Semantic Path-Based Learning for Review Volume Prediction

Ujjwal Sharma[1(✉)], Stevan Rudinac[1], Marcel Worring[1], Joris Demmers[2], and Willemijn van Dolen[1]

[1] University of Amsterdam, Amsterdam, The Netherlands
{u.sharma,s.rudinac,m.worring,w.m.vandolen}@uva.nl
[2] Monash University, Melbourne, Australia
joris.demmers@monash.edu

Abstract. Graphs offer a natural abstraction for modeling complex real-world systems where entities are represented as nodes and edges encode relations between them. In such networks, entities may share common or similar attributes and may be connected by paths through multiple attribute modalities. In this work, we present an approach that uses semantically meaningful, bimodal random walks on real-world heterogeneous networks to extract correlations between nodes and bring together nodes with shared or similar attributes. An attention-based mechanism is used to combine multiple attribute-specific representations in a late fusion setup. We focus on a real-world network formed by restaurants and their shared attributes and evaluate performance on predicting the number of reviews a restaurant receives, a strong proxy for popularity. Our results demonstrate the rich expressiveness of such representations in predicting review volume and the ability of an attention-based model to selectively combine individual representations for maximum predictive power on the chosen downstream task.

Keywords: Heterogeneous information networks · Metapaths · Deep learning on graphs · Venue popularity prediction

1 Introduction

Multimodal graphs have been extensively used in modeling real-world networks where entities interact and communicate with each other through multiple information pathways or *modalities* [1,23,31]. Each modality encodes a distinct *view* of the relation between nodes. For example, within a social network, users can be connected by their shared preference for a similar product or by their presence in the same geographic locale. Each of these semantic contexts links the same user set with a distinct edge set. Such networks have been extensively used for applications like semantic proximity search in existing interaction networks [7], augmenting semantic relations between entities [36], learning interactions in an unsupervised fashion [3] and augmenting traditional matrix factorization-based collaborative filtering models for recommendation [27].

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12035, pp. 821–835, 2020.
https://doi.org/10.1007/978-3-030-45439-5_54

Each modality within a multimodal network encodes a different semantic relation and exhibits a distinct *view* of the network. While such views contain relations between nodes based on interactions within a single modality, observed outcomes in the real-world are often a complex combination of these interactions. Therefore, it is essential to compose these complementary interactions meaningfully to build a better representation of the real world. In this work, we examine a multimodal approach that attempts to model the review-generation process as the end-product of complex interactions within a restaurant network.

Restaurants share a host of attributes with each other, each of which may be treated as a modality. For example, they may share the same neighborhood, the same operating hours, similar kind of cuisine, or the same 'look and feel'. Furthermore, each of these attributes only uncovers a specific type of relation. For example, a view that only uses the location-modality will contain venues only connected by their colocation in a common geographical unit and will prioritize physical proximity over any other attribute. Broadly, each of these views is characterized by a semantic context and encodes modality-specific relations between restaurants. These views, although informative, are *complementary* and only record associations within the same modality. While each of these views encodes a part of the interactions within the network, performance on a downstream task relies on a suitable combination of views pertinent to the task [5].

In this work, we use metapaths as a semantic interface to specify which relations within a network may be relevant or meaningful and worth investigating. We generate bimodal low-dimensional embeddings for each of these metapaths. Furthermore, we conjecture that their relevance on a downstream task varies with the nature of the task and that this task-specific modality relevance should be learned from data. In this work,

- We propose a novel method that incorporates restaurants and their attributes into a multimodal graph and extracts multiple, bimodal low dimensional representations for restaurants based on available paths through shared visual, textual, geographical and categorical features.
- We use an attention-based fusion mechanism for selectively combining representations extracted from multiple modalities.
- We evaluate and contrast the performance of modality-specific representations and joint representations for predicting review volume.

2 Related Work

The principle challenge in working with multimodal data revolves around the task of extracting and assimilating information from multiple modalities to learn informative joint representations. In this section, we discuss prior work that leverages graph-based structures for extracting information from multiple modalities, focussing on the auto-captioning task that introduced such methods. We then examine prior work on network embeddings that aim to learn discriminative representations for nodes in a graph.

