GraviTIE: Exploratory Analysis of Large-Scale Heterogeneous Image Collections

Sean T. Yang
University of Washington
Seattle, Washington
tyyang38@uw.edu

Luke Rodriguez
University of Washington
Seattle, Washington
rodriglr@uw.edu

Jevin D. West
University of Washington
Seattle, Washington
jevinw@uw.edu

Bill Howe
University of Washington
Seattle, Washington
billhowe@cs.washington.edu

ABSTRACT

We present GraviTIE (Global Representation and Visualization of Text and Image Embeddings, pronounced "gravity"), an interactive visualization system for large-scale image datasets. GraviTIE operates on datasets consisting of images equipped with unstructured and semi-structured text, relying on multi-modal unsupervised learning methods to produce an interactive *similarity map*. Users interact with the similarity map through pan and zoom operations, as well as keyword-oriented queries. GraviTIE makes no assumptions about the form, scale, or content of the data, allowing it to be used for exploratory analysis, assessment of unsupervised learning methods, data curation and quality control, data profiling, and other purposes where flexibility and scalability are paramount. We demonstrate GraviTIE on three real datasets: 500k images from the Russian misinformation dataset from Twitter, 2 million art images, and 5 million scientific figures. A screencast video is available at https://vimeo.com/310511187.

KEYWORDS

GraviTIE; MultiDEC; Visualizing Large-Scale Image Collections

ACM Reference Format:
Sean T. Yang, Luke Rodriguez, Jevin D. West, and Bill Howe. 2019. GraviTIE: Exploratory Analysis of Large-Scale Heterogeneous Image Collections . In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308558.3314142

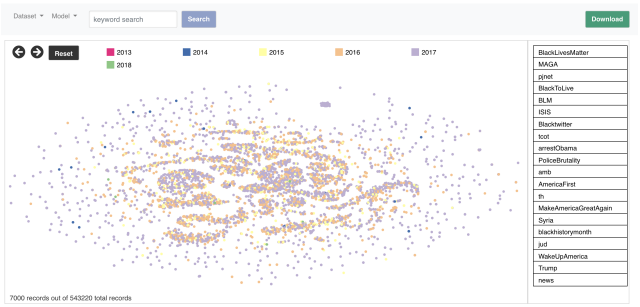


Figure 1: Main interface of GraviTIE. GraviTIE (Global Representation and Visualization of Text and Image Embedding) is an interactive web visualization application and discovery engine for large image-text datasets.

1 INTRODUCTION

Countless images exist on the web. These include photographs, memes, art images, scientific figures, and more. Analysis of these image datasets increasingly requires the application of large-scale machine learning methods to answer even relatively straightforward questions. For example, the dataset recently released by Twitter[1] contains rich information about Russian influence, but even conceptually simple questions such as "Do images posted from Russian accounts tend to differ from those posted by other accounts?" require a combination of unsupervised learning (to determine image similarity), NLP methods (to contextualize the image), and structured queries (to compare images by source and year).

These analysis tasks have several defining characteristics. First, tasks increasingly require some form of unsupervised learning. Ground truth labels or other obvious sources of supervision are rarely available. In these situations, interactive, qualitative exploration of the results is the only option. Second, the data is *multimodal*: each record consists of some combination of an image, free text, and a set of structured attributes. Each of these elements may be of arbitrary size, or altogether missing. Third, the datasets are large, making direct-browsing and main-memory approaches infeasible. For example, the Twitter dataset contains 9 million tweets with over 543k images, the Artstor dataset contains more than 2 million high-quality images of artwork along with text descriptions, and the Viziometrics [8] dataset consists of over 9 million figures from the scientific literature, along with their captions.

[1] https://about.twitter.com/en_us/values/elections-integrity.html#data

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3314142

In this demonstration, we present GraviTIE (**G**lobal **R**epresentation **a**nd **V**isualization of **T**ext and **I**mage **E**mbeddings, pronounced "gravity"), an interactive visualization system for large-scale image-text datasets. Fig. 1 shows the interface of GraviTIE.

For pre-processing, GraviTIE takes as input a large set of image-text pairs. It learns an embedding for the pair with clustering objective by running MultiDEC[14] on the text and image, such that the two models tend to produce the same distribution. The properties of MultiDEC are tightly coupled to the design of the interface. In particular, this model 1) accommodates multi-modal heterogeneous data while 2) maintaining reasonable performance at scale. The learned embedding is then dimension-reduced to two dimensions for display purposes.

