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

Bhargav Srinivasa Desikan
EPFL
Lausanne, Switzerland
bhargav.srinivasadesikan@epfl.ch

James Evans
University of Chicago
Chicago, USA
jevans@uchicago.edu

Abstract

The massive sensing of cultural action served through online platforms, and the aggregation of vast samples of cultural artifacts (e.g newspapers, images) has provided us with an unprecedented opportunity to embed cultural acts and artifacts in high dimensional representation space for analysis, comparison, and the efficient interactive elicitation of new cultural data, as from ordinal embedding. Existing research on culture and deep learning often embeds a single modality - text, images, more infrequently, networks and tables, rarely small-sample qualitative cultural associations, and never together. In this position paper, we propose that cultural data, their creation, proliferation, and consumption should be studied together and ongoingly elicited in the context of multi-modal neural representations. In short: embed everything! Cultural representations that align text, images, graphs, tables, elicited qualitative associations, and more can capture and compare complex cultural associations and reveal biases within dominant representations that neither capture nor serve those less present in training data. Within a high-dimensional representation space, we can use distance measures to study the full-spectrum influence of events, the dynamics of cultural change, and identify clear distinctions between representations native to different populations.

Introduction

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 In parallel, historical data is increasingly digitized, and access to millions of books, images, historical documents, patents is easier than ever before.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 pre-trained 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 (Hamilton, Leskovec, and Jurafsky 2016). 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 (Kozlowski, Taddy, and Evans 2018).

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 (Hamilton, Leskovec, and Jurafsky 2016). It is also possible to project words onto different cultural axes, as described in (Kozlowski, Taddy, and Evans 2018), 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. In the remainder of this position paper, we discuss early and related work in using such frameworks to extract cultural relationships; lay out a series of steps to operationalize the framework for samples of cultural acts and artifacts; discuss tasks and operations that can profitably be performed on these aligned spaces.

Related Work

The notion of creating a high-dimensional representation of actions and artifacts has been popular for capturing semantic representations among words. The notion of cultural spaces representing conceptual objects has its origins in the structuralism of early 20th century linguistics (De Saussure 1916), and mid-20th Century psychology (Osgood, Suci, and Tannenbaum 1957) and anthropology (Lévi-Strauss 1963; **?**), but large-scale cultural data was arguably first embedded and analyzed in high-dimensional space with the Latent Semantic Indexing (LSI) algorithm (Landauer and Dumais 1997). The notion of vector representations for words popularised by it has since been used extensively, with methods moving from linear algebra and matrix factorization to neural network approximations and extensions. (Mikolov et al. 2013) introduced *word2vec*, which furthered widespread use of such vectors and popularised embedding. *Glove* ((Pennington, Socher, and Manning 2014)) used both global and local occurrences to construct the cultural vector spaces, and *fastText* (Joulin et al. 2016) used them to classify text. These embedding methods were static in nature, but the elegance of high dimensional representations lies in the way they can be further modified: having their geometries entangled, morphed, and warped: work by (Faruqui et al. 2014), (Jauhar, Dyer, and Hovy 2015), (Levy and Goldberg 2014) are all examples, where syntactic structure or lexical information was used to change the nature of the embedding spaces; either to shift the associational distance between words that would not normally lie close to one another, or include other (e.g., syntactic) information in the embedding. These spaces can carry desirable and undesirable cultural bias (Caliskan, Bryson, and Narayanan 2017), which can be compensated for and corrected (Bolukbasi et al. 2016).