2.1 Graphs for Modelling Semantic Relationships

Graph-based learning techniques provide an elegant means for incorporating semantic similarities between multimedia documents. As such, they have been used for inference in large multimodal collections where a single modality may not carry sufficient information [2]. Initial work in this domain was structured around the task of captioning unseen images using correlations learned over multiple modalities (tag-propagation or auto-tagging). Pan *et al.* use a graph-based model to discover correlations between image features and text for automatic image-captioning [21]. Urban *et al.* use an *Image-Context Graph* consisting of captions, image features and images to retrieve relevant images for a textual query [32]. Stathopoulos *et al.* [28] build upon [32] to learn a similarity measure over words based on their co-occurrence on the web and use these similarities to introduce links between similar caption words. Rudinac *et al.* augment the *Image-Context Graph* with users as an additional modality and deploy it for generating visual-summaries of geographical regions [25]. Since we are interested in discovering multimodal similarities between restaurants, we use a graph layout similar to the one proposed by Pan *et al.* [21] for the image auto-captioning task but replace images with restaurants as central nodes. Other nodes containing textual features, visual features and users are retained. We also add categorical information like cuisines as a separate modality, allowing them to serve as semantic anchors within the representation.

2.2 Graph Representation Learning

Graph representation learning aims to learn mappings that embed graph nodes in a low-dimensional compressed representation. The objective is to learn embeddings where geometric relationships in the compressed embedding space reflect structural relationships in the graph. Traditional approaches generate these embeddings by finding the leading eigenvectors from the affinity matrix for representing nodes [16,24]. With the advent of deep learning, neural networks have become increasingly popular for learning such representations, jointly, from multiple modalities in an end-to-end pipeline [4,11,14,30,34].

Existing random walk-based embedding methods are extensions of the Random Walks with Restarts (RWR) paradigm. Traditional RWR-based techniques compute an affinity between two nodes in a graph by ascertaining the steady-state transition probability between them. They have been extensively used for the aforementioned auto-captioning tasks [21,25,28,32], tourism recommendation [15] and web search as an integral part of the PageRank algorithm [20]. Deep Learning-based approaches build upon the traditional paradigm by optimizing the co-occurrence statistics of nodes sampled from these walks. *DeepWalk* [22] uses nodes sampled from short truncated random walks as phrases to optimize a skip-gram objective similar to *word2vec* [17]. Similarly, *node2vec* augments this learning paradigm with second-order random walks parameterized by exploration parameters p and q which control between the importance of homophily and structural equivalence in the learnt representations [8]. For a homogeneous

network, random walk based methods like *DeepWalk* and *node2vec* assume that while the probabilities of transitioning from one node to another can be different, every transition still occurs between nodes of the same type. For heterogeneous graphs, this assumption may be fallacious as all transitions do not occur between nodes of the same type and consequently, do not carry the same semantic context. Indeed, our initial experiments with *node2vec* model suggest that it is not designed to handle highly multimodal graphs. Clements *et al.* [5] demonstrated that in the context of content recommendation, the importance of modalities is strongly task-dependent and treating all edges in heterogeneous graphs as equivalent can discard this information. *Metapath2Vec* [6] remedies this by introducing unbiased walks over the network schema specified by a *metapath* [29], allowing the network to learn the semantics specified by the metapath rather than those imposed purely by the topology of the graph. Metapath-based approaches have been extended to a variety of other problems. Hu *et al.* use an exhaustive list of semantically-meaningful metapaths for extracting Top-N recommendations with a neural co-attention network [10]. Shi *et al.* use metapath-specific representations in a traditional matrix factorization-based collaborative filtering mechanism [27]. In this work, we perform random walks on sub-networks of a restaurant-attribute network containing restaurants and attribute modalities. These attribute modalities may contain images, text or categorical features. For each of these sub-networks, we perform random walks and use a variant of the heterogeneous skip-gram objective introduced in [6] to generate low-dimensional bimodal embeddings. Bimodal embeddings have several interesting properties. Training relations between two modalities provide us with a degree of modularity where modalities can be included or held-out from the prediction model without affecting others. It also makes training inexpensive as the number of nodes when only considering two modalities is far lower than in the entire graph.

3 Proposed Method

In this section, we begin by providing a formal introduction to graph terminology that is frequently referenced in this paper. We then move on to detail our proposed method illustrated in Fig. 1.

3.1 Definitions

Formally, a heterogeneous graph is denoted by $G = (V, E, \phi, \sigma)$ where V and E denote the node and edge sets respectively. For every node and edge, there exists mapping functions $\phi(v) \rightarrow \mathcal{A}$ and $\sigma(e) \rightarrow \mathcal{R}$ where $\mathcal{A}$ and $\mathcal{R}$ are sets of node types and edge types respectively such that $|\mathcal{A} + \mathcal{R}| > 2$.

