Korea Advanced Institute of Technology

Department of Mathematical Sciences

Semi-supervised graph node classification of CORA dataset using GCN

Author:
Junghyun Lee

Instructor:
Prof. Chang Dong Yoo

Homework #4

EE531: Statistical Learning Theory, Fall 2019

December 23, 2019

Abstract

In this homework, I have implemented a simple version of GCN(Graph Convolutional Network) to do a semi-supervised graph node classification of CORA dataset. The algorithm is based upon the work by Kipf & Welling, 2016.

(Coding done in Google Colaboratory.)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This is one of the most important components of the dissertation. It should begin with a clear statement of what the project is about so that the nature and scope of the project can be understood by a lay reader. It should summarise everything you set out to achieve, provide a clear summary of the project's background and relevance to other work and give pointers to the remaining sections of the dissertation which contain the bulk of the technical material.

Further information can be found here: `https://goo.gl/k2huN9`.

1.1 LaTeX code examples and formatting tips

Hello, here's a citation [**?**]. References are stored in a Bibtex file. Google Scholar and IEEExplore allow you to download citations of papers in Bibtex format from their search engine. Some people use JabRef (`http://www.jabref.org`) to manage their database of references.

This is an inline equation $\Gamma(t) = K_i e^{\sin^2(\omega_t)}$. The first paragraph appears without indent but the following ones will have an indentation.

This is an actual named equation:

$$v(x) = \frac{1}{2}\sin(2\omega t + \phi)e^{-jst} \tag{1.1}$$

where ω is the angular speed. Notice that symbols liks ω should be written in italics whereas measurement units such as V for Volts appear as normal text. This paragraph didn't have an indentation because the first sentence was linked to the definition of equation (1.1). A code snippet for an example program is shown in Listing 1.1.

Listing 1.1: Source code for *hello.m*

```
for i:=maxint to 0 do
begin
{ do nothing }
end;
Write('Case insensitive ');
Write('Pascal keywords.');
```

The characteristic parameters of the system are sumarised in Table 1.1. A figure is shown Fig 1.1, we don't necessarily know if this figure will appear below, above or elsewhere; therefore, the text should never refer to the figure with sentences such as *"As shown here:"*.

Parameter	Value	Units
P	1	kW
Q	0	kVAr

Table 1.1: Characteristic parameters of the system

Figure 1.1: Felix the Cat

Sometimes, the symbols in an equation are defined as follows[1]:

$$V(t) = A\sin(\omega t + \theta_0) \tag{1.2}$$

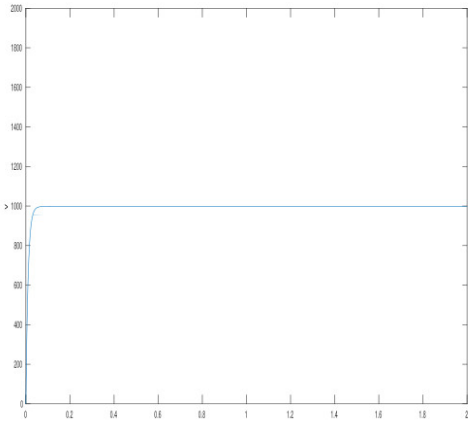
where V is a voltage waveform,
A is the amplitude of the voltage,
ω is the angular frequency,
t is the time.

1.1.1 A brief comparison between a proper plot and a horrible plot

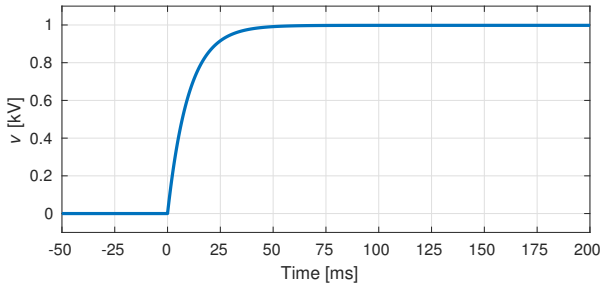
Figure 3.13 contains two plots of the same waveform. Subfigure 3.13(a) shows a badly formatted figure, Subfigure 3.13(b) shows a much better formatted figure. The problems with Subfigure 3.13(a), listed by order or relevance, are the following:

1. The font size is too small to be read properly.
2. The axes aren't labeled properly: the horizontal axis is not labeled and the units of the vertical axis are unknown. Further, symbols must be written in italics whereas numbers and units must be written as normal text.
3. The choice of limits for the axes is not good, the figure has wide useless empty spaces. The most relevant part of the waveform is the transient that happens between times t =0 and t =0.05 s, which is less than 10% of the timespan shown in the figure.
4. The figure has been scaled without keeping the original aspect ratio and fonts look narrower than they would if the figure had been scaled properly.
5. The plot doesn't have grid lines. This makes it hard to read the exact value (ie time, voltage) of points in the trace.
6. The width of the trace is too thin and may not be visible if printed in low resolution.
7. The choice of units of the vertical axis aren't the best. For example, in this case the plot would be easier to read if voltage had been expressed in kV instead of V.
8. The figure was exported as a bitmap (e.g. png, jpg, bmp) instead of being exported in vector format (e.g. eps, svg, pdf) and visual artifacts appear when the figure is scaled up or down in order to fit in the document.