Static word embedding models are the early prototypes of our *embedding everything* framework, and provide inspiration for our approach. Work by (Kozlowski, Taddy, and Evans 2019) used word embedding models trained on different historical time periods to study how different cultural dimensions related to each other. Word embedding models also popularized analogy tasks, enforcing a semantic algebra atop the geometry of the embedding. In the work by (Kozlowski, Taddy, and Evans 2018), these linear algebra operations are used to project words (generalised to concepts) onto axes to score words along these dimensions. This allows us to ask questions such as - where does basketball lie on the cultural axis of gender, and how has it changed over time? Another landmark paper in manipulating embedding spaces is the work by (Hamilton, Leskovec, and Jurafsky 2016), which used Procrustes alignment to measure distances between diachronic embedding spaces. Such manipulations allow us to measure cultural associations and differences in mathematically straightforward and conceptually elegant ways. An example of using such approaches for linguistic and cultural explorations can be seen in (Thompson, Roberts, and Lupyan 2020), where word embeddings are used to demonstrate how meanings of common words vary in ways that reflect user culture, history and geography.

Today, state-of-the-art embedding based approaches for natural language are more often contextual and dynamic, using large pre-trained language models such as BERT (Devlin et al. 2018) or the GPT family (Brown et al. 2020), built using variants of the Transformer model (Vaswani et al. 2017). Embeddings produced by these models are equally amenable to cultural analysis. It has been widely noted that these large models are prone to problems such as toxicity and biased training data (Bender et al. 2021), and that there have been efforts to de-bias the contextual stereotypes learned (Bartl, Nissim, and Gatt 2020). However, these biases reflect the social worlds on which the model was trained – for us to explore the nature of cultural associations, we want to work with the original models. Indeed, these large models contain spaces of cultural relations culture within them. There have already been attempts to use these embeddings and associated high dimensional spaces for knowledge discovery, such as work by (Tshitoyan et al. 2019) where they use unsupervised word embeddings to capture latent knowledge from material science literature for materials prediction. Combined corpora from chemistry and material science could approximate the knowledge space well enough to be able to uncover the underlying structure of the periodic table and structure-function relationships in materials. Exploring the topologies of these space can allow us to learn about the history and structure of ideas. Consider (Linzhuo, Lingfei, and James 2020), which demonstrates how centralized collaboration can reduce the space of ideas, and how these patterns generalize to other contexts in modern scholarship and science. Such research shows the relations that emerge in constructed high-dimensional spaces align with cultural categories and meanings.

All the examples above deal with single-modality embedding models, and allow us to explore cultural relationships within this modality. Theories of embodied cognition and multi-modal culture have motivated multi-modal embedding systems, which jointly learn representations of inherently multi-modal cultural objects from multiple input spaces. The earliest of such models used images in conjunction with text, aligned either with image captions or labels (Hwang and Grauman 2012; Rasiwasia et al. ; Gong et al. 2014; Socher et al. 2014), and are also referred to as *grounded models*. Approaches in the embodied cognition paradigm have used sources such as color to embed further information (Guilbeault et al. 2020; **?**), demonstrating that color adds important cognitively processed associational information to word representations. Today, state of the art multi-modal models for learning cultural data are built using Transformer (e.g, VilBERT (Lu et al. 2019)) based models, and the now famous image-to-text generation capacities of DALL-E come from such grounded, multi-modal representations (Radford et al. 2021; Ramesh et al. 2022).

Construction of such multi-modal models are still motivated by their increased performance in downstream tasks, such as classification or image generation. Shifting our focus to cultural objects and entities, opportunities for alignment increase. Specific text associated with an entity, such as e-mail transcripts of conversations between members of a company, or tweets associated with it, can add further socially situated information to a representation. Methods such as network and graph embeddings (Grover and Leskovec 2016), knowledge graph embeddings (Wang et al. 2017), and hyperbolic graph embeddings (Chami et al. 2019) all offer us methods of casting relationships in high dimensional spaces that capture variation with distance. There have been approaches that align Transformer based models with graph embeddings, such as TaBERT (Yin et al. 2020) or VGCNBERT (Lu, Du, and Nie 2020). High dimensional spaces are also modeled as architectures of cognition (Kelly et al. 2019), and there have been early attempts to include general purpose multi-modal embedding approach such as *X2vec* (Grohe 2020).

