Machine Learning for Graphs

Modality Encoder Dropout for Multi-Modal Relational Graph-Convolutional Neural Networks

Demothi Millarson

studentnumber: 2710380

Abstract. This work revisits "End-to-End Learning on Multimodal Knowledge Graphs" and reproduces the Multimodal Relational Graph-Convolutional Network (MR-GCN) pipeline. We provide an extension by introducing a Modality Encoder Dropout mechanism to fascilitate its generalization. Contemporary Knowledge Graphs contain heterogeneous data including numerical, temporal, textual, spatial and visual modalities. Message passing models designed to encapsulate these data can suffer from noninformative features disrupting the learning signal. Our reproduction implements the original MR-GCN pipeline with some implementation caveats. We implement dedicated modality-aware preprocessing and neural encoders to perform node classification on real-world Knowledge Graphs. Extending this framework, we propose a stochastic modality encoder dropout to enhance the learning of synergistic modality interactions during training, thereby escaping local optima and improving generalization to unseen data. The approach entails entirely foregoing the neural encoder step for some random subsample of the modality encoders. Experiments on a restrained DMG dataset indicate that our proposed mechanism decreases stability during training, increases volatility in experiments over less epochs. This behavior appears to allow for better optimization over longer training processes, however the addition of appropriate regularization methods-some of which were removed due to the restricted scope of the reproduction-could allow the MR-GCN to exhibit similar behavior in the no dropout setting.

Keywords: Multimodal Message Passing · Heterogeneous Knowledge Graphs · Optimization · Deep Learning

1 Introduction

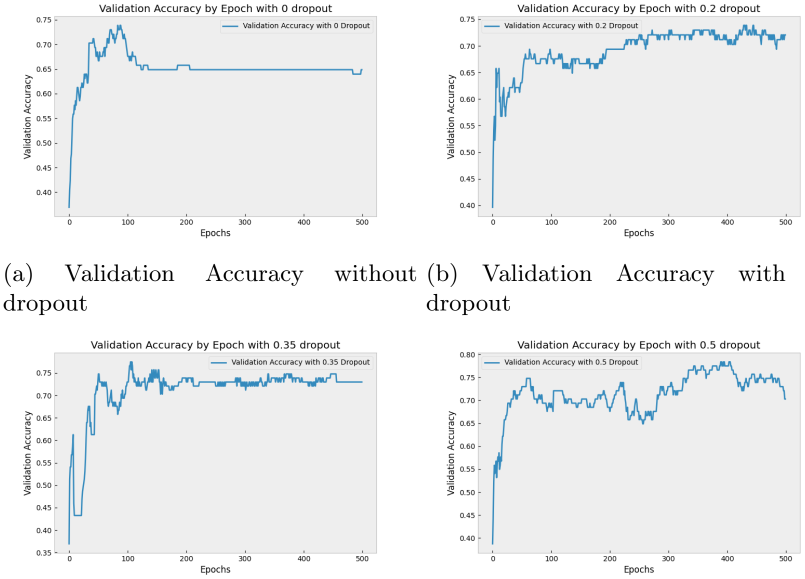
Knowledge graphs (KGs) are becoming more ubiquitous over the years. KGs are commonly used to formalize domain-specific information in a structured way. They facilitate easy inter-domain knowledge integration, query-answering, and other reasoning capabilities. Although machine learning methods may be used to overcome certain well-known issues with knowledge graphs, such as the identification of incomplete or uncertain data [5, 8], the limitations in reasoning abilities of deep learning methods by themselves may be enhanced by knowledge graphs, restricting them to methodical reasoning structures. Since KGs are especially well-suited for encoding diverse types of data within a rigorous relational structure, challenges and opportunities emerge when integrating machine learning methods with KGs to effectively reason across multimodal data. The paper proposes the MR-GCN [10], a method that uses dedicated encoders to process multimodal data, including numerical-, temporal-, spatial, textual- and image data to generate embeddings for end-to end systems. Additionally, a relational model leverages different edge weights to capture relations between objects. While MR-GCN represents a significant advancement over traditional GCNs, it still experiences some limitations. One of these limitations is its below expected performance when non-informative modalities are included. To mitigate this short-coming we propose Modality Encoder dropout. A method that excludes a random subset of the encoders from the training process each epoch. The aim of this approach is to enhance the model's robustness to noisy data, as well as to stabilize the training process and improve generalization to unseen data, with minimal computational overhead. Additionally, we would like to augment the training process of the MR-GCN in such a way that it learns to synergistically combine modalities.

