S7: Graph Information Bottleneck – Sufficient Learning of Representation

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Scribe: 7

Student: Pranjal Umesh Kalekar

CS 7840: Foundations and Applications of Information Theory (fa24)

https://northeastern-datalab.github.io/cs7840/fa24/

Lecturers: Wolfgang Gatterbauer, Javeed Aslam

Dec 10, 2024

Connecting Measure of Divergence to Measure of Similarity

- AS we know Kl divergence tells us how different the output probability distribution is from the original distribution, hence we could use it regularize the loss function.
- The information bottleneck principle aims to find a compressed representation of an input variable that retains as much information as possible about a relevant output variable.
- If measure of difference can be use as a regulariser then a measure that keeps control on compression and relevance can be utilized in the similar manner, right?
- KL Divergence is used to estimate the upper bound of the mutual information between graph-structured data and node representations in GIB.

GIB For Regularization and Overfitting Control:

Graph Information Bottleneck:

$$\min_{\mathbb{P}(Z|\mathcal{D})\in\Omega} \mathrm{GIB}_{\beta}(\mathcal{D}, Y; Z) \triangleq [-I(Y; Z) + \beta I(\mathcal{D}; Z)]$$

GIB's principle means finding a balance between relevance (how well the representation captures information about the output) and compression (how much the representation reduces the input's complexity).

$D \rightarrow actual\ distribution, Y \rightarrow final\ distribution, Z \rightarrow Current\ State\ of\ prediction$.

This forms the Markov chain $D \rightarrow Z \rightarrow Y$,

tracing his back to the Information bottleneck and by controlling the balance between relevance between Target distribution and the current state $Z \rightarrow I(Y,Z)$and compression between Z and $D \rightarrow I(D,Z)$will let us have a hold of over fitting using a hyperparameter β.

Prior distribution is assumed to be Gaussian for Z. The prediction loss is calculated by cross entropy within target and current state and compression difference is calculated by KL divergence.

Implementation Of this concept on PubMed Dataset

Creation of GIB-Loss function:

```
def compute_gib_loss(self, z, y, prior_dist, target_dist, edge_index):
    # Cross-entropy for I(Y; Z_X^(L)) - Task relevance
    prediction_loss = F.cross_entropy(z, y)

    # KL divergence for I(D; Z_X^(L)) - Compression term
    kl_div = torch.distributions.kl_divergence(target_dist, prior_dist).mean()

    # Combine losses
    gib_loss = prediction_loss + self.beta * kl_div
    return gib_loss
```

Balanced Accuracy for baseline model: 0.80

Balanced Accuracy for GIB_loss Model: 0.88

Effective 10% increase in the performance!

Code available at: [GitHub]

Further Advancements With This Idea:

- I set priors to gaussian distribution for Z but would be interesting to try different distributions.
- Would be interesting to observe behavior of attention in this training – GIB paper has mentioned a trieal with GAT layers namely GIB-Cat and GIB-Bern for categorical and Bernoulli probability distributions.
- Here I primarily applied GIB to the task of node classification. However, the principles of GIB could be extended to other graph-related tasks, such as:
  - Link prediction: Predicting the existence of edges between nodes.
  - Graph classification: Assigning labels to entire graphs.
- Would be interesting to observe effect GIB on class imbalance- Graphs are very versatile in application and imbalance is one of the very commonly occurring cases even in indrustrial, medical and alchemical environments.

References:

- https://arxiv.org/abs/2010.12811, Graph Information Bottleneck, Tailin Wu, Hongyu Ren, Pan Li, Jure Leskovec, Oct 2020
- Part of KLCE loss and Graph Information Bottleneck (GIB) on Network Datasets For Class Imbalance for CS7840-Fall24, Dec 09, 2024, Northeastern University