For a heterogeneous graph $G = (V, E, \phi, \sigma)$, a network schema is a *metagraph* $M_G = (\mathcal{A}, \mathcal{R})$ where $\mathcal{A}$ is the set of node types in V and $\mathcal{R}$ is the set of edge types in E. A network schema enumerates the possible node types and edge types that can occur within a network.

A metapath $\mathcal{M}(\mathcal{A}_1, \mathcal{A}_n)$ is a path on the network schema M_G consisting of a sequence of ordered edge transitions: $\mathcal{M}(\mathcal{A}_1, \mathcal{A}_n) : [\mathcal{A}_1 \rightarrow \mathcal{A}_2 \rightarrow \ldots \rightarrow \mathcal{A}_n]$.

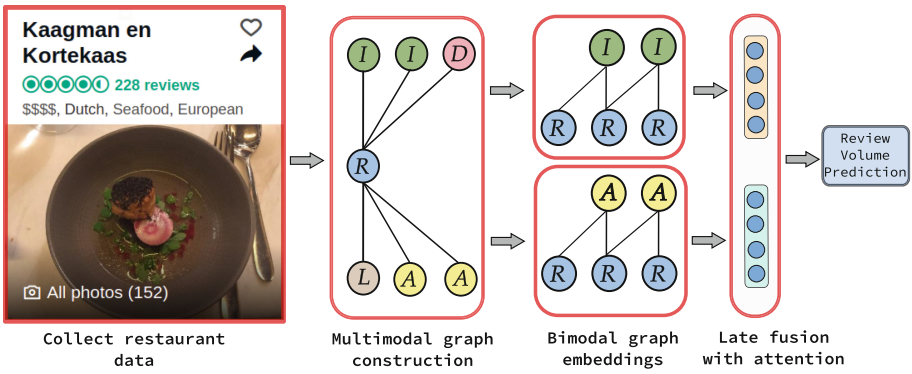

Fig. 1. Model pipeline: We use TripAdvisor to collect information for restaurants in Amsterdam. Each venue characteristic is then embedded as a separate node within a multimodal graph. In the figure above R nodes denote restaurants, I nodes denote images for a restaurant, D nodes are review documents, A nodes are categorical attributes for restaurants and L nodes are locations. Bimodal random walks are used to extract pairwise correlations between nodes in separate modalities which are embedded using a heterogeneous skip-gram objective. Finally, an attention-based fusion model is used to combine multiple embeddings together to regress the review volume for restaurants.

3.2 Graph Nodes

Let $G = (V, E)$ be the heterogeneous graph with a set of nodes V and edges E. We assume the graph to be undirected as linkages between venues and their attributes are inherently symmetric. Below, we describe the node types used to construct the graph (cf. Figs. 1 and 2).

1. **Restaurant Nodes:** For each of the N restaurants in a city, we introduce a node.
2. **Categorical Attribute Nodes:** A categorical feature node is added for each of the categories below.
   - **Cuisines:** Product/Cuisine type served in the restaurant.
   - **Meals:** Meals (Breakfast, Lunch, Dinner, etc.) which the restaurant serves.
   - **Features:** Additional services/attributes for the restaurants.
   - **Symbolic Price:** A discretized price bracket for venues.
3. **Location:** We split the region under consideration into rectangular geographical cells and bin restaurant locations into these cells. We add a node for every geographical cell that contains at least one restaurant within it.
4. **Image Features:** Images of a restaurant reflect its visual look-and-feel. These images may depict e.g. the interior, food and drinks, or people enjoying their meal. To extract a high-level semantic representation from these images, we deploy the ResNet-18 network [9] trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset [26] with 1000 semantic concepts