1.1 Contribution

- In this work we propose and test the idea of modality-specific encoder dropout mechanism for deep end-to-end learning on multimodal knowledge graphs. This approach aims to force the network to parse less informative modalities and discover less dominant gradient signals to enhance synergy between heterogeneous features.
- We introduce encoder-level dropout as a regularization method to stabilize training and allow for loss landscape shifts to escape local optima during training.
- We carry out systematic experiments on the DMG dataset using different encoder-level dropout configurations to evaluate its impact on deep learning. The experiments are reproducible through the code provided on Github

2 Related Work

Earlier work in multimodal message passing by [4] proposed the MM-GNN for vision question-answering. Their method aims to extract multi-modal informa-



(a) Validation Accuracy without dropout

(b) Validation Accuracy with 0.2 dropout

(c) Validation Accuracy with 0.35 dropout

(d) Validation Accuracy with 0.5 dropout

Fig. 1: Validation accuracy by Encoder Dropout Configuration over 500 Epochs

tion from visual data. Their approach includes creating a multimodal knowledge graph latent space that is reasoned over by a modality-aware attention aggregation. Their approach to modality fusion is an iterative bi-directional fusion during message passing. Although this allows their model to implicitly model conditional dependency structures within the graph, the lack of relational awareness renders it less suitable for reasoning over KGs. A key takeaway from this approach is that its iterative refinement allows for modelling of more fine-grained multimodal interactions. Furthermore, the emphasis on translating embeddings into each other's representational spaces may lead to information loss.

Earlier work by [9] focused on incorporating multimodal data in an informative fashion, encoding literals separately from nodes, using the late fusion approach, in contrast to [4] to more truthfully represent semantic discrepancies between modalities. Character-level tokenization, trigonometric temporal encodings, and terminal-point aware spatial encodings introduced to handle textual, temporal and spatial vectorization are introduced into the pipeline to leverage our domain knowledge specific to highly structured semantic web formalisms. They advocate an end-to-end approach in order to align the encoders to the downstream task. This approach still lacked some of the more specialized neural encoder architectures proposed in [10].

3 Background

In this section, we will explain some of the foundations that underlie both the MR-GCN framework and our extension with modality encoder dropout. We review the structure and semantics of knowledge graphs and outline the machine learning methods specifically graph neural networks for heterogeneous data that forms the backbone of our approach.

3.1 Knowledge Graphs

KGs offer a structured way to represent entities and their relationships. In a KG, nodes represent entities and edges denote the semantic relations between them. This structure is often encoded using triples (subject, predicate, object). Common formalisms such as the Resource Description Framework (RDF) allow for processing of large amounts of data, while retaining semantic meaning through rigorously defined semantics. For example

$$x \, \text{rdf:type} \, Y$$

precisely means

$$x \in Y$$

This well defined structure allows us to capture complex semantics in a machine processable way. Over time these KGs have grown in suffistication, allowing for the inclusion of heterogeneous literals that may carry vital domain-specific information for accomplishing downstream tasks. Thus, the need for effective multimodal message passing has become more ubiquitous.

