAT=T	Text is supplemented with an exchangeable image	Search tasks in educational settings; Text books
Anchorage	CMI=1, SC=1, STAT=1	Image is supplemented with a caption describing visual concepts	Search tasks in educational settings, e.g., definitions or explanations
Contrasting	CMI=1, SC=-1, STAT=0	Modalities complement each other, but contain contrasting details	Quality check; Filter for retrieval tasks or recommender systems
Bad Illustration	CMI=1, SC=-1, STAT=T	Given visual example is ill composed, unusual of ambiguous	Quality check; Filter for retrieval tasks or recommender systems
Bad Anchorage	CMI=1, SC=-1, STAT=1	A given caption describes details of displayed information incorrectly	Quality check; Filter for retrieval tasks or recommender systems

A suggestion in class for modeling which captures many-to-many connections was having an alternative CLIP training paradigm, where you train w/ multiple possible captions as gold captions to match an image with, instead of just one. This is similar to a distillation setup where you learn from a soft probability distribution over the set of labels, although here all captions are matched independently and the probability mass is not shared amongst alternatives.

3 How can we better learn connections that happen at a very fine-grained and compositional level?

Finally, our discussions move onto how we can better train models that can capture a diverse range of semantic connections across modalities as observed in Table 1. We discuss how contrastive learning over image-text matching pairs, as done for CLIP [Radford et al., 2021], is not enough to learn finer-grained compositional reasoning, as observed in our second paper reading [Yuksekgonul et al., 2022]. Even when the word-order in text or the patches in images are completely unordered, models can still perform extremely well at image-text retrieval. This means that pre-training on a retrieval loss cannot guarantee that our representations are any better than a bag of words model [Yuksekgonul et al., 2022]. However, as done in CLIP, contrastive learning is much easier to scale than pre-training on a generative loss, and is hence widely observed in practice.

References

Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. *arXiv preprint arXiv:2303.03323*, 2023.

Roland Barthes. *Image-music-text*, volume 6135. Macmillan, 1977.

Paul Hager, Martin J Menten, and Daniel Rueckert. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23924–23935, 2023.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*, 2022.

Zudi Lin, Erhan Bas, Kunwar Yashraj Singh, Gurumurthy Swaminathan, and Rahul Bhotika. Relaxing contrastiveness in multimodal representation learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2227–2236, 2023.

Emily E Marsh and Marilyn Domas White. A taxonomy of relationships between images and text. *Journal of documentation*, 59(6):647–672, 2003.

Radan Martinec and Andrew Salway. A system for image–text relations in new (and old) media. *Visual communication*, 4(3):337–371, 2005.

Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. Characterization and classification of semantic image-text relations. *International Journal of Multimedia Information Retrieval*, 9:31–45, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

Jinghao Zhou, Li Dong, Zhe Gan, Lijuan Wang, and Furu Wei. Non-contrastive learning meets language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11028–11038, 2023.