The GraviTIE interface features a large similarity map to display thousands of nodes, where each of the node represents an image positioned by a multi-modal vector embedding. Similar images are clustered together on this similarity map which helps the users efficiently navigate a large number of images. Users can refine the current view using a combination of free-text queries (with appropriate operators “must include”, “might contain” or “exclude”) and structured query filters. Only a sample of the images is displayed to ensure performance at scale and to afford hover interactions with individual images; displaying all images would lead to extremely dense regions with significant occlusion. Users can pan, zoom, filter, and search iteratively, gradually refining the set of displayed images and inspecting the images in each cluster via hover interactions at any time. Each view is associated with a unique url (i.e. RESTful), making the system stateless and programmable, allowing refinements to be saved and shared with collaborators. The data displayed in each view can be directly downloaded to provide custom curated subsets for specific tasks. For example, an art historian can construct a dataset of impressionist paintings by selecting particular clusters and filtering for a particular time period. The RESTful API design also affords additional downstream applications to be developed using the search and download functions.

GraviTIE is designed to be adaptable to a number of applications across multiple modalities and communities. In its current form, art historians are using GraviTIE to explore influence patterns in art, computer scientists are using it to qualitatively assess unsupervised learning methods, and social scientists are using it to understand the behavior of Russian actors on Twitter.

2 RELATED WORK

Analysis and visualization of large scale images have been receiving great attention from researchers. Hochman et al. [7] explore over 2 million photos on Instagram by sorting the images in terms of their properties (e.g., hue or create time) and showing the complete image sets of different cities to reveal the local culture and social patterns. The T-SNE Map [2], one of Google Arts & Culture experiments by Diagne et al., is a 3D interactive platform created by computing the visual similarity of art works and grouping the art works with t-SNE algorithm. PixPlot[3] extends the idea by applying UMAP [10] algorithm which can scale large data set and cluster million of data points and allows users to implement custom datasets. These applications provide a simple platform to explore large scale image sets, but do not support query and do not exploit text or structured metadata to assist search and curation. Wang et al. [13] introduced iMap, a treemap-based visualization for navigating image search and clustering results, but the number of images displayed is limited. iGraph [5] constructs a similarity graph by computing the affinity between images, the relationship between texts, and the connection between images and texts, which results in a system that requires enormous amount of computation.

[2]https://artsexperiments.withgoogle.com/tsnemap/
[3]http://dhlab.yale.edu/projects/pixplot/

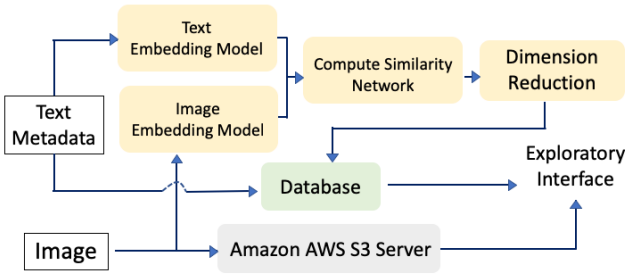


Figure 2: GraviTIE system overview. Yellow: learning the similarity map. Green: learned embeddings and associated metadata are stored in a database. Gray: images are stored in a cloud-based object store.

3 GRAVITIE SYSTEM OVERVIEW

Figure 2 summarizes the architecture of GraviTIE. The pipeline can roughly be divided into three parts: Constructing similarity map, textual and numeric data storage, and image storage.