[1]Some authors like to define their symbols this way.



(a) A horrible one.



(b) A proper one.

Figure 1.2: A figure with two subfigures.

Figure 1.3: Here's a large drawing of Felix the Cat that wouldn't fit in a portrait page

1.2 Objectives

1.3 Challenges

1.4 Contributions

Chapter 2

Background

The background section of the dissertation should set the project into context by relating it to existing published work which you read at the start of the project when your approach and methods were being considered. There are usually many ways of solving a given problem, and you shouldn't just pick one at random. Describe and evaluate as many alternative approaches as possible. The background section is often included as part of the introduction but can be a separate chapter if the project involved an extensive amount of research.

The published work may be in the form of research papers, articles, text books, technical manuals, or even existing software or hardware of which you have had hands-on experience. Don't be afraid to acknowledge the sources of your inspiration; you are expected to have seen and thought about other people's ideas; your contribution will be putting them into practice in some other context. However, you must avoid plagiarism: if you take another person's work as your own and do not cite your sources of information/inspiration you are being dishonest; in other words you are cheating.

Chapter 3

Evaluation

For my evaluation, I've used **Tensorboard** for plotting the learning(loss) curve and performance(accuracy) curve for both training and validating processes. Also, I've used external open source library(**pretty-print-confusion-matrix**[1]) for plotting the confusion matrix for the test set.

In all the plots, x-axis corresponds to the epoch. As for the graphs,

- Accuracy graph
  - Dark red: train accuracy
  - Bright red: validation accuracy
- Loss graph
  - Dark blue: train loss
  - Bright blue: validation loss