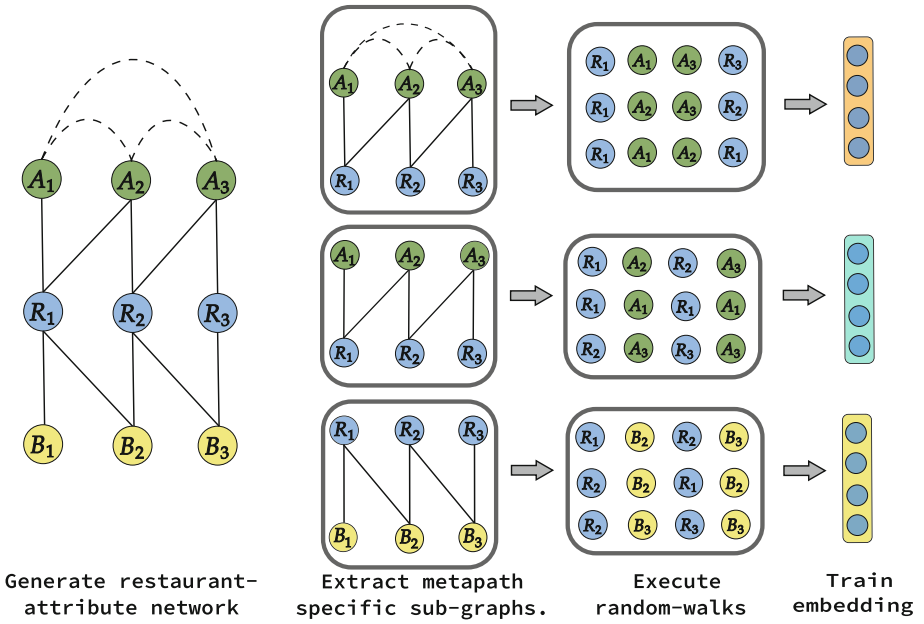

Fig. 2. Pipeline for embedding modality-specific sub-networks: In this illustrative example, we use a graph with 2 modalities. $\{R_1, \ldots, R_3\}$ are restaurant nodes, $\{A_1, \ldots, A_3\}$ are nodes from a modality with categorical information and $\{B_1, \ldots, B_3\}$ are nodes from a modality with similarity linkages. Random walks are performed over modality-specific sub-networks for each metapath and each model is trained separately to create metapath-specific embeddings.

and use the penultimate layer output as a compressed low-dimensional representation for the image. Since the number of available images for each venue may vary dramatically depending on its popularity, adding a node for every image can lead to an unreasonably large graph. To mitigate this issue, we cluster image features for each restaurant using the K-Means algorithm and use the cluster centers as representative image features for a restaurant, similar to Zahálka *et al.* [35]. We chose K = 5 as a reasonable trade-off between the granularity of our representations and tractability of generating embeddings for this modality.

5. **User Review Features:** The way patrons write about a restaurant and the usage of specialized terms can contain important information about a restaurant that may be missing from its categorical attributes. For example, usage of the Indian cottage cheese 'Paneer' can be found in similar cuisine types like Nepali, Surinamese, etc. and user reviews talking about dishes containing 'Paneer' can be leveraged to infer that Indian and Nepali cuisines share some degree of similarity. To model such effects, we collect reviews for every restaurant. Since individual reviews may not provide a comprehensive unbiased picture of the restaurant, we chose not to treat them individually,

but to consider them as a single document. We then use a distributed bag-of-words model from [13] to generate low-dimensional representations of these documents for each restaurant. Since the reviews of a restaurant can widely vary based on its popularity, we only consider the 10 most recent reviews for each restaurant to prevent biases from document length getting into the model.

6. **Users:** Since TripAdvisor does not record check-ins, we can only leverage explicit feedback from users who chose to leave a review. We add a node for each of the users who visited at least two restaurants in Amsterdam and left a review.

3.3 Graph Edges

Similar to [25,28,32], we introduce two kinds of edges in our graph:

1. **Attribute edges:** These are *heterogeneous* edges that connect a restaurant node to the nodes of its categorical attributes, image features, review features and users. In our graph, we instantiate them as undirected, unweighted edges.
2. **Similarity edges:** These are *homogeneous* edges between the feature nodes within a single modality. For image features, we use a radial basis function as a non-linear transformation of the euclidean distances between image feature vectors.

$$S_I(I_j, I_k) = \exp\left(-\frac{\|I_j - I_k\|^2}{2\sigma^2}\right) \tag{1}$$

For document vectors, we use cosine similarity to find restaurants with similar reviews.

$$S_T(T_j, T_k) = \frac{T_j \cdot T_k}{\|T_j\|\|T_k\|} \tag{2}$$

Adding a weighted similarity edge between every node in the same modality would yield an extremely dense adjacency matrix. To avoid this, we only add similarity links between a node and its K nearest neighbors in each modality. By choosing the nearest K neighbors, we make our similarity threshold adaptive allowing it to adjust to varying scales of distance in multiple modalities.

