COS4312 Deep Learning Lab5 IMDB

1. แบ่งข้อมูล 5,000 รีวิวสำหรับ training set 5,000 รีวิวสำหรับ validation set และ 40,000 รีวิวสำหรับ testing

https://colab.research.google.com/github/bpppoh/Mylearning/blob/main/4312/week6/week5_lab1.ipynb

2. แบ่งข้อมูล 15,000 รีวิวสำหรับ training set 10,000 รีวิวสำหรับ validation set และ 25,000 รีวิวสำหรับ testing

https://colab.research.google.com/github/bpppoh/Mylearning/blob/main/4312/week6/week5_lab2.ipynb

3. เปรียบเทียบและวิเคราะห์ผล

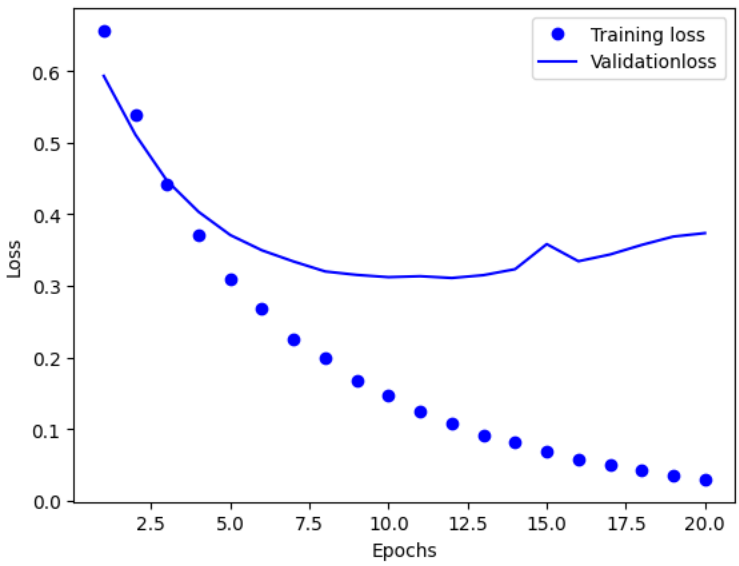
โค้ดจากไฟล์ทั้งสอง เป็นการนำข้อมูลคำจาก IMDB มาเทรน model โดยมีวัตถุประสงค์เพื่อการเรียนรู้บทบาท และความสำคัญของ Validation set โดยมีรายละเอียดการเทรนดังนี้

กลุ่มแรก แบ่งข้อมูลออกเป็น 3 ชุด ดังนี้ ชุดข้อมูลเพื่อเทรน 5,000 ข้อมูล , ชุดข้อมูลเพื่อการทำ Validate 5,000 ข้อมูล และชุดข้อมูลเพื่อทดสอบ 40,000 ข้อมูล เทรนโดยมีรายละเอียดดังนี้

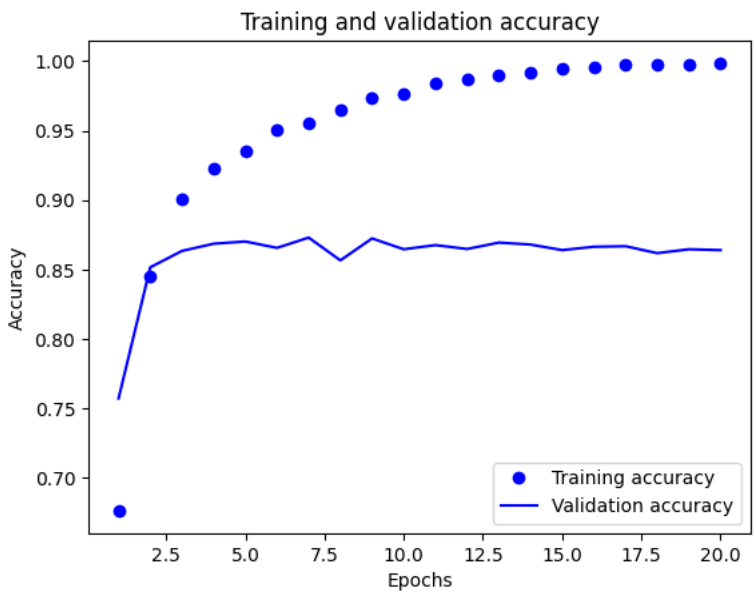
- Hidden Layer 2 Layers
- Layer 1 ชั้นมี Neural จำนวน 16 ตัว
- แต่ละ Neural ใช้ activation function เป็น Relu (Rectified Linear Unit)
- Epoch = 20
- Batch Size = 512

