How to go to HELL?

with NMT Start-up Work Experience

Choi Minjoo (All for One)

Choi Minjoo
(Judy Choi)

- Careers
  - 2012 ~ 2019 **Malware Analyst**
  - 2018 **Working Holiday** in France
  - 2020 ~ **M.S** in Kangwon Univ
    - Intelligence Software Lab
      - NLP (Machine Translation)
  - 2021 ~ 2022 **Bering Lab (NLP Researcher & Engineer)**

- SNS
  - https://www.facebook.com/minjoo.choi.562/
  - https://github.com/Judy-Choi

Contents

1. Instruction

1. Instruction

Quiz

- Which is the ‘impossible’ AI service ‘now’?
  1) Translation service for specialist
  2) Electrical circuit design on chip (ex : TPU)
  3) Cancer cell detection
  4) Malware detection
  5) Perfect autonomous driving

1. Instruction

Quiz

1) Which is the ‘impossible’ AI service ‘now’?
    1) Translation service for specialist
    2) Electrical circuit design on chip (ex : TPU)
    3) Cancer cell detection
    4) Malware detection
    5) Perfect autonomous driving

        https://www.dispatch.co.kr/2196537

Set your Goal

Define clearly the problem

1. Instruction

How to go to HELL?

- Have only ‘Idea’
- Just try
- Judge only by your own thoughts
- If there are no problem, release quickly

1. Instruction

How to don't go to HELL?

- Have only 'Idea'
  - Realize your ideas
- Just try
  - Set the purpose of the experiment and consider the key points
- Judge only by your own thoughts
  - Evaluate the results clearly (by Evaluation score, Test set by human)
- If there are no problem, release quickly
  - Test enough (Quality Assurance)

2. Dataset

2. Dataset

How to go to HELL?

- I've already been to hell… 😂
- ParaCrawl Corpus is enough large 🤓
  - Over 20,000,000 pairs (EN-UZ)
  - So I just go ahead!
- But NMT model training acc score was dynamic & poor… 😔
  - Finally I got the BLEU score **9** 🤪

Big Data != Better Performance

2. Dataset

How to don't go to HELL?

- Filter noise data
  - Rule, LASER…

- Use clean dataset
  - AI Hub
    - Only 1,600,000 Pairs (KO-EN)
    - NMT model training acc score was progressive & good
    - Finally I got the BLEU score **35.97** 🥳

2. Dataset

Filtering

- Rule
  - Filter by rule
    - Length of sentence
    - Remove blank, special character..
  - Fast, Simple, Clear
  - Perl script

- Embedding & Similarity
  1. Filter by Deep Learning
     - Multilingual, Semantic
     - Open source
       - Universal Encoder (A little old method)
       - LASER (by Facebook AI)
  2. Calculate Similarity
     - Cosine Similarity
     - FAISS

2. Dataset

How to go to HELL?

- Big data file causes excessive file I/O
  - Too slow to I/O 🥺
  - Not enough memory 😇

- This delays our dinner time 😢

2. Dataset

How to don't go to HELL? : Handling Big Data

- Algorithm
  - Reduce time complexity
    - Remove nested loop
      - ex) double 'for' loop
  - Reduce space complexity
    - Reduce memory usage
      - ex) unnecessary copy
- Data Structure
  - Which is faster in Python?
    - list
    - dictionary
    - set
- Pandas & Numpy
  - DataFrame

2. Dataset

Coding Test

- How can we get intersection of 2 dataset?
  - A = [1, 2, 3, 4, 5]
  - B = [3, 4, 5, 6, 7]

2. Dataset

Coding Test

- Q) How can we get intersection of 2 dataset?
  - A = [1, 2, 3, 4, 5]
  - B = [3, 4, 5, 6, 7]
  - But it's too naive approach!

```
save = []

for a in A:
    for b in B:
        if a == b:
            save.append(a)
```

2. Dataset

Coding Test

- Q) How can we get intersection of 2 dataset?
  - A = [1, 2, 3, 4, 5]
  - B = [3, 4, 5, 6, 7]

```
set_A = set(A)
set_B = set(B)

save = set_A.intersection(B)
```

  - More simple & fast!

2. Dataset

Coding Test

- Q) How can we pick the column in parallel corpus?

Name	Age	Sex	Score
RM	20	Male	50
J-Hope	24	Male	40
Judy	33	Female	60
Jimin	31	Male	10

2. Dataset

Coding Test

- Q) How can we pick the column in parallel corpus?
  - Pandas (& Numpy)

```
import pandas as pd
df = pd.read_csv('Dataset.csv')

columns_df = df[['Name', 'Score']]
```

Name	Age	Sex	Score
RM	20	Male	50
J-Hope	24	Male	40
Judy	33	Female	60
Jimin	31	Male	10

3. Model

3. Model

How to go to HELL?

- Limit of deep learning
  let me go to HELL… 😇
  - Too much training time
  - Too much resource
  - Not easy to debug
    - Usually it makes me go to HELL 🐞

3. Model

How to don't go to HELL?

- Limit of deep learning
  let me go to HELL… 😇
  - Too much training time
  - Too much resource
  - Not easy to debug

- Reduce Batch size
  - But performance could be poor
- Reduce Model size
  - Improve model architecture
- Early-Stopping
- Schedule your experiment plan

3. Model

How to don't go to HELL?

- Limit of deep learning
  let me go to HELL… 😇
  - Too much training time
  - Too much resource
  - Not easy to debug

- Reduce Batch size
  - But performance could be poor
- Reduce Model size
  - Improve model architecture
- Model Parallelism
  - Models sharded across multiple GPUs

3. Model

How to don't go to HELL?

- Limit of deep learning
  let me go to HELL… 😇
  - Too much training time
  - Too much resource
  - Not easy to debug

- Visualization
  - TensorBoard
- Check Dataset
  - Is it clean?
- Optimization Tool
  - WandB

Try Quickly & Fail Quickly

4. Appendix (Tip)

4. Appendix

Development Tip : Linux command

- nohup (no hangups)
  - Command that runs process even after logging out
  - Can save output to log file
  - Helps automation

  - Usage : nohup script.sh > output.txt &

4. Appendix

Development Tip : Linux command

- File
  - head / tail
  - wc
  - find
  - cp
  - sh
- Directory
  - mv
  - rm (-rf)
- Server
  - scp

4. Appendix

Management Tip

- Version
  - Dataset
  - Model
- Storage
  - Server
  - AWS S3
  - NAS
- Collaborate Tool
  - Notion
  - Jira
  - Github
- Security
  - Protect against hacking ☠️

4. Appendix

Get support from

- School
  - Server (GPU)
  - Professor
  - Senior
- Government or Company
  - Server (GPU)
  - Education
  - Business support
  - Start-up investment
  - Working space

Happy Deep Learning!

Q&A