Related literature has become increasingly rich with attempts to create multi-modal representations, as well as approaches to extract cultural information and relationships from these embeddings. However, these approaches have not been integrated in a theoretically grounded manner to perform cultural analysis, and measure the effects of events on cultural representations. We spend the next two sections describing potential approaches to create multi-modal, high-dimensional representations of cultural entities (digital twins), and the nature of relationships and bias we can uncover with these representations.

Embedding Everything: Multi-Modal Cultural Representations

Cultural associations and relationships are complex and inherently multi-modal. To capture this, our representations of cultural entities must also be complex and multi-modal. We propose a three-step process to compare between and within cultural entities along multiple axes - aggregate, integrate, align. We illustrate this approach with three examples of cultural subjects/objects - artists, cities, and fruits.

Aggregate

The first step is to aggregate each modality by label. For artists, this could be: a dataset of artwork, a dataset of artist biographies, and a network of influences and associations. The first step would be to assign each modality to different cultural identities in the set. With every identity in the set linked to different data samples, now we aggregate these representations. This can be done in multiple ways. One is to create a representation for each data point - for example, each artwork, or each description of the artist. We then choose an aggregation technique, such as a weighted average. We then have one representation for each modality. In the case of artists, each artist could have one aggregated style vector or aggregated color vector, as described in (Srinivasa Desikan, Shimao, and Miton 2022), along with a document vector of the artist biographical data or personal papers and interpersonal networks. We refer to this step of collecting all datapoints per modality for each cultural identity as *aggregation*. We note that only such aggregated representations would adequately capture complex intersectional relationships that cultural entities lie in.

Integrate

Once we have embedded each identity in our set within different modalities, we can integrate our representation. One approach to integration is to simply concatenate the representations of each modality. In this case, we would have separate representations for each datapoint associated with our object. For example, for fruits–a category at the intersection of structure (biological seed-bearing entities) and function (humans eat them), let us use images of fruits, their position in the hierarchy of biological classification (i.e genus), their chemical composition (e.g mol2vec (Jaeger, Fulle, and Turk 2018)), and possibly their co-purchase within a shopping basket (e.g., Nielsen shopping data). Integration here would involve concatenating an aggregate image embedding, topological embedding, chemical representation, and co-purchase embedding and mapping it to the associated fruit. Another approach is to perform a joint-learning of our representation. Such an approach would involve training the model with a downstream task or objective - for our fruits, this could be to predict the rate of consumption by region. We refer to this step of creating one multi-modal representation for each identity as integration.

Align

After each identity in our set of cultural subjects or objects has a corresponding embedding, we can begin to measure relationships within a domain. For example, if we have aggregated and integrated representations of cities, we would now be able to perform ordinal embedding tasks, and analogy operations. This allows us to ask questions such as: what is the equivalent of Pittsburgh in Spain, or pairs of cities in South East Asia are culturally similar, but topologically different. These comparisons can only be made if each of the datapoints associated with the entity is collected simultaneously, as the representations are only useful when compared with each other. It is here that the Procrustes alignment trick from (Hamilton, Leskovec, and Jurafsky 2016) comes in handy, and allows us to align embedding spaces across modalities such as time or language, or we can use deep learning to optimize alignments (Milbauer, Mathew, and Evans 2021). This allows us to ask questions such as - in the last ten years, which Australian city became economically closer to Singapore? It also allows us to elicit new cultural information and identify its difference from the majority within the models. In this way, even ”small” cultural data can be used to diagnose and potentially de-bias ”large” cultural data within the original samples and models.

Measuring Dynamics and Processes of Cultural Change

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 (Kozlowski, Taddy, and Evans 2018; Nelson 2021; Peng et al. 2021) will be supercharged and vastly extended with multi-modal representations. Measuring culture (Mohr et al. 2020) 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.