โดยได้ผลลัพธ์การเทรนออกมามีลักษณะดังนี้

- Accuracy ในแต่ละ Epoch มีค่าเพิ่มขึ้น โดยเริ่มที่ 0.9004 ใน Epoch ที่ 1 และสิ้นสุดที่ 0.9972 ใน Epoch ที่ 20
- Loss ในแต่ละ Epoch มีค่าลดลงโดยเริ่มที่ 0.6568 ใน Epoch ที่ 1 และสิ้นสุดที่ 0.0295 ใน Epoch ที่ 20
- Validated Accuracy มีแนวโน้มเพิ่มขึ้นในช่วง Epoch ที่ 1 – 2 และตั้งแต่ Epoch ที่ 3 เป็นต้นไป ค่าแกว่งตัวอยู่ในช่วง 0.85 – 0.87
- Validated Loss มีแนวโน้มลดลงตั้งแต่ Epoch ที่ 1 – 10 และตั้งแต่ Epoch ที่ 11 เป็นต้นไป ค่ามีแนวโน้มค่อย ๆ เพิ่มขึ้นเรื่อย ๆ ตั้งแต่ Epoch ที่ 11 เป็นต้นไป



กราฟเปรียบเทียบค่า Training loss และ Validation loss ในแต่ละ Epoch ของกลุ่มที่ 1



กราฟเปรียบเทียบค่า Training accuracy และ Validation accuracy ในแต่ละ Epoch ของกลุ่มที่ 1

กลุ่มสอง แบ่งข้อมูลออกเป็น 3 ชุด ดังนี้ ชุดข้อมูลเพื่อเทรน 15,000 ข้อมูล , ชุดข้อมูลเพื่อการทำ Validate 10,000 ข้อมูล และชุดข้อมูลเพื่อทดสอบ 25,000 ข้อมูล เทรนโดยมีรายละเอียดดังนี้

- Hidden Layer 2 Layers

- Layer 1 ชั้นมี Neural จำนวน 16 ตัว

- แต่ละ Neural ใช้ activation function เป็น Relu (Rectified Linear Unit)

- Epoch = 20

- Batch Size = 512

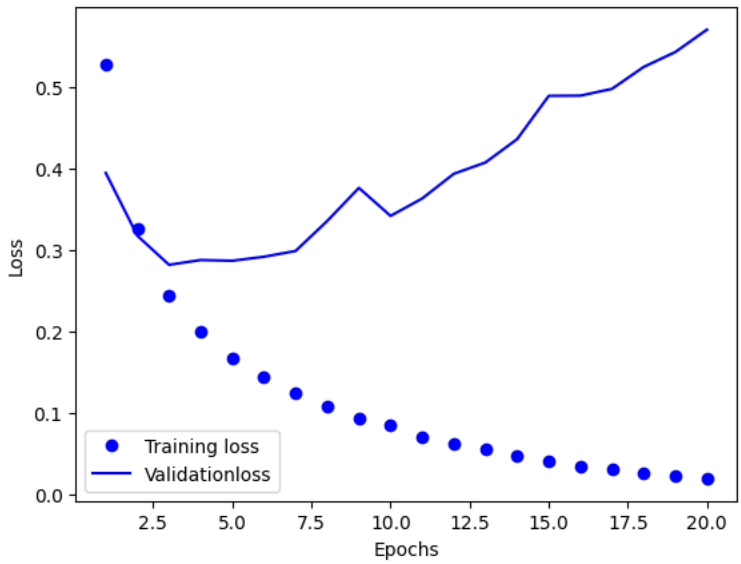
โดยได้ผลลัพธ์การเทรนออกมามีลักษณะดังนี้

- Accuracy ในแต่ละ Epoch มีค่าเพิ่มขึ้นเรื่อย ๆ โดยเริ่มที่ 0.7743 ใน Epoch ที่ 1 และสิ้นสุดที่ 0.9971 ใน Epoch ที่ 20