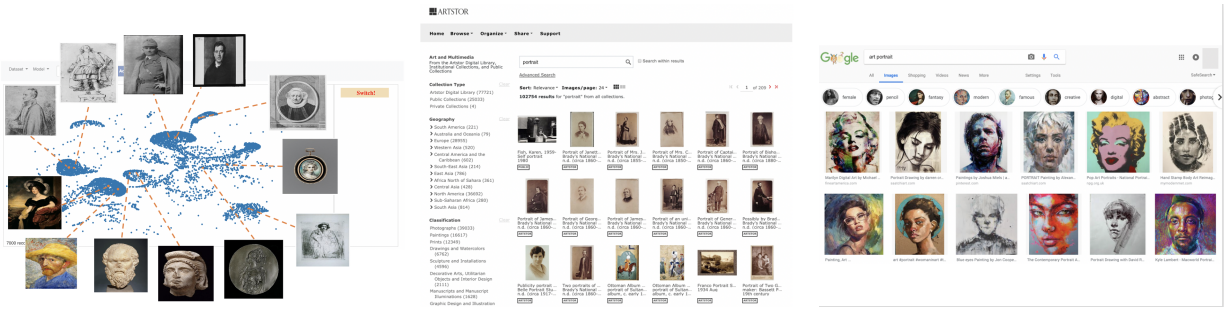
Learning the Similarity Map: The front end interface of GraviTIE takes a 2-d vector of coordinates for each image-text pair to construct the similarity map visualization. In this demonstration, all datasets are pre-processed by the same pipeline in which we first separately embed image and text, acquire a combined representation vector for each text and image pair, and finally apply a dimensional reduction algorithm to obtain the desired 2-d numeric vectors. We use MultiDEC [14] to learn representations of these image-caption pairs. MultiDEC is a method for clustering image-caption pairs by iteratively computing an auxiliary target distribution and matching both image distribution and text distribution to the target. Because MultiDEC produces distinct and semantically meaningful representations for both image and text, it is a natural candidate for this task. Users can also visualize the visual features and textual features alone.

Textual and Numeric Data Storage: The metadata, textual description, and the 2-d numeric vector for each image-text pair is saved in a MySQL database. GraviTIE also utilizes MySQL full text search (boolean mode) to support keyword search.

Image Storage: GraviTIE uses Amazon AWS S3 to store millions of image for display purposes.

4 GRAVITIE DEMONSTRATION DATASETS

We will demonstrate the GraviTIE system on three large-scale datasets to show the versatility and the breadth of GraviTIE.



(a) Search for "portrait" on GraviTIE. (b) Search for "portrait" on Artstor. (c) Search for "portrait" on Google Images.

Figure 3: Search results for "portrait" using (a) GraviTIE, (b) Artstor, and (c) Google Images. GraviTIE summarizes the space of relevant images; different clusters represent different types of portraits. Artstor and Google Images each show only a small set of top-ranked images, and each engine uses a different opaque ranking function, making the results unpredictable.

Russian Tweets. Twitter [4] released over 9 millions tweets from 3,841 accounts that are believed to be connected to Russian Internet Research Agency. This dataset has received significant attention from researchers to understand foreign influence on US politics. We process and display all 543,220 tweets with images on GraviTIE to investigate relationships from a visual perspective.

In this particular demonstration, we first embed images with ResNet-50 [6] model, pre-trained on 1 million ImageNet dataset[4], and acquire a text representation with a Glove [12] model that is pre-trained on 2 billion tweets. We then apply *multiDEC*[15] to learn a joint representation for the combined image-text pair. Finally, UMAP [10] is performed to produce 2-d vectors for visualization.

Art Imagery. GraviTIE includes 2,144,301 art images from the ArtStor collection. Artstor is a non-profit organization that creates digital libraries for scholars arts, humanities, social sciences and architecture. They approached us looking for new ways to explore their large image sets that went beyond simple metadata queries. For these images, the metadata is sparse, so the results of the queries are unreliable. A goal of GraviTIE is to utilize the full coverage of arbitrary data, with no assumptions on completeness. We also want to visually convey the results of unsupervised image clustering. A group of art historians are currently using GraviTIE to find patterns of influence and providing us with feedback.

We extracted visual features for the art images using pre-trained ResNet-50. We concatenate values of title, art classification, creator, repository, and date to serve as the description of each image. We then trained a FastText [2] model based on the description and acquired the textual/meta-data representation for each art work. Similar to the pipeline for the twitter data, we applied *multiDEC* followed by *UMAP* to obtain the final 2-d representation.

Scientific Figures. In addition to the art images and twitter data sets, we use a large database of scientific figures that we have been extracted from PubMed, arXiv and other repositories of the scientific literature. The goal is to understand how the use of different types of figures vary over time and across fields. The example dataset includes over 5 million figures from 800k scientific papers from the arXiv.org. ArXiv provides the raw pdfs; we extract images and captions from these pdfs using pdffigures2.0 [3] and then follow the similarity map pipeline described previously.

