TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN – ĐHQG TP.HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN CÔNG NGHỆ PHẦN MỀM

TÀI LIỆU HƯỚNG DẪN
CÀI ĐẶT VÀ BIÊN DỊCH MÃ NGUỒN

XÂY DỰNG MÔ HÌNH DỊCH MÁY TỪ TIẾNG ANH SANG TIẾNG VIỆT

KHOÁ LUẬN TỐT NGHIỆP

GVHD:
TS. Ngô Huy Biên

Sinh viên thực hiện:
1612689 – Trương Phạm Nhật Tiến
1612689@student.hcmus.edu.vn
1612726 – Nguyễn Minh Trí
1612726@student.hcmus.edu.vn

Tp. Hồ Chí Minh, tháng 6 năm 2020

Mục lục

1. Giới thiệu 4

1.1 Mục đích 4

1.2 Yêu cầu môi trường 4

2. Cài đặt môi trường, công cụ 5

3. Biên dịch mã nguồn 6

Bảng mô tả thay đổi tài liệu

Ngày	Phiên bản	Mô tả	Người viết
16/06/2020	1.0	Hướng dẫn cách cài đặt môi trường và biên dịch mã nguồn của mô hình dịch máy	Trương Phạm Nhật Tiến Nguyễn Minh Trí
17/06/2020	1.1	Cập nhật thông tin môi trường cài đặt	Trương Phạm Nhật Tiến Nguyễn Minh Trí

1. Giới thiệu

1.1 Mục đích

Tài liệu hướng dẫn cũng như cung cấp những thông tin cần thiết, chi tiết về các bước để cài đặt môi trường và biên dịch mã nguồn của mô hình dịch máy từ tiếng Anh sang tiếng Việt.

1.2 Yêu cầu môi trường

MacOS, Ubuntu hoặc Linux và cũng có thể sử dụng dịch vụ Google Colab

RAM : 25GB

2. Cài đặt môi trường, công cụ

2.1 Chuẩn bị

- Máy tính sử dụng hệ điều hành MacOS, Ubuntu hoặc Ubuntu

2.2 Cài đặt môi trường

- Ở trong bài này nhóm phát triển sẽ hướng dẫn cài đặt môi trường Python và các thư viện hỗ trợ trên các máy tính (sử dụng Google Colab được cung cấp sẵn nên ta có thể chuyển sang phần biên dịch mã nguồn để tiếp tục).
- Bước 1: Ta cài đặt Python 3.5 và môi trường ảo. Ta chạy 2 câu lệnh sau trên command:
  - **"sudo apt-get update"**
  - **"sudo apt install python3-dev python3-pip virtualenv"**
- Bước 2: Tạo môi trường ảo env bằng câu lệnh:
  - **"virtualenv env"**
  - Ta được kết quả có một thư mục env được tạo ra.

```
tientruongphamnhat@model:~$ virtualenv env
Running virtualenv with interpreter /usr/bin/python2
New python executable in /home/tientruongphamnhat/env/bin/py
Also creating executable in /home/tientruongphamnhat/env/bin
Installing setuptools, pkg_resources, pip, wheel...done.
tientruongphamnhat@model:~$ ls
env
```

- Bước 3: Kích hoạt môi trường ảo env:

- **“source env/bin/activate”**
- Môi trường đã được kích hoạt có dấu hiệu như sau:

```
tientruongphamnhat@model:~$ source env/bin/activate
(env) tientruongphamnhat@model:~$
```

- Bước 4: Cài đặt các thư viện hỗ trợ:
  - **“sudo pip3 install --upgrade pip”**
  - **“sudo pip3 install --upgrade setuptools”**
  - **“sudo -H pip3 install tensorflow==1.15”**
  - **“sudo pip3 install sklearn nltk gensim matplotlib”**
- Vậy là chúng ta đã cài đặt môi trường để huấn luyện mô hình.

3. Biên dịch mã nguồn

- Tại đây, chúng ta sẽ tải mã nguồn mô hình từ github về máy:
  - **“git clone https://github.com/nmtri1912/Model.git”**
- **”cd Model”**

```
(env) tientruongphamnhat@model:~$ git clone https://github.com/nmtri1912/Model.git
Cloning into 'Model'...
remote: Enumerating objects: 52, done.
remote: Counting objects: 100% (52/52), done.
remote: Compressing objects: 100% (39/39), done.
remote: Total 52 (delta 21), reused 41 (delta 12), pack-reused 0
Unpacking objects: 100% (52/52), done.
Checking connectivity... done.
(env) tientruongphamnhat@model:~$ ls
Model  env
(env) tientruongphamnhat@model:~$ cd Model/
(env) tientruongphamnhat@model:~/Model$ ls
README.md  imgs  predict.py  train.py  train_load.py
```