[1] `https://github.com/wcipriano/pretty-print-confusion-matrix`

3.1 Original model

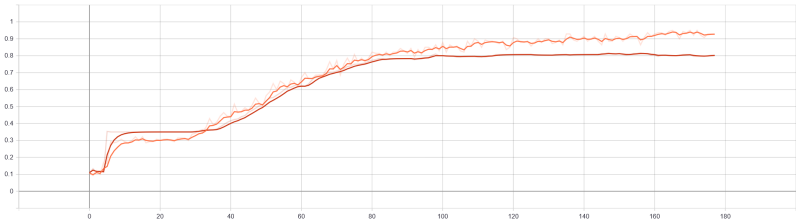


Figure 3.1: Accuracy graph

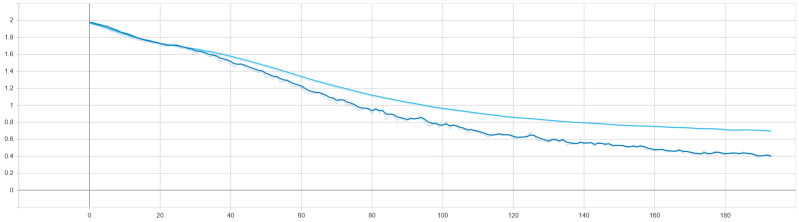


Figure 3.2: Loss graph

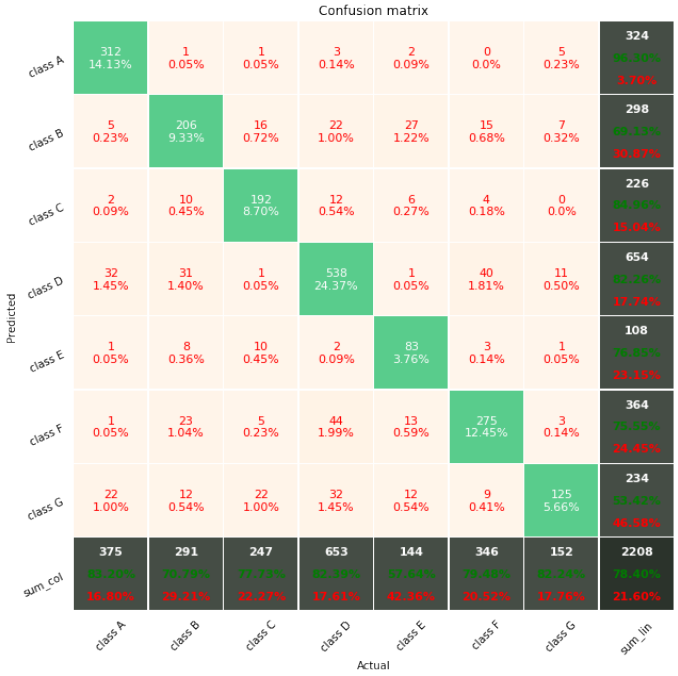


Figure 3.3: Confusion matrix

3.2 Modified Model (Ver. 1)

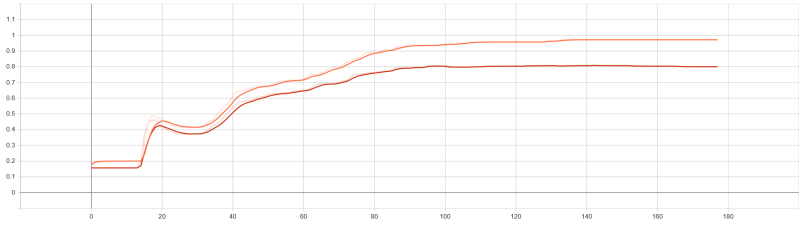


Figure 3.4: Accuracy graph

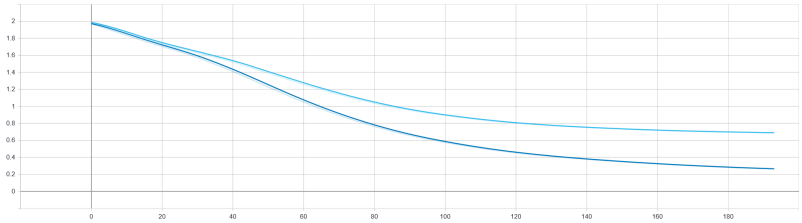


Figure 3.5: Loss graph

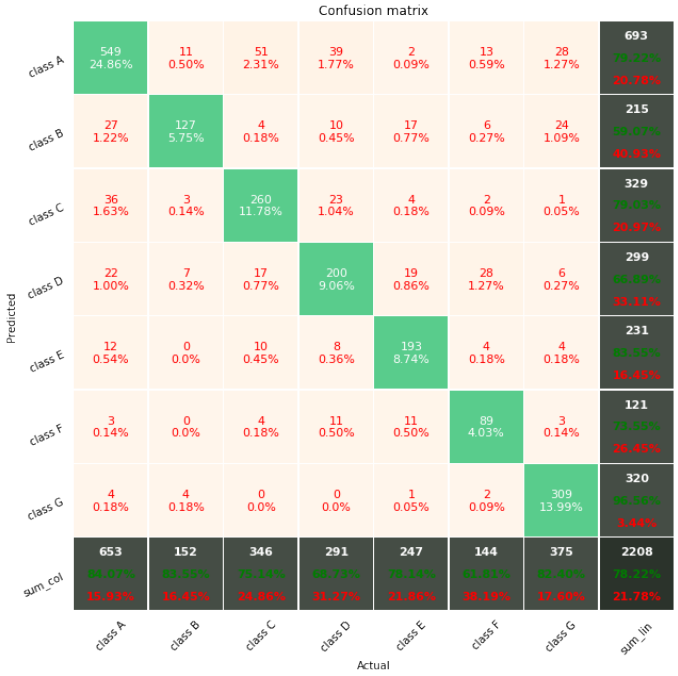


