Multimodal graph networks

Overview

- Multimodal survey
    - Multimodal disciplines
    - Multimodal representations
- Multimodal VAE
- GNN survey
    - 3 frameworks
- Papers
    - Neuromatch
    - Graph matching networks
    - Graph Auto Encoder
    - Neural relational inference
- Discussion
    - Input/output, latent
    - Which graphs
    - Graph similarities

Multimodal learning

- Multimodal problems
  - Alignment
    - Explicit (goal of model)
    - Implicit (side-effect of model/loss)
  - Translation
- Further disciplines
  - **Representations**
  - Co-Learning (Transfer learning)
  - Fusion (Inference)
- Fusion types
  - Early fusion
    - Concat (or something easy) almost at start
  - **Late fusion**
    - **Train unimodal first, then combine later**

Multimodal representations

- Representation wishlist
  - Similarity in latent space means similarity in concept space
  - Usefulness for discriminative tasks
  - **Can miss modalities**
  - **Can fill modalities**
- Representation types
  - **Joint representation → in same space**
    - **Autoencoder**
  - Coordinated representation → in comparable space
    - Canonical correlation analysis

MVAE

$$p_\theta(x_1, x_2, ..., x_N, z) = p(z)p_\theta(x_1|z)p_\theta(x_2|z) \cdots p_\theta(x_N|z)$$

- Assume modalities are conditionally independent
- Approximate inference network

  → Product of experts

$$p(z|x_1, ..., x_N) \propto \frac{\prod_{i=1}^{N} p(z|x_i)}{\prod_{i=1}^{N-1} p(z)} \approx \frac{\prod_{i=1}^{N} [\tilde{q}(z|x_i)p(z)]}{\prod_{i=1}^{N-1} p(z)} = p(z)\prod_{i=1}^{N} \tilde{q}(z|x_i).$$

- ELBO
  - Not 2^N, but in O(N)

$$\text{ELBO}(x_1, ..., x_N) + \sum_{i=1}^{N} \text{ELBO}(x_i) + \sum_{j=1}^{k} \text{ELBO}(X_j)$$

3 GNN-Frameworks

- Message Passing (simple, popular)
  - Edge update
    - M tangled NN for all edges
  - Node update
    - U tangled NN for all nodes
  - End: global aggregation
- Non-local NN (graph attention)
  - f → attention
  - g → NN

$$\mathbf{m}_v^{t+1} = \sum_{w \in \mathcal{N}_v} M_t\left(\mathbf{h}_v^t, \mathbf{h}_w^t, \mathbf{e}_{vw}\right)$$

$$\mathbf{h}_v^{t+1} = U_t\left(\mathbf{h}_v^t, \mathbf{m}_v^{t+1}\right)$$

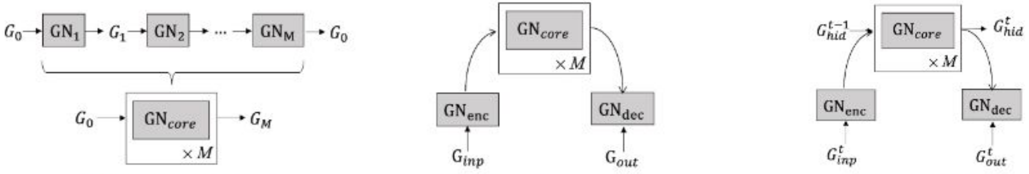
$$\hat{\mathbf{y}} = R(\{\mathbf{h}_v^T | v \in G\})$$

$$\mathbf{h}_i' = \frac{1}{\mathcal{C}(\mathbf{h})} \sum_{\forall j} f(\mathbf{h}_i, \mathbf{h}_j) g(\mathbf{h}_j)$$

3 GNN-Frameworks

- Graph Networks (general)
  - Edge updates
    - Every edge
    - All incident edges of a node
    - All edges to global
  - Node updates
    - Every node
  - Global update
    - One global from
      - Prev. global
      - All nodes
      - All edges