- Cài đặt gdown, unzip và để và giải nén word_embedding:
  - **“pip3 install gdown”**
  - **“sudo apt-get install unzip”**
- Tiến hành tải word_embedding:
  - **“gdown https://drive.google.com/uc?id=1FbUCVL6UKDC-yyi72VUVdR4RYYaVAFQf”**
  - **“unzip data_webb.zip”**

```
(env) tientruongphamnhat@model:~/Model$ gdown https://drive.google.com/uc?id=1FbUCVL6UKDC-yyi72VUVdR4RYYaVAFQf
Downloading...
From: https://drive.google.com/uc?id=1FbUCVL6UKDC-yyi72VUVdR4RYYaVAFQf
To: /home/tientruongphamnhat/Model/data_webb.zip
3.43GB [00:10, 334MB/s]
(env) tientruongphamnhat@model:~/Model$ unzip data_webb.zip
Archive:  data_webb.zip
  inflating: data/dev-2012-en-vi/tst2012.vi
  inflating: data/dev-2012-en-vi/tst2012.en
  inflating: data/test-2013-en-vi/tst2013.vi
  inflating: data/test-2013-en-vi/tst2013.en
  inflating: data/train-en-vi/train.vi
  inflating: data/train-en-vi/train.en
  inflating: word_embedding/model_vn.bin
  inflating: word_embedding/model_en.bin
```

- Ta tiến hành tại 2 thư : **“vocab_english”** và **“vocab_vietnamese”**
  - **“mkdir vocab_english”**
  - **“mkdir vocab_vietnamese”**
  - Ta được thư mục Model với các thư mục con và file như sau:

```
(env) tientruongphamnhat@model:~/Model$ mkdir vocab_english c
(env) tientruongphamnhat@model:~/Model$ mkdir vocab_vietnamese
(env) tientruongphamnhat@model:~/Model$ ls
NMT.ckpt.data-00000-of-00001  README.md   data_webb.zip  train.py       vocab_vietnamese
NMT.ckpt.index                checkpoint  imgs           train_load.py  word_embedding
NMT.ckpt.meta                 data        predict.py     vocab_english
(env) tientruongphamnhat@model:~/Model$
```

- Tiến hành huấn luyện mô hình
  - **python3 train.py**

    --language_src data/train-en-vi/train.en

    --language_targ data/train-en-vi/train.vi

    --vocab_src vocab_english/

    --vocab_targ vocab_vietnamese/

    --word_emb_src word_embedding/model_en.bin

    --word_emb_targ word_embedding/model_vn.bin

    --num_layer 2

    --num_hiddens 512

    --learning_rate 0.001

    --keep_prob 0.85

    --beam_width 10

    --batch_size 128

    --checkpoint NMT.ckpt

- Mô hình tiến hành huấn luyện và:

```
[nltk_data] Downloading package punkt to
[nltk_data]     /home/tientruongphamnhat/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

```
step 1: loss = 9.653509140014648
step 2: loss = 9.51797866821289
step 3: loss = 8.091070175170898
step 4: loss = 7.290840148925781
step 5: loss = 7.1438517570495605
```

- Sau khi chạy xong mỗi epoch ta sẽ có kết quả các file checkpoint như sau:

- Như vậy ta đã tiến hành xong quá trình huấn luyện mô hình dịch máy từ tiếng Anh sang tiếng Việt. Để sử dụng mô hình ta xem thêm tài liệu **"Huong_Dan_Trien_Khai"**

4. Hướng dẫn sử dụng mô hình để dịch một file từ tiếng Anh sang tiếng Việt

- Đầu tiên ta cần tải Language Model được nhóm sinh viên phát triển để sử dụng trong lúc dịch:
  - "**gdown** https://drive.google.com/file/d/1-2wQJY5N1HWbFdDuznypMDfP42eWt0gA/view?usp=sharing
- Sau đó ta chạy đoạn lệnh:
  - **python3 predict.py**

    --language_src data/test-2013-en-vi/tst2013.en

    --language_targ data/ test-2013-en-vi/tst2013.vi

    --vocab_src vocab_english/

    --vocab_targ vocab_vietnamese/

    --word_emb_src word_embedding/model_en.bin

    --word_emb_targ word_embedding/model_vn.bin

    --num_layer 2

    --num_hiddens 512

--learning_rate 0.001

--keep_prob 0.85

--beam_width 10

--batch_size 128

--checkpoint NMT.ckpt

- Có thể ta có gặp trường hợp không đủ bộ nhớ. Khi đó ta tiến hành tách file **"tst2013.en"** thành các file nhỏ với kích thước không quá 500 dòng.