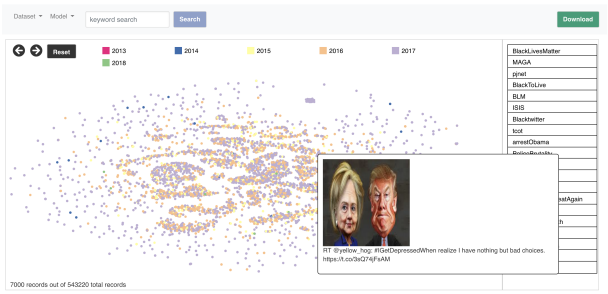
[4]https://about.twitter.com/en_us/values/elections-integrity.html#data

5 KEY FEATURES

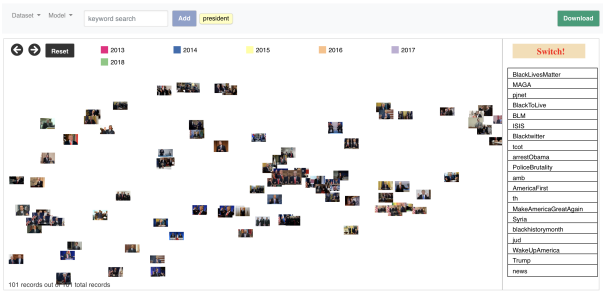
GraviTIE supports interactive queries and the exploration images *collections*. The central feature is a similarity map supporting mixed queries, panning, zooming and visual inspection of individual images. Each image-text pair is represented as a single glyph colored by a user-configurable attribute and situated in the 2-D plane based on the similarity with other image-text pairs. Users can hover over each glyph to view the corresponding image and text content associated with that image (Fig. 4a). The query system (Fig. 4c) affords dataset refinement and curation via keyword search or structured predicates. Users can customize the current view through a highlighting feature (Fig. 4d). These customizations can easily be saved and shared with other users since each view is associated with a unique url. Below we highlight other features of the system:

Similarity Map. : The similarity map is what distinguishes the design of GraviTIE from other systems for interacting with large image collections. It makes no assumptions about the images (unlike, say Building Rome in A Day [1]), and can make use of multi-modal information if available. The similarity map can also be used as a replacement for conventional list-based search interfaces: by providing a visual summary of all results rather than the top few items in a ranked list, we allow the user to process information and identify patterns faster than scanning a list of textual information [11].

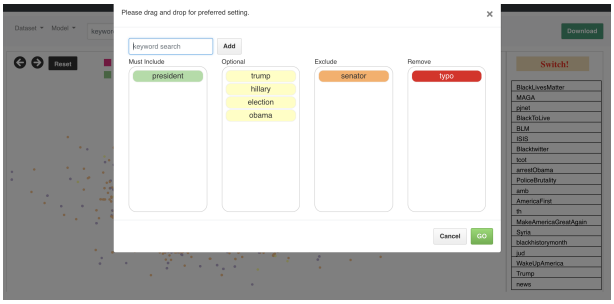
Consider the simple experiment in Figure 3. We search for the keyword *portrait* on three platforms: GraviTIE, Artstor, and Google Images. GraviTIE reveals that there are several dense clusters of similar images, and upon inspection, these clusters are easily identified as corresponding to print portraits, painted portraits, photograph portraits, and other semantic types. However, the Artstor search results display only the top k images ranked by query relevance, which in this case consist entirely of print portraits. The ranking function used by Google Images makes another seemingly arbitrary choice, displaying entirely painted portraits. Not only do both of these sets of results seem incorrect in general, but even for those situations where they are what the user wants, GraviTIE can emulate similar results within a few hover-and-zoom interactions.



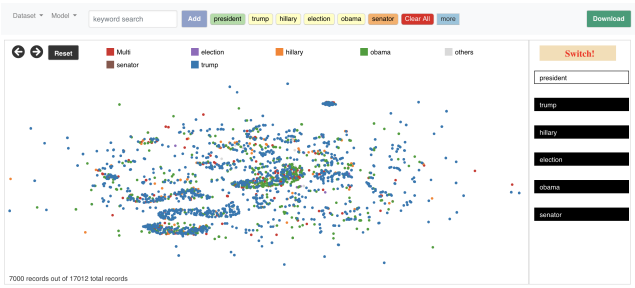
(a) Similarity Map: Each point represents an image-text pair. The user can hover on a point to display the visual and textual content.



(b) Direct inspection: When the number of records is less than 300, thumbnail images are displayed instead of glyphs.



(c) Advanced search: Keyword semantics are configurable.



(d) Highlighting: Items with a selected keyword are highlighted.

