**11-877 Advanced Multimodal Machine Learning** **Spring 2023**

# Week 3: Modality Connections

*Instructors: Louis-Philippe Morency and Paul Liang* *Synopsis Leads: Durvesh Malpure, Haofei Yu*

*Edited by Paul Liang* *Scribes: Suzzane Nie, Sean Chang*



Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2023/

**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?
2. What is the difference between interactions and connections? How can you formally define the two and their corresponding differences?
3. How to discover modality connections in the data when given limited data or large-scale data? Are connections always strong and one-to-one?
4. What formalism or framework could be used to formalize cross-modal connections? How can we subsequently define estimators, where we can accurately quantify the presence of each type of connection given a dataset? How much knowledge of each modality do we need in order to estimate modality connections?
5. Linking back to week 2's discussion on heterogeneity: How would you relate the concepts of heterogeneity and connections?
6. How is heterogeneity affecting the study of crossmodal connections and inversely, how connections should be taken into consideration when heterogeneity is studied? Are connections also present in homogenous settings?

As background, students read the following papers:

1. (Required) 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.
2. (Required) Characterization and Classification of Semantic Image-text Relations [Otto et al., 2020]: This paper instead takes a semantic view on connections and discusses potential metrics for classifying them.
3. (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.
4. (Suggested) Order-Embeddings of Images and Language [Vendrov et al., 2015]: Learning representations that respect semantic order relationships.
5. (Suggested) Corpus-based Learning of Analogies and Semantic Relations [Turney and Littman, 2005]: Learning semantic relations from data.
6. (Suggested) Multimodal Neurons in Artificial Neural Networks [Goh et al., 2021]: Visualizing connections learned in neural networks.

7. (Suggested) Learning Aligned Cross-Modal Representations from Weakly Aligned Data [Castrejon et al., 2016]: Alignment with only weak connections.
8. (Relevant) Toward Causal Representation Learning [Schölkopf et al., 2021]: General review paper on causality + ML, useful for thinking about causal relationships and how we can discover them.
9. (Relevant) A Review of Relational Machine Learning for Knowledge Graphs [Nickel et al., 2015]: Review paper on semantic relations, graph formalisms, and relational ML.
10. (Relevant) Image-Music-Text [Heath et al., 1977]: 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. Not an ML paper but worth a quick read.
11. (Relevant) Unsupervised Alignment of Embeddings with Wasserstein Procrustes [Grave et al., 2019]: How can we do alignment without knowing the explicit connection pairs?
12. (Relevant) On Variational Bounds of Mutual Information [Poole et al., 2019]: More details on information-theoretic formalisms of statistical connections.
13. (Relevant) 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 taxonomy of modality connections

Figure 1: When seeing a muffin image, human beings would naturally connect it with its smell and its taste.

Many objects in our real world have multimodal components and these components are inherently connected with each other. For example, videos include audio, image, and language modalities that are all connected with the other. Moreover, multimodality is related to how human beings actually think and feel. For example, when looking at an image of a delicious muffin, we naturally create connections between the image and its taste and smell in our brains. These connections between modalities enable us to integrate different multiple partially-observed views into one joint distribution that is the most complete representation of our real world.

We summarize a taxonomy of the different ways in which modalities can be connected with each other in Table 1, categorized broadly into statistical and semantic perspectives. Statistical connections can information-theoretic [Tian et al., 2020], feature-based [Du et al., 2021], and symmetric or asymmetric [Martinec and Salway, 2005]. When considered from the semantic view, semantic hierarchies, causal, logical, or knowledge-driven connections can also exist.

## 1.1 Modality connections and heterogeneity

What is the relationship between connections and heterogeneity? We observe that heterogeneity is generally at the surface level, as seen in week 2's discussion, while the connections are much more complex, requiring the learning of feature spaces on modalities and joint representations across modalities [Tian et al., 2020]. Therefore, modeling heterogeneity can be seen as a 'precursor' for understanding connections [Tian et al., 2020]. Furthermore, heterogeneity can be inferred using only the marginals $p(x_i)$ or $p(x_i, y)$, but understanding connections requires knowing the joint distribution across modalities $p(x_1, x_2)$. Finally, modalities can be homogeneous and connected (i.e., semantic connections between different sentences), and can also be heterogeneous and unconnected.

## 1.2 Modality connections and interactions

Connections exist in the modalities and data itself, while interactions arise from modeling 2 modalities for an end task. However, modalities may not need to be connected for them to interact. For example, consider a car's dead battery or a blocked fuel pump. Ordinarily, we assume that battery death and fuel pump blockage

| Perspective | Category | Definition |
|---|---|---|
| Statistical | Information Connection | The distributions of different modalities have overlapped information, with some parts of each distribution being paired with each other. |
| | Feature Connection | Different modalities provide diverse features in the representation space, with some features paired with features in another modality. |
| | Asymmetric Connection | Modality connections can be unbalanced and asymmetrical, with connections in many modalities dominated by one specific modality statistically. |
| Semantic | Hierarchical Connection | Modality connections can be hypernym-hyponym relationships. Hierarchical semantic connections can be defined by considering one modality as the hypernym and another modality as the hyponym. |
| | Causal Connection | When considering each modality as a causal variable, data in multiple modalities can have causal relationships with each other in a structural model if they are statistically dependent. |
| | Knowledge Connection | Concepts in multiple modalities are shared and connected in various semantic relationships (e.g., function or use). |

Table 1: A taxonomy of modality connections spanning statistical and semantic perspective, and across various granularities.

are independent events, but knowing that the car fails to start, if an inspection shows the battery to be in good health, we can conclude that the fuel pump must be blocked. Therefore, in these common cause scenarios, modalities that are initially unconnected become connected and interact when the label is observed. Similarly, modalities can be connected but do not interact, such as semantic connections present in images and text that describe task-irrelevant information.

