9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Aggregate, Integrate and Align to Embed Everything: A Multi-Modal Framework for Measuring Cultural Dynamics

Deep Learning, Computational Social Science, Cultural Dynamics, Methodological Framework, Multi-modality

Extended Abstract

A deluge of digital content is generated daily by web-based platforms and sensors that capture digital traces of communication and connection, and complex states of society, the economy, and the world. In parallel, historical data is increasingly digitized, and access to millions of books, images, historical documents, patents is easier than ever before. Web archives such as Reddit and Wikipedia offer massive samples of structured textual data with rich labels. Images can be scraped from google search results, with aligned captions, and graphs of scientific citations and financial transactions and trade data are readily available.

Indeed, many of these datasets have become the training data for large pretrained models of language and images. These models are often used for various predictive downstream tasks, such as image classification or question answering. What we propose is instead to *aggregate* each of these datasets by a cultural label representing a fundamentally multi-modal entity, such as a person, place, or object. After aggregating multi-modal data by cultural label, we then *integrate* the data across modalities by learning joint representations. For an author, for example, it could be include not only the content of their books, but also autobiographical information, their position within a network of letter correspondences, and their associations by others on social media. Similarly, a cultural city representation might include tabular demographic information, economic indicators, as well as images of the city, and text (tweets, posts, Wikipedia articles) associated with the city. By jointly learning these representations, these cultural samples can now be used for a variety of tasks, from the elicitation of new cultural data (e.g., Is London culturally closer to Paris or New York?) to projections on cultural axes of interest (which author has rendered the most violent characters?) to identifications of the loci of cultural divergence. We refer to these aggregated and integrated multi-modal representations as digital twins of culture.

Once we have a set of cultural objects and their representations, it is possible for us to align different cultural worlds. For example, we could align artists over time, and observe their movements in style space, analogous to observing semantic drift in diachronic word embeddings **[1]** . Alignment of embedding spaces need not only occur across time; and other axes such as language, or location can also be used. In this way, it is possible to compare how different cultures represent the same concept **[2]**.

This process of aggregating, integrating, and aligning provides us with a framework to compare different cultural samples, across axes of time, language, or location. With aligned embedding spaces, we can now measure distances between cultural objects within and between worlds, similar to the distance-based measures described in **[1]**. It is also possible to project words onto different cultural axes, as described in **[2]**, where the authors project words (such as sports, or music genres) onto axes of gender, class, and race to track the changing relationships between cultural dimensions within society.

We argue that such a framework would help model the factors and extent of cultural change, and the distinctions and biases underlying specific cultural representations produced by Silicon Valley companies from selective and often privileged consumers underlying training data. Such a framework could also lend itself to causal analysis, by observing the positions of cultural objects before and after experiments are conducted.

After we have aggregated, integrated, and aligned our data and identities, we have opened ourselves to a world of potential analysis. Diachronic ordinal embedding tasks will allow us to measure how relationships between triplets change over time, and we can measure movements of entities along the axes of our choosing. Previous work in unraveling cultural artifacts from word embedding models **[2, 3, 4]** will be supercharged and vastly extended with multi-modal representations. Measuring culture **[5]** with such neural approaches allows for complex patterns to emerge that may be missed with simpler approaches, to identify complex conflicts between embedded data, and allow for the diagnosis of cultural bias in the representations underlying modern recommendation engines that extend well beyond their training domains, and may be irrelevant or exercise unanticipated and unwanted cultural influence.

We note here the limitations of using a purely quantitative approach to measuring culture. Keeping in line with work on Computational Grounded Theory **[6]**, any kind of analysis must use human knowledge and hermeneutic skills to augment computational analysis. While the framework we describe has high potential to uncover new relationships among cultural artefacts as well as their dynamics, they must be guided by theory. Indeed, work on the Geometry of Culture **[2]**, for example, uses neural embedding models to investigate theories of social class: without the qualitative knowledge of different theories of class and how they might be operationalised in an embedding space, the study would not be possible.

We also note that while there has been significant efforts in explainability and interpretability of neural models, the pitfalls of such models have been well documented, along the axis of both text **[7]** and images **[8]**. However, the inherent biases encoded in these models allow for analysis of the ways humans construct biases in image and text. Indeed, while we do not recommend using such models with confidence on downstream tasks of societal impact, we believe in the strong potential of such models for social scientific and cultural reflections.

Conclusion

Cultural markers for identities of diverse kinds are fundamentally multi-modal and complex. Measuring relationships between groups of identities in a set of cultural subjects and objects is difficult in settings that do not account for multi-modality and emergent behaviour. Neural models have been shown to handle diverse data sources and account for complex behaviour. By representing cultural identities in high-dimensional cultural spaces, we can use ordinal embedding tasks and distance measures to study relationships between cultural acts and artifacts that allow comparison, bias identification, and potential correction or speciation. We propose a method of aggregation, integration and alignment: embed everything! to study the dynamics of culture.

References

[1] Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1489-1501. 2016.

[2] Kozlowski, Austin C., Matt Taddy, and James A. Evans. "The geometry of culture: Analyzing the meanings of class through word embeddings." American Sociological Review 84, no. 5 (2019): 905-949.

[3] Nelson, Laura K. "Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century US South." Poetics 88 (2021): 101539.

[4] Peng, Hao, Qing Ke, Ceren Budak, Daniel M. Romero, and Yong-Yeol Ahn. "Neural embeddings of scholarly periodicals reveal complex disciplinary organizations." Science Advances 7, no. 17 (2021): eabb9004.

[5] Mohr, John W., Christopher A. Bail, Margaret Frye, Jennifer C. Lena, Omar Lizardo, Terence E. McDonnell, Ann Mische, Iddo Tavory, and Frederick F. Wherry. Measuring culture. Columbia University Press, 2020.

[6] Nelson, Laura K. "Computational grounded theory: A methodological framework." Sociological Methods Research 49, no. 1 (2020): 3-42.

[7] Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?." In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 610-623. 2021.

[8] Crawford, Kate, and Trevor Paglen. "Excavating AI: The politics of images in machine learning training sets." Ai Society 36, no. 4 (2021): 1105-1116.