3.4 Bimodal Graph Embeddings

Metapaths can provide a modular and simple interface for injecting semantics into the network. Since metapaths, in our case, are essentially paths over the modality set, they can be used to encode inter-modality correlations. In this work, we generate embeddings with two specific properties:

1. All metapaths are binary and only include transitions over 2 modalities. Since venues/restaurants are always a part of the metapath, we only include one other modality.

2. During optimization, we only track the short-range context by choosing a small window size. Window size is the maximum distance between the input node and a predicted node in a walk. In our model, walks over the metapath only capture short-range semantic contexts and the choice of a larger window can be detrimental to generalization. For example, consider a random walk over the **Restaurant - Cuisine - Restaurant** metapath. In the sampled nodes below, restaurants are in red while cuisines are in blue.

   [Kediri, Indonesian Cuisine, Indonesian Palace, Chinese Cuisine, China Town, Fast Food Cuisine, McDonalds]

   Optimizing over a large context window can lead to McDonald's (fast-food cuisine) and Kediri (Indonesian cuisine) being placed close in the embedding space. This is erroneous and does not capture the intended semantics which should bring restaurants closer only if they share the exact attribute.

We use the metapaths in Table 1 to perform unbiased random walks on the graph detailed in Sect. 3.2. Each of these metapaths enforces similarity based on certain semantics.

Table 1. Metapaths used for the Restaurant-Attributes multimodal network.

No.	Metapath
1	Venues - Cuisines - Venues
2	Venues - Facilities - Venues
3	Venues - Meals - Venues
4	Venues - Price - Venues
5	Venues - Location cell - Venues
6	Venues - Users - Venues
7	Venues - Location cell - Location cell - Venues
8	Venues - Img. Feats. - Img. Feats - Venues
9	Venues - Review Feats. - Review Feats - Venues

We train separate embeddings using the heterogeneous skip-gram objective similar to [6]. For every metapath, we maximize the probability of observing the heterogeneous context $N_a(v)$ given the node v. In Eq. (3), $\mathcal{A}_m$ is the node type-set and V_m is the node-set for metapath m.

$$\arg\max_{\theta} \sum_{v \in V_m} \sum_{a \in \mathcal{A}_m} \sum_{c_a \in N_a(v)} \log p(c_a | v; \theta) \tag{3}$$

3.5 Modality Fusion

The original Metapath2vec model [6] uses multiple metapaths [29] to learn separate embeddings, some of which perform better than the others. On the DBLP

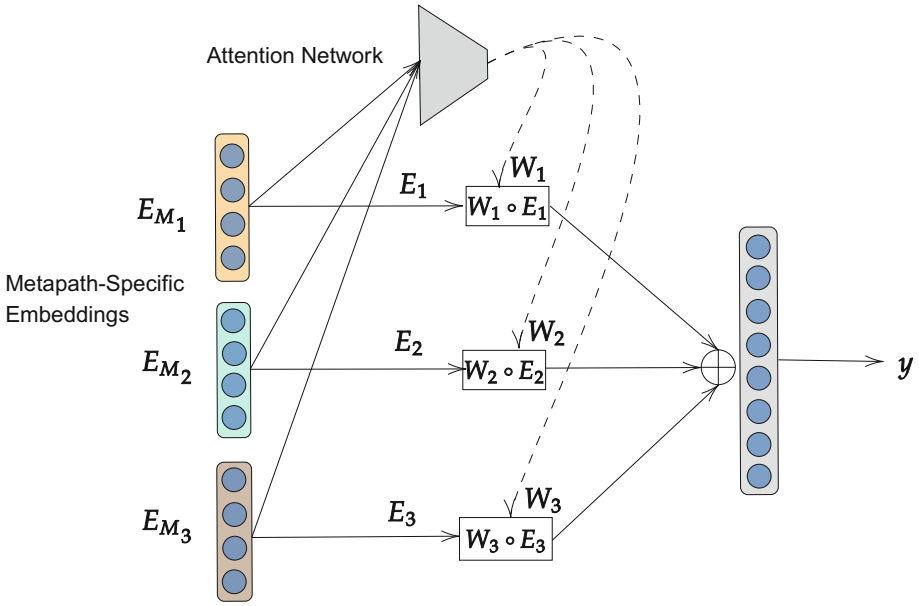

Fig. 3. Attention-weighted modality fusion: metapath-specific embeddings are fed into a common attention mechanism that generates an attention vector. Each modality is then reweighted with the attention vector and concatenated. This joint representation is then fed into a ridge regressor to predict the volume of ratings for each restaurant.

bibliographic graph that consists of Authors (A), Papers (P) and Venues (V), the performance of their recommended metapath 'A-P-V-P-A' was empirically better than the alternative metapath 'A-P-A' on the node classification task. At this point, it is important to recall that in our model, each metapath extracts a separate *view* of the same graph. These views may contain *complementary* information and it may be disadvantageous to only retain the best performing view. For an optimal representation, these complementary views should be fused.