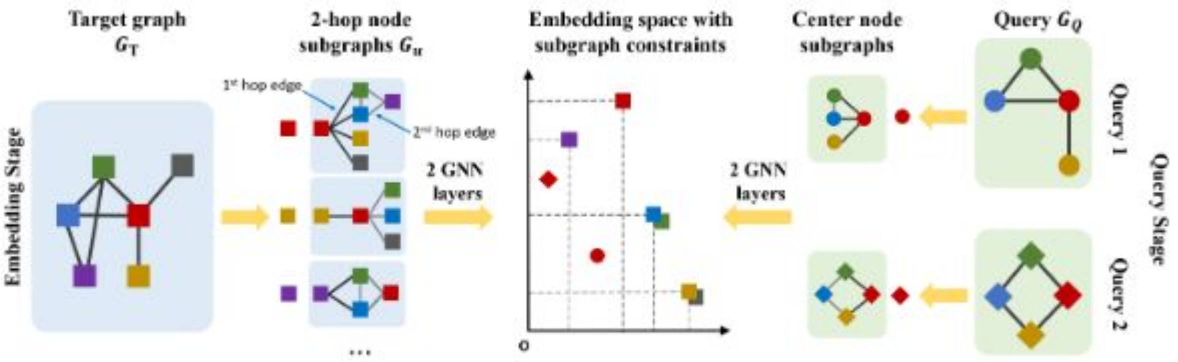
$$
\begin{aligned}
\mathbf{e}'_k &= \phi^e\left(\mathbf{e}_k, \mathbf{h}_{r_k}, \mathbf{h}_{s_k}, \mathbf{u}\right) & \bar{\mathbf{e}}'_i &= \rho^{e\to h}\left(E'_i\right)\\
\mathbf{h}'_i &= \phi^h\left(\bar{\mathbf{e}}'_i, \mathbf{h}_i, \mathbf{u}\right) & \bar{\mathbf{e}}' &= \rho^{e\to u}\left(E'\right)\\
\mathbf{u}' &= \phi^u\left(\bar{\mathbf{e}}', \bar{\mathbf{h}}', \mathbf{u}\right) & \bar{\mathbf{h}}' &= \rho^{h\to u}\left(H'\right)
\end{aligned}
$$



(a) Sequential GN blocks (b) Encode-process-decode (c) Recurrent GN blocks

Fig. 3. Examples of architectures composed by GN blocks. (a) The sequential processing architecture; (b) The encode-process-decode architecture; (c) The recurrent architecture.

NeuroMatch



- Message Passing
- Curriculum training
-
- Check all neighbourhoods and try to match them in embedding space
- Much faster compared to combinatorial
-
- Custom Hinge loss in embeddings
    - Z is 2-dimensional

$$\mathcal{L}(z_q, z_u) = \sum_{(z_q, z_u) \in P} E(z_q, z_u) + \sum_{(z_q, z_u) \in N} \max\{0, \alpha - E(z_q, z_u)\}, \text{where}$$

$$E(z_q, z_u) = ||\max\{0, z_q - z_u\}||_2^2$$

Graph Matching Networks

- Message Passing
- Additional Edges
    - Between graphs
    -
- Embeddings only in contrast to another graph
    - No standalone-embedding, if using a reference graph, then just another NN

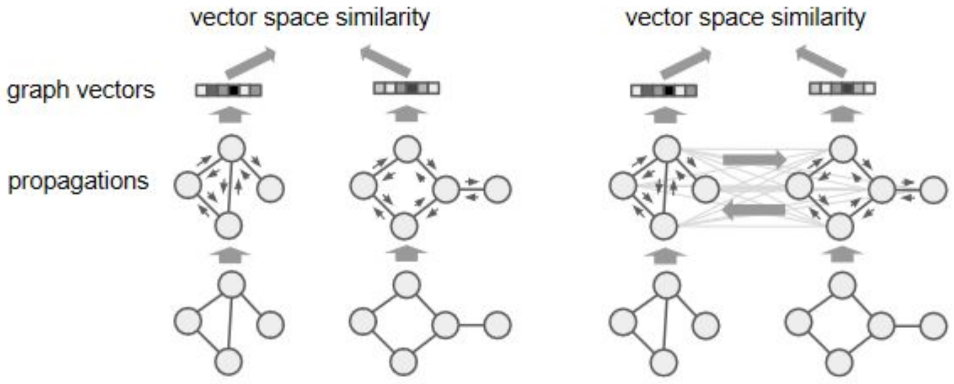


Figure 2. Illustration of the graph embedding (left) and matching models (right).

Graph Auto Encoder

- Exists
    - by Kipf & Welling
- Try to model the adjacency matrix A
    - Amount of nodes given
- μ,σ estimated with GCN
- Every node has a z
- p(Z) was a weakness
    - Keep graph sparse

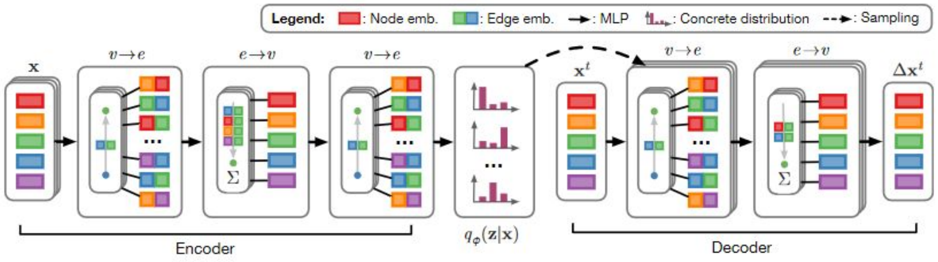
$$q(\mathbf{Z} \,|\, \mathbf{X}, \mathbf{A}) = \prod_{i=1}^{N} q(\mathbf{z}_i \,|\, \mathbf{X}, \mathbf{A}), \quad \text{with} \quad q(\mathbf{z}_i \,|\, \mathbf{X}, \mathbf{A}) = \mathcal{N}(\mathbf{z}_i \,|\, \boldsymbol{\mu}_i, \operatorname{diag}(\boldsymbol{\sigma}_i^2)).$$

$$p(\mathbf{A} \,|\, \mathbf{Z}) = \prod_{i=1}^{N} \prod_{j=1}^{N} p(A_{ij} \,|\, \mathbf{z}_i, \mathbf{z}_j), \quad \text{with} \quad p(A_{ij} = 1 \,|\, \mathbf{z}_i, \mathbf{z}_j) = \sigma(\mathbf{z}_i^\top \mathbf{z}_j),$$

