01204466 การเรียนรู้เชิงลึก (Deep Learning)

รายงานโครงงาน : Sentiment Analysis on IMDB Movie Reviews using TextCNN

ผู้จัดทำ : ธนภัทร กาญจนรุจิวุฒิ

รหัสนิสิต : 6610505403

ภาควิชา : วิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์

ผู้สอน : Paruj Ratanaworabhan

ภาคต้น ปีการศึกษา 2568

บทคัดย่อ

โครงงานนี้มีวัตถุประสงค์เพื่อวิเคราะห์อารมณ์ของข้อความรีวิวจาก IMDB โดยจำแนกรีวิวออกเป็นเชิงบวกและเชิงลบ โดยการประยุกต์ใช้โมเดล Convolutional Neural Network ตัวผู้จัดทำต้องการศึกษาความสามารถของ Deep Learning ในทำความเข้าใจภาษาและอารมณ์ของข้อความ ผลการทดลองแสดงให้เห็นว่า TextCNN สามารถแยกแยะอารมณ์ของรีวิวได้ โดยได้ค่าความถูกต้องเฉลี่ย 81.6%

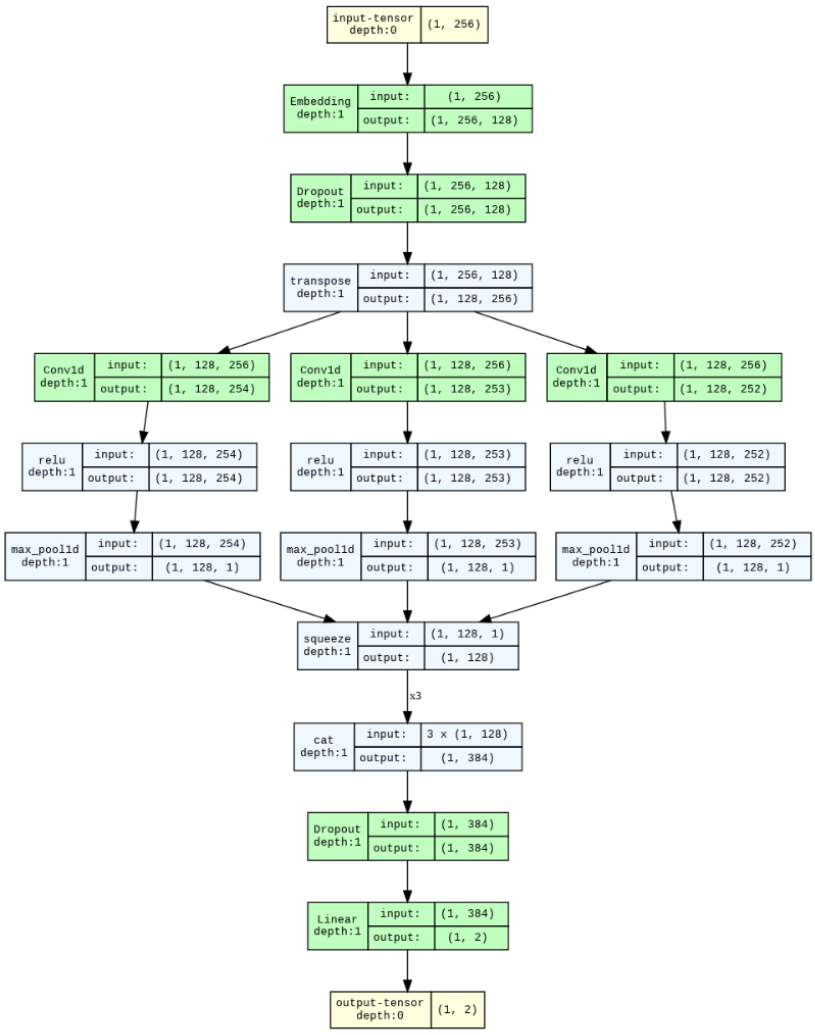
เหตุผลที่ใช้ Deep Learning แทนวิธีอื่น และการเปรียบเทียบ

จากการศึกษาพบว่า ก่อนมี Deep Learning วิธีที่นิยมคือการใช้ Bag-of-Words ซึ่งแม้จะใช้ง่ายและเทรนได้เร็ว แต่ไม่สามารถจับรูปประโยค หรือบริบทได้จริงๆ ในทางกลับกัน Deep Learning โดยเฉพาะ CNN สามารถเรียนรู้ feature ของคำได้ ทำให้เข้าใจอารมณ์ของ text ได้ดีกว่า