Figure 3.6: Confusion matrix

3.3 Modified Model (Ver. 2)

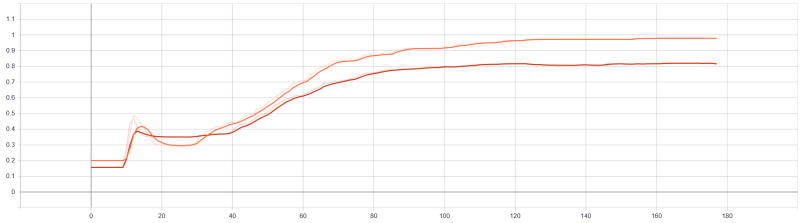


Figure 3.7: Accuracy graph

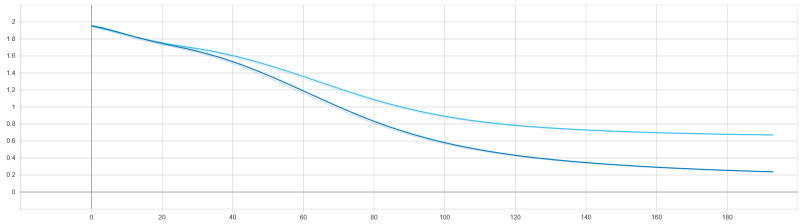


Figure 3.8: Loss graph

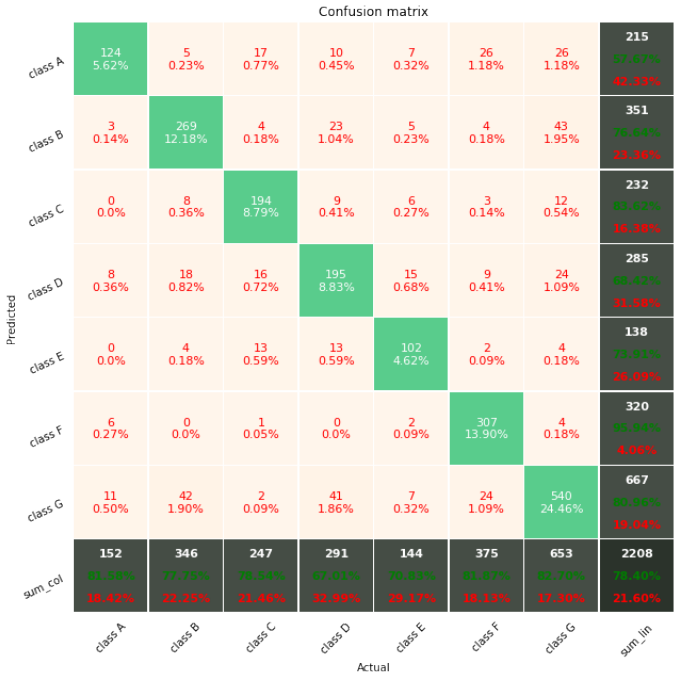


Figure 3.9: Confusion matrix

3.4 Modified Model (Ver. 3)

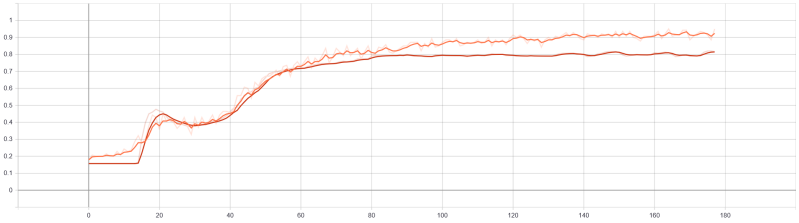


Figure 3.10: Accuracy graph

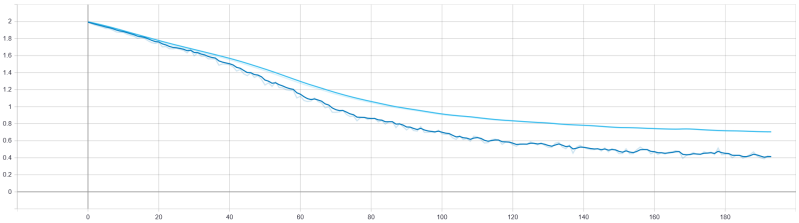


Figure 3.11: Loss graph

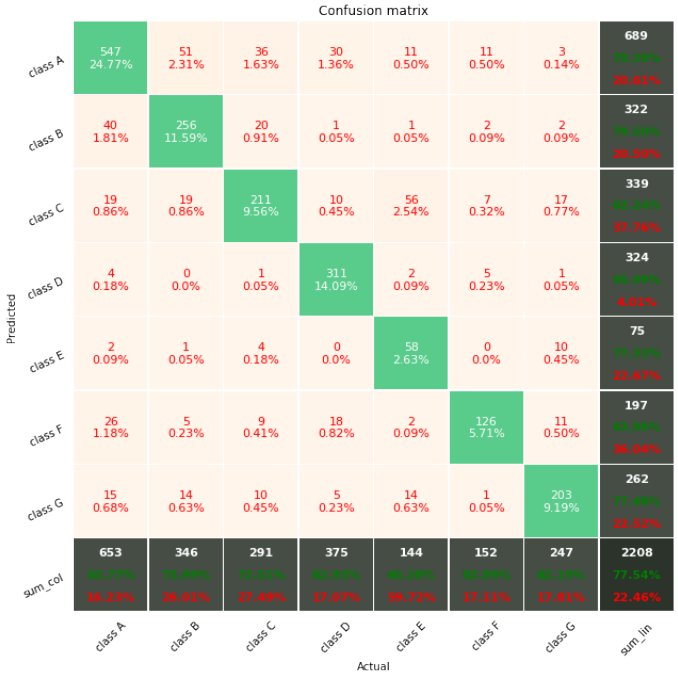


Figure 3.12: Confusion matrix

Chapter 4

Conclusion