3.2 Graph Neural Networks for Heterogeneous Data

Graph Convolutional Neural Networks (GCNs) extend the idea of convolutions to apply to graphs. Since nodes in a graph can have an arbitrary amount of neighbours, some aggregation operation must take place to transform this various-sized convolution to the tabular format expected by traditional Deep Learning models. The method proposed in [7] uses the following message passing mechanism

$$H^{(l+1)} = \sigma\Big(AH^{(l)}W\Big)$$

where A is the Adjacency matrix, $H^{(l)}$ is the feature matrix at layer l, and σ is some nonlinearity. This lays the ground work for the models used in this work. Specifically to the R-GCN, which extends this idea with relation-specific weights, allowing for a more expressive encapsulation of relational structure. It is performed using

$$H^{(l+1)} = \sigma\Big(\sum_{r \in R} A_r \, H^{(l)} \, W_r\Big)$$

where A_r is the row- and column normalized adjacency matrix for relation r. As explained in [8]

4 Research Reproduction

4.1 An overview of the MR-GCN pipeline

Our reproduction involves the reconstruction of the end-to-end multi-modal system as described in [10]. The preprocessing steps are: normalizing the numerical data, centering the spatial data around zero, and labeling the shape and sub-shape terminal points using a 3 class one-hot encoding as described in [1]. Temporal data are encoded trigonometrically using the following function.

$$f_{\text{trig}}(\phi, \psi) = \left[\sin\left(\frac{2\pi\phi}{\psi}\right), \cos\left(\frac{2\pi\phi}{\psi}\right)\right]$$

This results in a vector of size $2 * n$ where n is the size an input vector of

$$\left[century, decade, year, month, day, hour\right]$$

Centuries are processed linearly ranging from -99 to 99. Other temporal entries are processed by the trigonometric function. For textual information, character-level tokenization is used. Visual information is transformed from URL-safe bytes into RGB-Tensors. This concludes the preprocessing vectorization. It is crucial to emphasize that the preprocessing vectorization methods are designed to be injective and maximally informative, ensuring that the unique properties of each modality are preserved throughout the transformation. An injective mapping guarantees that distinct inputs result in distinct vectors, enhancing the expressiveness of the downstream R-GCN. The multimodal Relational Graph Convolutional Network (MR-GCN) architecture uses dedicated neural encoders for all modalities, except numerical. Temporal Convolutional Networks (TCN) are used for spatial and textual encoding. The MobileNet architecture [6] is used for image encoding. Temporal encoding is done with a Multi-Layer Perceptron (MLP). These embeddings form a late-fusion [2] feature matrix, where each embedding type inhabits its own column space. The feature matrix is concatenated with the identity matrix, such that

$$H_0 = \left[I \; F\right]$$

where I is the identity matrix and F is the original feature matrix.

The message passing algorithm follows

$$H_1 = \sigma\left(\sum_{r \in R} A_r H_I W_r^I + A_r H_F^0 W_r^F\right)$$

where $A_r H_F^0 W_r^F$ is the aggregation of the multimodal features H_0. This function is applied to the first layer.

Any subsequent layer l is updated as follows

$$H_{l+1} = \sigma\left(\sum_{r \in R} A_r H_l W_r^I\right)$$

4.2 Reproduction Caveats

Several adjustments were made to the scope of the reproduction due to CUDA environment challenges and limited computational resources. Inverse edge relations were excluded, as well as basis decomposition and block decomposition in the message passing system. The reproduction used dense tensors instead of sparse tensors, restricting scalability, but allowing for manageable reproduction within the time frame. The visual encoder differs from the original MobileNet architecture in that it cannot perform batch normalization due to small batches. The 'large' text encoder uses the smaller architecture from the 'medium' text encoder to allow the entire end-to-end MR-GCN model to fit in working memory. The reproduction is limited to node classification on the DMG dataset in a split literals setting.

5 Research Extension

In this paper the MR-GCN framework is extended by incorporating a modality encoder dropout mechanism. In the original design, each modality-specific encoder contributes to the formation of the feature matrix. In some scenarios non-informative or noisy modalities can degrade performance and reduce generalization. The dropout mechanism randomly excludes a subset of modality encoders during each training epoch. The intuition is that by stochastically dropping entire modalities, the overall model learns to express less informative features to the best of its ability, relying less on the modality that contributes most to the model's performance. This method aims to enhance the synergistic effect between features that was missing in the original work. Another intuition is that 'shifting' the loss landscape between epochs helps the model escape local optima during training. Besides these effects, similar to regular dropout, modality-specific dropout may have a regularizing effect. This may result in better generalization to unseen data. The modality encoder dropout is implemented in several configurations, defined by the dropout probability $d \in \{0, 0.2, 0.35, 0.5\}$. When an encoder is dropped out, its contribution to the feature matrix is left out entirely, saving on computational complexity. The intuition for why encoder-level dropout may be more useful than just feature-level dropout is that it forces the Graph Network to iteratively rediscover different multimodal associations through the remaining modalities.