Model Architecture

Text CNN

- Embedding Layer: แปลงคำเป็นเวกเตอร์ขนาด 128 มิติ
- Convolution Layer: ใช้ kernel ขนาด 3, 4, 5 เพื่อดู pattern ของคำ
- Max Pooling Layer: ดึง feature เด่นที่สุดของแต่ละ filter ออกมา
- Dropout Layer (p=0.5): ป้องกัน overfitting
- Fully Connected Layer: รวม feature ทั้งหมดเพื่อจำแนกอารมณ์เป็น 2 คลาส (positive/negative)
- Activation Function: ใช้ ReLU ใน convolution layer และ Softmax ใน output



รายละเอียดของโค๊ด

ส่วนการจัดการข้อมูล:

ใช้ IMDBDataset ในการอ่านไฟล์รีวิวจากโฟลเดอร์ pos/neg, ทำ tokenization, แปลงคำเป็น index และ padding ความยาวเท่ากัน

ส่วนการสร้างโมเดล:

สร้างคลาส TextCNN(nn.Module) โดยกำหนด embedding, convolution layer, pooling และ fully connected layer

ส่วนการเทรน:

ใช้ optimizer AdamW, loss function CrossEntropyLoss และเทรนด้วย GradScaler บน GPU

มี Early Stopping และ ReduceLROnPlateau เพื่อกัน overfitting

ส่วนการประเมินผล:

ประเมินด้วย classification_report (accuracy, precision, recall, f1-score) และ Confusion Matrix

รายละเอียดการ Train และ Dataset

ใช้ชุดข้อมูล IMDB Movie Review Dataset (25,000 ตัวอย่างสำหรับ train + 25,000 สำหรับ test) ซึ่งแบ่งเป็นรีวิวเชิงบวกและเชิงลบอย่างละเท่าๆ กัน

Hyperparameters :

- Vocabulary size = 20,000
- Embedding dimension = 128
- Batch size = 128
- Epoch = 8 (Early stopping ใช้ patience = 2)
- Optimizer = AdamW (lr=0.002)
- Regularization = Dropout 0.5 + label smoothing 0.05

Evaluation And Conclusion

	precision	recall	f1-score	support
neg	0.8107	0.8249	0.8177	12500
pos	0.8218	0.8074	0.8145	12500
accuracy			0.8161	25000
macro avg	0.8162	0.8161	0.8161	25000
weighted avg	0.8162	0.8161	0.8161	25000

Confusion matrix:
```
[[10311  2189]
 [ 2408 10092]]
```

จากการทดลองใช้โมเดล TextCNN ในการวิเคราะห์อารมณ์ของข้อความรีวิวภาพยนตร์ พบว่าโมเดลสามารถเรียนรู้อารมณ์จาก text ได้จริง โมเดลมีค่าความถูกต้องเฉลี่ย 81.6% และให้ค่า Precision, Recall, และ F1-score ที่มีความสมดุลระหว่างรีวิวเชิงบวกและเชิงลบ แสดงให้เห็นว่าโมเดลมีความแม่นยำ สามารถนำไปใช้ต่อยอดได้จริง โดยไม่เกิดการ overfitting

ถึงแม้ว่าผลลัพธ์จะอยู่ในระดับที่น่าพอใจ แต่ยังสามารถพัฒนาให้ดียิ่งขึ้นได้ในอนาคต ไม่ว่าจะผ่าน embedding ที่ผ่านการเทรนมาแล้ว อย่าง GloVe หรือ FastText เพื่อให้โมเดลเข้าใจความหมายของคำได้ลึกขึ้น , การเพิ่มความหลากหลายของข้อมูลด้วย Text Augmentation

บทความอ้างอิงและงานที่เกี่ยวข้อง

1. Yoon Kim. “Convolutional Neural Networks for Sentence Classification.” *EMNLP 2014.*
2. PyTorch Documentation — https://pytorch.org/docs/stable/
3. IMDB Dataset — https://ai.stanford.edu/~amaas/data/sentiment/
4. GeeksforGeeks — https://www.geeksforgeeks.org/nlp/text-classification-using-cnn/