In this work, we employ an embedding-level attention mechanism similar to the attention mechanism introduced in [33] that selectively combines embeddings based on their performance on a downstream task. Assuming S to be the set of metapath-specific embeddings for metapaths $m_1, m_2, \ldots, m_N$, following the approach outlined in Fig. 3, we can denote it as:

$$S = \{E_{m_1}, E_{m_2}, \cdots, E_{m_N}\} \tag{4}$$

We then use a two-layer neural network to learn an embedding-specific attention A_{m_n} for metapath m_n:

$$A_{m_n} = h \cdot ReLU(W \cdot E_{m_n} + b) \tag{5}$$

Further, we perform a softmax transformation of the attention network outputs to an embedding-specific weight

$$w_{m_n} = \frac{\exp(A_{m_n})}{\sum_{k=1}^{N} \exp(A_{m_k})} \tag{6}$$

Finally, we concatenate the attention-weighted metapath-specific embeddings to generate a fused embedding

$$O = [w_{m_1}.E_{m_1}|w_{m_2}.E_{m_2}|\ldots|w_{m_N}.E_{m_N}] \tag{7}$$

4 Experiments

We evaluate the performance of the embedding fusion model on the task of predicting the volume (total count) of reviews received by a restaurant. We conjecture that the volume of reviews is an unbiased proxy for the general popularity and footfall for a restaurant and is more reliable than indicators like ranking or ratings which may be biased by TripAdvisor's promotion algorithms. We use the review volume collected from TripAdvisor as the target variable and model this task as a regression problem.

4.1 Experiment Setup

Data Collection. We use publicly-available data from TripAdvisor for our experiments. To build the graph detailed in Sect. 3.2, we collect data for 3,538 restaurants in Amsterdam, The Netherlands that are listed on TripAdvisor. We additionally collect 168,483 user-contributed restaurant reviews made by 105,480 *unique* users, of which only 27,318 users visit more than 2 restaurants in the city. We only retain these 27,318 users in our graph and drop others. We also collect 215,544 user-contributed images for these restaurants. We construct the restaurant network by embedding venues and their attributes listed in Table 1 as nodes.

Bimodal Embeddings. We train separate bimodal embeddings by optimizing the heterogeneous skip-gram objective from Eq. (3) using stochastic gradient descent and train embeddings for all metapaths enumerated in Table 1. We use restaurant nodes as root nodes for the unbiased random walks and perform 80 walks per root node, each with a walk length of 80. Each embedding has a dimensionality of 48, uses a window-size of 5 and is trained for 200 epochs.

Embedding Fusion Models. We chose two fusion models in our experiments to analyze the efficacy of our embeddings:

1. Simple Concatenation Model: We use a model that performs a simple concatenation of the individual metapath-specific embeddings detailed in Sect. 3.4 to exhibit the baseline performance on the tasks detailed in Sect. 4. Simple concatenation is a well-established additive fusion technique in multimodal deep learning [18,19].

Table 2. Performance of individual metapath-specific embeddings on the review volume prediction task

Metapath	Mean squared error	Mean absolute error	Coefficient of determination
Venues - Cuisines - Venues	2.56	1.24	0.22
Venues - Facilities - Venues	1.93	1.09	0.40
Venues - Meals - Venues	2.41	1.23	0.28
Venues - Price - Venues	2.00	1.09	0.36
Venues - Location cell - Venues	3.13	1.45	0.09
Venues - Users - Venues	3.38	1.54	0.01
Venues - Location cell - Location cell - Venues	2.75	1.36	0.08
Venues - Img. Feats. - Img. Feats - Venues	2.73	1.33	0.12
Venues - Review Feats. - Review Feats - Venues	2.68	1.32	0.224
Modality concatenation + Ridge regression	1.29	0.89	0.62
Attention weighted modality concatenation + Ridge regression	0.99	0.77	0.70

2. Modality Fusion Model: We use our attention-based modality fusion model detailed in Sect. 3.5 to exhibit the effects of a learnable weighted-fusion of modalities on the evaluation tasks.