Figure 4: Some features of GraviTIE: (a) Similarity map, (b) Direct inspection, (c) Advanced search, (d) Highlighting

Scalability, Query, and Reproducibility. : The similarity map also directly affords scalability. By simply applying a sampling mechanism, we can retain the overall structure of the (embedded) image space while controlling system performance and not occluding individual images. By combining offline pre-processing and sampling, we maintain high performance interactions for datasets of arbitrary scale. Cluster-aware sampling approaches and their impact on interactivity and user task performance remain future work.

GraviTIE supports exploring a set of related keywords at a time. Given a set of user-provided search terms, the user can force each term to be included or excluded. For instance, the results of the query result of Fig.4c would definitely include the term *president*, might contain *trump*, *hilary*, *election*, and *obama*, and would exclude *senator*. In addition, selecting a keyword highlights the elements in the similarity map associated with that keyword, as in Fig. 4d. To initialize this interaction, users can toggle between default interesting keywords (eg. Fig. 4b shows the top 20 frequent hashtags) or type their own (eg. Fig. 4d). If an item includes multiple highlighted keywords, it is indicated with a distinguished color.

Every view in GraviTIE is associated with a unique url to afford collaboration, provenance, and sharing. The browser history can be used to retrace steps, and particular views can be saved simply by copying and pasting links. In addition, a user can download the corresponding dataset associated with any view for further analysis. This feature has been valuable in supporting research by providing an easy way to construct high-quality training datasets from important subsets of images. For example, GraviTIE can be used to quickly find a large set of phylogenetic tree images in the scientific literature for information extraction research[9].

6 CONCLUSION

We have built a fully functional prototype of GraviTIE with several large-scale, image sets to explore. The tool is publicly available and easily accessible in both desktop and mobile environments. For the conference, we plan to have several examples of how the tool can be used to identify and zoom into clusters of similar images, the text associated with these images and nuances in the output of these different unsupervised algorithms. Conference attendees will be able to initiate our pre-baked queries but also initiate their own queries. As an example, we will default to the term "portrait" in the art image set. This will illustrate the distinct types of portraits that exits in the 2.5 million ArtStor images, as mentioned above. Early feedback from collaborators in art history suggest that they can use GraviTIE to quickly understand the coverage of the database in certain time periods and styles (e.g., how many portrait archetypes exist in 20th century paintings). Our goal for the demonstration is to show how GraviTIE works and to gather feedback from users on what kind of functionality is needed to help them explore image sets in ways not available.

7 ACKNOWLEDGEMENT

We would like to thank Artstor who provided their valuable art collection to visualize on GraviTIE.

REFERENCES

[1] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M Seitz, and Richard Szeliski. 2009. Building rome in a day. In *ICCV*. IEEE, 72–79.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[3] Christopher Clark and Santosh Divvala. 2016. PDFFigures 2.0: Mining figures from research papers. In *JCDL*. IEEE, 143–152.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. Ieee, 248–255.

[5] Yi Gu, Chaoli Wang, Jun Ma, Robert J Nemiroff, David L Kao, and Denis Parra. 2017. Visualization and recommendation of large image collections toward effective sensemaking. *Information Visualization* 16, 1 (2017), 21–47.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[7] Nadav Hochman and Lev Manovich. 2013. Zooming into an Instagram City: Reading the local through social media. *First Monday* 18, 7 (2013).

[8] Poshen Lee, Jevin West, and Bill Howe. 2016. VizioMetrix: A Platform for Analyzing the Visual Information in Big Scholarly Data. In *BigScholar Workshop (co-located at WWW)*.

[9] Poshen Lee, T. Sean Yang, Jevin West, and Bill Howe. 2017. PhyloParser: A Hybrid Algorithm for Extracting Phylogenies from Dendrograms. (2017).

[10] Leland McInnes and John Healy. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[11] Douglas L Nelson, Valerie S Reed, and John R Walling. 1976. Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning and Memory* 2, 5 (1976), 523.

[12] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[13] Chaoli Wang, John P Reese, Huan Zhang, Jun Tao, Yi Gu, Jun Ma, and Robert J Nemiroff. 2015. Similarity-based visualization of large image collections. *Information Visualization* 14, 3 (2015), 183–203.

[14] Sean Yang, Kuan-Hao Huang, and Bill Howe. 2019. MultiDEC: Multi-Modal Clustering of Image-Caption Pairs. *arXiv preprint arXiv:1901.01860* (2019).

[15] Sean Yang, Kuan-Hao Huang, and Bill Howe. 2019. MultiDEC: Multi-Modal Clustering of Image-Caption Pairs. (2019). arXiv:arXiv:1901.01860