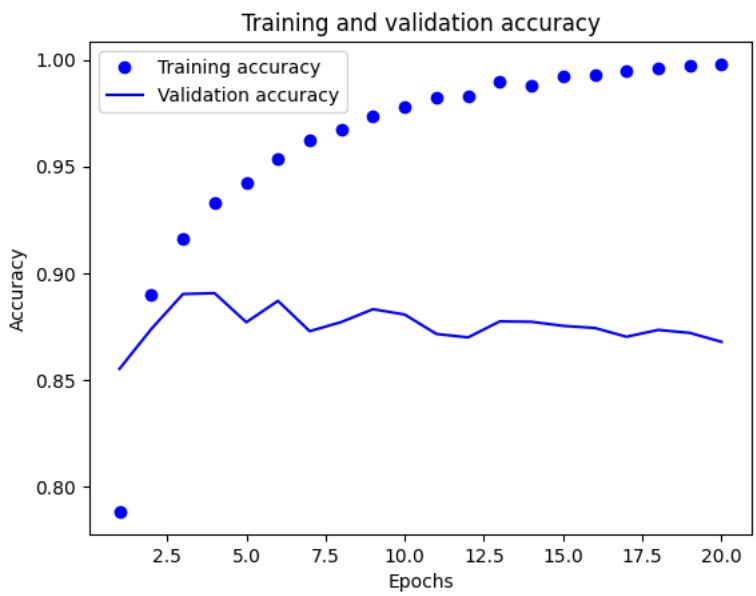
- Loss ในแต่ละ Epoch มีค่าลดลงเรื่อย ๆ โดยเริ่มที่ 0.5277 ใน Epoch ที่ 1 และสิ้นสุดที่ 0.0194 ใน Epoch ที่ 20

- Validated Accuracy มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ ในช่วง Epoch ที่ 1 - 4 และเริ่มมีแนวโน้มน้อยเรื่อย ๆ ตั้งแต่ Epoch ที่ 4 เป็นต้นไป

- Validated Loss มีแนวโน้มลดลงในช่วง Epoch ที่ 1 – 3 และเริ่มมีแนวโน้มเพิ่มขึ้นในช่วง Epoch ที่ 4 เป็นต้นไป



กราฟเปรียบเทียบค่า Training loss และ Validation loss ในแต่ละ Epoch ของกลุ่มที่ 2



กราฟเปรียบเทียบค่า Training accuracy และ Validation accuracy ในแต่ละ Epoch ของกลุ่มที่ 2

สรุปได้ว่า กลุ่มแรก ที่มี Training Data เพียง 5,000 ข้อมูล มีแนวโน้มที่จะ Overfit น้อยกว่า กลุ่มสองที่มี Training Data เป็น 15,000 เนื่องจาก Batch size ของทั้งสองกลุ่มมีค่าเป็น 512 เท่ากัน ทำให้กลุ่มที่สองมีจำนวนรอบในการปรับ Weight หรือกระบวนการ Optimization มากกว่ากลุ่มที่หนึ่ง จึงมีแนวโน้มที่จะเกิด Overfitting สูงกว่า

แนวทางการแก้ไขภาวะ Overfitting ในกลุ่มที่สอง ดำเนินการได้โดยการเพิ่มค่า Batch size ในกลุ่มที่สอง ให้มีค่าสูงขึ้น เพื่อลดจำนวนรอบการ Optimization ให้น้อยลง