5.1 Experiments & Results

The experiments used a connected random subsample of the DMG dataset of 4000 edges instead of the entire dataset of 777,000 edges. This reproduction only focuses on the DMG dataset, rather than a broad mixture of real-word and synthetic datasets as in [10]. The subsample used in the experiments is included on Github as a pickle file for reproducibility. The train, validation and test set splits adhered to the original split proposed in [3].

The first experiment Two experiments were conducted. In the first experiment 2, the MR-GCN was trained over 100 epochs six times for every dropout configuration $d \in \{0, 0.2, 0.35, 0.5\}$. Due to a constrained setting and limited runs, and a discrepancy in the initial experiment pipeline, sample size and the amount of epochs per run are reduced. Something else to note is that basis- and block decomposition could have a regularizing effect that is not present in this reproduction, leading to an over-estimation of the effects of Encoder-level dropout. Overall, modality encoder dropout does not appear to improve model performance over short training durations on a small dataset. The approach both reduces the overall performance of the model on the test set, and makes the performance more volatile.

The second experiment In the second experiment 1, the MR-GCN was trained for an extended duration of 500 epochs, using the same dropout configurations $d \in \{0, 0.2, 0.35, 0.5\}$ as in the initial setup. The aim was to evaluate whether longer training durations allow the model to harness the expected benefits of modality encoder dropout. While the no dropout configuration plateaued early at approximately 65%, failing to improve beyond that point, all configurations employing encoder dropout continued to improve. Validation accuracy reached above 70% over the 500 epoch period. The dropout configurations show higher variance, indicating exploratory behavior that suggests the model may not have finished training, and would still be able to find lower optima in the loss landscape. These configurations with dropout achieved higher accuracy scores on the test set, as seen in 1.

Limitations and evaluation The reduced scope of the experiment should be noted in evaluating these results. A smaller subset of the dataset was used to perform the experiments, restricting the statistical power of these results. Furthermore due to the restricted scope of the reproduction, certain regularization techniques were omitted from the model. These regularization techniques could significantly improve the validation- and test accuracy of the model in the no dropout configuration, likely reducing the difference in performance between configurations Therefore, some caution is advised in interpreting the results of this experiment.

Dropout Rate	**Epochs**	**Test Accuracy**
0.00	500	0.7319
0.20	500	0.7938
0.35	500	0.8144
0.50	500	0.7938

Table 1: Test set accuracy for different Modality Encoder Dropout rates after 500 epochs of training using Adam with a learning rate of 0.01.

6 Discussion & Conclusion

Although the reproduction was successfully implemented, using dense tensors and reducing the complexity of some encoders and the R-GCN may have significantly altered the results of the experiments. The regularization effects of augmenting the model with Modality Encoder Dropout may be over-estimated, due to not incorporating the R-GCN regularization methods and limiting the normalization layers of the visual encoder. The modality encoder dropout solution should have been compared to a robust pipeline that incorporates batch-normalization and consistent dropout layers. The results from the second experiment indicate that shifting the loss landscape by modality allows the model to escape local optima and effectively perform deep learning on multimodal graphs. All tests were done in a setting including all modalities.

6.1 Future work

Future work should provide an ablation study for a sound feature-level dropout architecture to an encoder-level dropout. Another avenue worth exploring would be to enhance the aggregation step with variance-based aggregation to provide a more informative description of the message population to a downstream message passing layer. Should computational resources allow, augmenting the feature vector with relation-aware Local Clustering Coefficients during a preprocessing step may allow the R-GCN more discriminative power between relational structures, thus improving expressivity. Methods that require multi-modal latent spaces might benefit from combining iterative refinement with late fusion by encoding feature spaces seperately and iteratively refining a conditionally dependent interaction space before feeding the feature vector to a down-stream message passing model. This could counteract the over-smoothing problem GCNs tend to face, while preserving a rich interaction structure. The promising trends observed in our experiments calls for a more detailed ablation study in a full-scope reproduction setting. Such a study should aim to verify whether the improvement in validation accuracy is due to enhanced synergistic interactions between modalities. The individual and synergistic contributions of each modality should be tested with the proposed mechanism in place. These interactions should be measured against a setting with robust regularization in place.