## 2 Modeling and quantifying modality connections

We list some ways that enable training models to discover modality connections.

**Contrastive learning** at different granularities can be used to capture strong or weak connections. Contrastive learning requires domain knowledge to define semantic positive and negative pairs, typically via data augmentation [Tian et al., 2020]. Ideally, the augmented positive and negative views should overlap as much as the actual degree of connections in the data (i.e., redundancy).

**Large-scale noisy alignment**: CLIP [Radford et al., 2021] used 400 million image-text pairs collected from the Internet and pre-trained the model based on contrastive learning. Even though noisy connections exists in large datasets, these costs are outweighed by the benefits of data diversity. As long as there is enough data, useful connections can also be discovered from noisy relations.

In Table 2, we summarize some methods that have been discussed to measure modality connections. For `cosine similarity` and `wasserstein distance`, it focuses on the representation space and representation distance between different modalities is determined by a combination of model initialization and contrastive learning optimization [Liang et al., 2022]. Considering from a probability perspective, `mutual information` can be one possible way to describe the modality connections. While this metric is not specific to a typical downstream task, `loss different` is typically related to one downstream task. The drop or improvement of downstream task loss gives a clear signal of whether one modality is critical for the task's performance. Moreover, quantification can gain information from the model results. In transformer architecture, `attention score` between different modalities can be explained as a connection between different modalities. Compared

| Metric Name | Definition | Explanation |
|---|---|---|
| Cosine Similarity | $\mathbf{S}_{i,j} = \text{Normalize}(\text{Enc}(x_{A,i})) \cdot \text{Normalize}(\text{Enc}(x_{B,j}))$ | It computes representation similarity between different modalities. High cosine similarity between similarilty in feature space. |
| Wasserstein Distance | $\mathbf{W}_{i,j} = \min_{\mathbf{P} \in \mathbf{P_n}} \sum_{i,j} P_{ij} \|x_i - y_j\|_2^2$ | It computes the squared Wasserstein distance between two sets of points $\mathbf{X}$ and $\mathbf{Y}$. If we consider each data in one modality as a point, Wasserstein distance help us to measure the distance between sets of points when having no explicit knowledge of the connection pairs. |
| Mutual Information | $\mathbf{I}(X_A; X_B) = \mathbb{E}_{p(x_A, x_B)} \log \frac{q(x_A \mid x_B)}{p(x_A)}$ | Information theory can be used to describe the statistical connections between different modality distributions. |
| Loss Difference | $\Delta\mathbf{L}_{i,j} = \|\mathbf{L}(x_{A,i}, x_{B,j}) - \mathbf{L}(\texttt{[MASK]}, x_{B,j})\|_p^2$ | If we mask one modality and its performance drops a lot, the modality probably has strong connections with the others. |
| Attention Score | $\mathbf{A}_{i,j} = \sum_{\text{Layer}=1}^{N} \frac{1}{2}\text{Attn}(\mathbf{Q}_{A,i}, \mathbf{K}_{B,j}) + \frac{1}{2}\text{Attn}(\mathbf{Q}_{A,j}, \mathbf{K}_{B,i})$ | If modalities share a dimension like time series, we can use attention scores as a metric to test where they attend to each other. |
| Human Annotation | $\mathbf{R} = \text{Rank}((A_1, B_1), \ldots, (A_n, B_n))$ | Human annotators can be tasked to judge semantic connections such as correspondence, causal, or temporal relations. Each annotator should rank and check which pairs are more strongly related. |

Table 2: Metrics for quantifying modality connections. $A$ and $B$ represents two modalities. $h$ represents the representations given by model. $x$ represents the input data for the model. $y$ indicates the labels. $f$ represents the multimodal model combining two modalities.

with the attention score which is gained from the model's perspective, more accurate connection quantification can be gained from the human perspective. We can easily design a ranking task for human annotations to rank and give a data-level ranking score based on `human annotation`. Attention-based connection metric is more straightforward and easy to do but not accurate while a human annotation is explainable and accurate while being expensive and hard to implement.

# References

Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2940–2949, 2016.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Yue Wang, Yang Yuan, and Hang Zhao. Modality laziness: Everybody's business is nobody's business. 2021.

Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030. https://distill.pub/2021/multimodal-neurons.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, 2019.

Stephen Heath et al. Image-music-text. *London: Fontana*, pages 78–118, 1977.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning, 2022. URL https://arxiv.org/abs/2203.02053.

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.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2015.

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.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

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.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 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.

Peter D Turney and Michael L Littman. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60:251–278, 2005.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.