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

Bartl, M.; Nissim, M.; and Gatt, A. 2020. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 1–16.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, 4349–4357.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.

Chami, I.; Ying, Z.; Ré, C.; and Leskovec, J. 2019. Hyperbolic graph convolutional neural networks. In *Advances in neural information processing systems*, 4868–4879.

De Saussure, F. 1916. Nature of the linguistic sign. *Course in general linguistics* 1:65–70.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Faruqui, M.; Dodge, J.; Jauhar, S. K.; Dyer, C.; Hovy, E.; and Smith, N. A. 2014. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

Gong, Y.; Ke, Q.; Isard, M.; and Lazebnik, S. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision* 106(2):210–233.

Grohe, M. 2020. word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 1–16.

Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864.

Guilbeault, D.; Nadler, E. O.; Chu, M.; Sardo, D. R. L.; Kar, A. A.; and Srinivasa Desikan, B. 2020. Color associations in abstract semantic domains. *Cognition* 201:104306.

Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Hwang, S. J., and Grauman, K. 2012. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International journal of computer vision* 100(2):134–153.

Jaeger, S.; Fulle, S.; and Turk, S. 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling* 58(1):27–35.

Jauhar, S. K.; Dyer, C.; and Hovy, E. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *proceedings of the 2015 conference of the north American chapter of the Association for Computational Linguistics: human language technologies*, 683–693.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kelly, M. A.; Arora, N.; West, R.; and Reitter, D. 2019. High dimensional vector spaces as the architecture of cognition.

Kozlowski, A. C.; Taddy, M.; and Evans, J. A. 2018. The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*.

Kozlowski, A. C.; Taddy, M.; and Evans, J. A. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review* 84(5):905–949.

Landauer, T. K., and Dumais, S. T. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104(2):211.

Lévi-Strauss, C. 1963. Structural analysis in linguistics and in anthropology. *Semiotics-An Introductory Anthology* 110–128.

Levy, O., and Goldberg, Y. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 302–308.

Linzhuo, L.; Lingfei, W.; and James, E. 2020. Social centralization and semantic collapse: Hyperbolic embeddings of networks and text. *Poetics* 101428.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 13–23.

Lu, Z.; Du, P.; and Nie, J.-Y. 2020. Vgcn-bert: augmenting bert with graph embedding for text classification. In *European Conference on Information Retrieval*, 369–382. Springer.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Milbauer, J.; Mathew, A.; and Evans, J. 2021. Aligning multidimensional worldviews and discovering ideological differences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4832–4845.

Mohr, J. W.; Bail, C. A.; Frye, M.; Lena, J. C.; Lizardo, O.; McDonnell, T. E.; Mische, A.; Tavory, I.; and Wherry, F. F. 2020. Measuring culture. In *Measuring Culture*. Columbia University Press.

Nelson, L. K. 2021. Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century us south. *Poetics* 88:101539.

Osgood, C. E.; Suci, G. J.; and Tannenbaum, P. H. 1957. *The measurement of meaning*. Number 47. University of Illinois press.

Peng, H.; Ke, Q.; Budak, C.; Romero, D. M.; and Ahn, Y.-Y. 2021. Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Science Advances* 7(17):eabb9004.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, 251–260.

Socher, R.; Karpathy, A.; Le, Q. V.; Manning, C. D.; and Ng, A. Y. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* 2:207–218.

Srinivasa Desikan, B.; Shimao, H.; and Miton, H. 2022. Wikiartvectors: Style and color representations of artworks for cultural analysis via information theoretic measures. *Entropy* 24(9):1175.

Thompson, B.; Roberts, S. G.; and Lupyan, G. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour* 4(10):1029–1038.

Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; and Jain, A. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* 571(7763):95–98.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Wang, Q.; Mao, Z.; Wang, B.; and Guo, L. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering* 29(12):2724–2743.

Yin, P.; Neubig, G.; Yih, W.-t.; and Riedel, S. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314*.