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{Z}|\mathbf{X},\mathbf{A})}\left[\log p(\mathbf{A} \,|\, \mathbf{Z})\right] - \mathrm{KL}\left[q(\mathbf{Z} \,|\, \mathbf{X}, \mathbf{A}) \,||\, p(\mathbf{Z})\right],$$

$$\hat{\mathbf{A}} = \sigma(\mathbf{Z}\mathbf{Z}^\top), \quad \text{with} \quad \mathbf{Z} = \mathrm{GCN}(\mathbf{X}, \mathbf{A}).$$

Neural Relational Inference



- Message passing
- Fully connected graph
  - But some edges are “no edge”
- Edges have types, which enforce a different cell architecture
-
- Encoder is a GNN, as in the image
  - Z has an entry for each edge
- Z is a graph
- Decoder predicts multiple timesteps from a single latent representation
  - Also is a GNN, can be extended with RNN cells

The ELBO objective, Eq. 3, has two terms: the reconstruction error $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$ and KL divergence $\mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})]$. The reconstruction error is estimated by:

$$-\sum_j \sum_{t=2}^{T} \frac{||\mathbf{x}_j^t - \boldsymbol{\mu}_j^t||^2}{2\sigma^2} + \text{const} \tag{18}$$

while the KL term for a uniform prior is just the sum of entropies (plus a constant):

$$\sum_{i \neq j} H(q_\phi(\mathbf{z}_{ij}|\mathbf{x})) + \text{const}. \tag{19}$$

Input, output, latent?

- Not main focus
    - (GNNs already used for multimodal)
    - Image → graph exists
    - Text → graph exists
    - Modality → graph exists (generally)
- Operations
    - Graph → graph
    - Graph → set of node representations
    - Graph → vector/global embedding
- Where latent combination?
    - Graph x Graph
    - Embedding x Embedding
- What combination?
    - Neural
    - Exists a multimodal convolutional kernel for images
    - Combinatorial (à la **[Weisfeiler-Lehman]**)

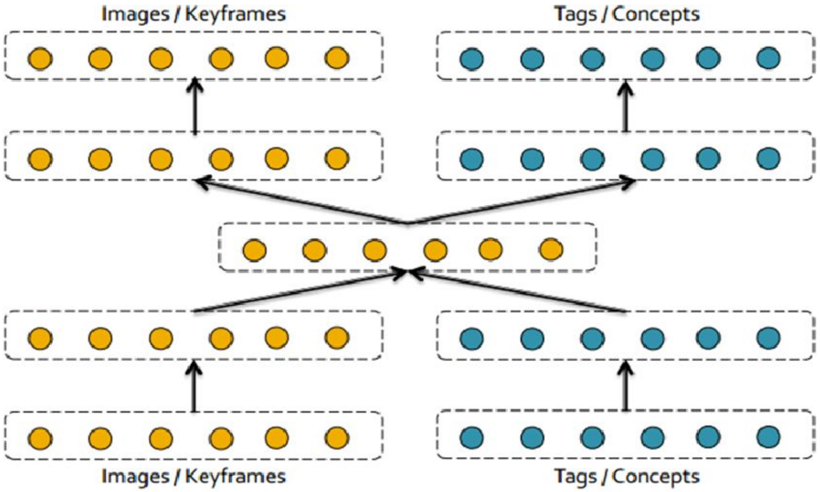


Figure 3. Multimodal Stacked Denoising Autoencoders

$$\mathbf{m}_v^{t+1} = \sum_{w \in \mathcal{N}_v} M_t\left(\mathbf{h}_v^t, \mathbf{h}_w^t, \mathbf{e}_{vw}\right)$$

$$\mathbf{h}_v^{t+1} = U_t\left(\mathbf{h}_v^t, \mathbf{m}_v^{t+1}\right)$$

$$\hat{\mathbf{y}} = R(\{\mathbf{h}_v^T \mid v \in G\})$$

What graphs to focus on?

- Graphs that represent a real situation
  - General graphs
  - Typed entities
- Dynamic graphs
- HMM graphs
  - DAG
- Causal graphs
  - DAG
- Have full adjacency matrix
  - For spectral methods
    - Still popular?

Graph similarities between modalities assumptions

- Same graph, different “graph center”
- Similar types
- Same nodes
- Same edges
- **None, but same “real” origin**