References

1. van 't Veer, R., Bloem, P., Folmer, E.: Deep Learning for Classification Tasks on Geospatial Vector Polygons, (2019). arXiv: 1806.03857 [stat.ML]. https://arxiv.org/abs/1806.03857.
2. Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2019)

3. Bloem, P. *et al.*: kgbench: A Collection of Knowledge Graph Datasets for Evaluating Relational and Multimodal Machine Learning. In: Eighteenth Extended Semantic Web Conference - Resources Track (2021). https://openreview.net/forum?id=yeK_9wxRDbA
4. Gao, D. *et al.*: Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12743–12753 (2020). https://doi.org/10.1109/CVPR42600.2020.01276
5. Hogan, A. *et al.*: Knowledge Graphs. ACM Comput. Surv. **54**(4) (2021). https://doi.org/10.1145/3447772
6. Howard, A.G. *et al.*: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, (2017). arXiv: 1704.04861 [cs.CV]. https://arxiv.org/abs/1704.04861.
7. Kipf, T.N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks, (2017). arXiv: 1609.02907 [cs.LG]. https://arxiv.org/abs/1609.02907.
8. Schlichtkrull, M. *et al.*: Modeling Relational Data with Graph Convolutional Networks, (2017). arXiv: 1703.06103 [stat.ML]. https://arxiv.org/abs/1703.06103.
9. Wilcke, W.X. *et al.*: End-to-End Entity Classification on Multimodal Knowledge Graphs. ArXiv **abs/2003.12383** (2020). https://api.semanticscholar.org/CorpusID:214693100
10. Wilcke, W.X. *et al.*: End-to-End Learning on Multimodal Knowledge Graphs, (2023). arXiv: 2309.01169 [cs.LG]. https://arxiv.org/abs/2309.01169.

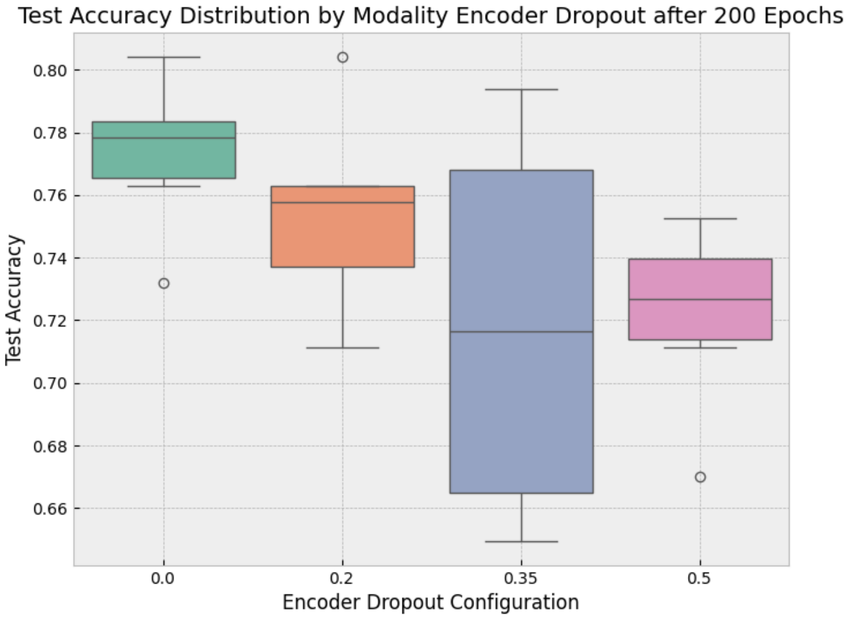


Fig. 2: Test-set Accuracy by Encoder Dropout Configuration