Each of the models uses a ridge regression algorithm to estimate the predictive power of each metapath-specific embedding on the volume regression task. This regressor is jointly trained with the attention model in the Attention-Weighted Model. All models are optimized using stochastic gradient descent with the Adam optimizer [12] with a learning rate of 0.1.

4.2 Results and Findings

In Table 2, we report the results from our experiments on the review-volume prediction task. We observe that metapaths with nodes containing categorical attributes perform significantly better than vector-based features. In particular, categorical attributes like Cuisines, Facilities, and Price have a significantly higher coefficient of determination (R^2) as compared to visual feature nodes. It is interesting to observe here that nodes like locations, images, and textual reviews are far more numerous than categorical nodes and part of their decreased performance may be explained by the fact that our method of short walks may not be sufficiently expressive when the number of feature nodes is large. In

addition, as mentioned in related work, we performed these experiments with the *node2vec* model, but since it is not designed for heterogeneous multimodal graphs, it yielded performance scores far below the weakest single modality.

A review of the fusion models indicates that taking all the metapaths together can improve performance significantly. The baseline simple concatenation fusion model, commonly used in literature, is considerably better than the best-performing metapath (Venues - Facilities - Venues). The attention based-model builds significantly over the baseline performance and while it employs a similar concatenation scheme as the baseline concatenation model, the introduction of the attention module allows it to handle noisy and unreliable modalities. The significant increase in the predictive ability of the attention-based model can be attributed to the fact that while all modalities encode information, some of them may be less informative or reliable than others, and therefore contribute less to the performance of the model. Our proposed fusion approach is, therefore, capable of handling weak or noisy modalities appropriately.

5 Conclusion

In this work, we propose an alternative, modular framework for learning from multimodal graphs. We use metapaths as a means to specify semantic relations between nodes and each of our bimodal embeddings captures similarities between restaurant nodes on a single attribute. Our attention-based model combines separately learned bimodal embeddings using a late-fusion setup for predicting the review volume of the restaurants. While each of the modalities can predict the volume of reviews to a certain extent, a more comprehensive picture is only built by combining complementary information from multiple modalities. We demonstrate the benefits of our fusion approach on the review volume prediction task and demonstrate that a fusion of complementary views provides the best way to learn from such networks. In future work, we will investigate how the technique generalises to other tasks and domains.

References

1. Abrach, H., et al.: MANTIS: system support for MultimodAl NeTworks of in-situ sensors. In: Proceedings of the Second ACM International Workshop on Wireless Sensor Networks and Applications, WSNA 2003 (2003)
2. Arya, D., Rudinac, S., Worring, M.: HyperLearn: a distributed approach for representation learning in datasets with many modalities. In: MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia (2019). https://doi.org/10.1145/3343031.3350572
3. Battaglia, P., Pascanu, R., Lai, M., Rezende, D.J., Kavukcuoglu, K.: Interaction networks for learning about objects, relations and physics. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS 2016, pp. 4509–4517. Curran Associates Inc., USA (2016). http://dl.acm.org/citation.cfm?id=3157382.3157601

4. Chang, S., Han, W., Tang, J., Qi, G.J., Aggarwal, C.C., Huang, T.S.: Heterogeneous network embedding via deep architectures. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015). https://doi.org/10.1145/2783258.2783296
5. Clements, M., De Vries, A.P., Reinders, M.J.T.: The task-dependent effect of tags and ratings on social media access. ACM Trans. Inf. Syst. (2010). https://doi.org/10.1145/1852102.1852107
6. Dong, Y., Chawla, N.V., Swami, A.: Metapath2vec: scalable representation learning for heterogeneous networks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017). https://doi.org/10.1145/3097983.3098036
7. Fang, Y., Zhao, X., Huang, P., Xiao, W., de Rijke, M.: M-HIN: complex embeddings for heterogeneous information networks via metagraphs. In: Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, pp. 913–916. ACM, New York, NY, USA (2019). https://doi.org/10.1145/3331184.3331281, http://doi.acm.org/10.1145/3331184.3331281
8. Grover, A., Leskovec, J.: Node2vec: scalable feature learning for networks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016). https://doi.org/10.1145/2939672.2939754
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016). https://doi.org/10.1109/CVPR.2016.90
10. Hu, B., Shi, C., Zhao, W.X., Yu, P.S.: Leveraging meta-path based context for top-n recommendation with a neural co-attention model. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2018). https://doi.org/10.1145/3219819.3219965
11. Huang, F., Zhang, X., Li, C., Li, Z., He, Y., Zhao, Z.: Multimodal network embedding via attention based multi-view variational autoencoder. In: ICMR 2018 - Proceedings of the 2018 ACM International Conference on Multimedia Retrieval (2018). https://doi.org/10.1145/3206025.3206035
12. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic gradient descent. In: ICLR: International Conference on Learning Representations (2015)
13. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: 31st International Conference on Machine Learning, ICML 2014 (2014)
14. Li, Z., Tang, J., Mei, T.: Deep collaborative embedding for social image understanding. IEEE Trans. Pattern Anal. Mach. Intell. (2018). https://doi.org/10.1109/TPAMI.2018.2852750
15. Lucchese, C., Perego, R., Silvestri, F., Vahabi, H., Venturini, R.: How random walks can help tourism. In: Baeza-Yates, R., et al. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 195–206. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28997-2_17
16. McAuley, J., Leskovec, J.: Image labeling on a network: using social-network metadata for image classification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 828–841. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_59
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (2013)

18. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011 (2011)
19. Ouyang, W., Chu, X., Wang, X.: Multi-source deep learning for human pose estimation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2014). https://doi.org/10.1109/CVPR.2014.299
20. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Technical report, 1999-66. Stanford InfoLab, November 1999. Previous number: SIDL-WP-1999-0120. http://ilpubs.stanford.edu:8090/422/
21. Pan, J.Y., Yang, H.J., Faloutsos, C., Duygulu, P.: GCap: graph-based automatic image captioning. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (2004). https://doi.org/10.1109/CVPR.2004.353
22. Perozzi, B., Al-Rfou, R., Skiena, S.: DeepWalk: online learning of social representations. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2014). https://doi.org/10.1145/2623330.2623732
23. Pirson, I., Fortemaison, N., Jacobs, C., Dremier, S., Dumont, J.E., Maenhaut, C.: The visual display of regulatory information and networks. Trends Cell Biol. (2000). https://doi.org/10.1016/S0962-8924(00)01817-1
24. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science (2000). https://doi.org/10.1126/science.290.5500.2323
25. Rudinac, S., Hanjalic, A., Larson, M.: Generating visual summaries of geographic areas using community-contributed images. IEEE Trans. Multimed. (2013). https://doi.org/10.1109/TMM.2013.2237896
26. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
27. Shi, C., Hu, B., Zhao, W.X., Yu, P.S.: Heterogeneous information network embedding for recommendation. IEEE Trans. Knowl. Data Eng. (2019). https://doi.org/10.1109/TKDE.2018.2833443
28. Stathopoulos, V., Urban, J., Jose, J.: Semantic relationships in multi-modal graphs for automatic image annotation. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 490–497. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78646-7_47
29. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: PathSim: meta path-based top-k similarity search in heterogeneous information networks. In: Proceedings of the VLDB Endowment (2011)
30. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: LINE: large-scale information network embedding. In: WWW 2015 - Proceedings of the 24th International Conference on World Wide Web (2015). https://doi.org/10.1145/2736277.2741093
31. Uchida, K., Sumalee, A., Watling, D., Connors, R.: Study on optimal frequency design problem for multimodal network using probit-based user equilibrium assignment. Transp. Res. Rec. (2005). https://doi.org/10.3141/1923-25
32. Urban, J., Jose, J.M.: Adaptive image retrieval using a graph model for semantic feature integration. In: Proceedings of the ACM International Multimedia Conference and Exhibition (2006). https://doi.org/10.1145/1178677.1178696
33. Wang, X., et al.: Heterogeneous graph attention network. In: The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019 (2019). https://doi.org/10.1145/3308558.3313562
34. Yang, C., Liu, Z., Zhao, D., Sun, M., Chang, E.Y.: Network representation learning with rich text information. In: IJCAI International Joint Conference on Artificial Intelligence (2015)

35. Zahalka, J., Rudinac, S., Worring, M.: Interactive multimodal learning for venue recommendation. IEEE Trans. Multimed. (2015). https://doi.org/10.1109/TMM.2015.2480007
36. Zhang, D., Yin, J., Zhu, X., Zhang, C.: MetaGraph2Vec: complex semantic path augmented heterogeneous network embedding. In: Phung, D., Tseng, V.S., Webb, G.I., Ho, B., Ganji, M., Rashidi, L. (eds.) PAKDD 2018. LNCS (LNAI), vol. 10938, pp. 196–208. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93037-4_16