Madeline Younes
z5208494

Thesis Description

To develop a novel approach to Dialectal Identification (DID) through utilising transfer learning techniques for low resource dialects, particularly Arabic. There has been a significant amount of research into transfer learning for English based downstream tasks while the technology has not progressed to the same extent for low resource languages. In addition to this there has been limited research into using transfer based learning for LID and DID tasks. This thesis will investigate an accurate and reliable method to segment conversations of dialectal Arabic into homogenous dialectal segments and identify the dialect of each segment. The dialectal identifier should accurately distinguish the four major Arabic dialects and then further extended to be able to distinguish between seventeen regional dialects.

Meeting Minutes Logs

Week 2 Supervisor Meeting

Attendees: Dr Beena Ahmed, Madeline Younes
Date: 7/05/22
Agenda:

- Feedback from Thesis A
- Progress Check in
    - Script that relabels the training, dev, test using the umbrella terms, ensures that the data provided for each dialect is balanced based on a given time.
    - Started on going through hugging face tutorials to design a xlsr model + linear layer + softmax classifier

Action Items:

- Organise a meeting time with Renee Lu to learn how to use Katana and the Hugging face libraries.
- Ensure that I have access to Katana, if not request access.

Week 3 Supervisor Meeting

Attendees: Dr Beena Ahmed, Madeline Younes
Date: 14/05/22
Agenda:

- Progress Check in
    - Continued to work on an training implementation using the hugging face tutorials

- Wrote a custom filter and tested it on various audio files, not sure if it will cause distortion to the results. Side lining as a module that may or may not be integrated.
- Not sure how to manage and use the large dataset.

Action Items:

- Find code that can be used for bench marking
- Download dataset onto Katana (if too large to download the training dataset download just the dev and test portion of the dataset)
- Complete writing code to setup Wav2Vec
- Test using a custom made filter rather than the noise reduce function for the filtering code.
- Prep Slides for Monday meeting with Dubber summarising Thesis A and my progress so far.

Week 3 Wav2Vec Tutorial

Attendees: Renee Lu, Madeline Younes, Mahnee Przibilla, Rachel Gray
Date: 17/05/22
Agenda:

- Listen to the presentation which will cover:
  - Katana
  - Fine-tuning wav2vec 2.0 with hugging face
  - Questions
- Slides with resources:

  https://docs.google.com/presentation/d/1eUX6Hzslo2Hi_X7t5UseqfFIneVZiOG8xn6tMEgCDeY/edit?usp=sharing

- Questions
  - How did Renee handle batching?
    - Dataset was fairly small used the same method for batching as the hugging face tutorial
  - Did Renee add a downstream model in addition to the pretrained? If so how was this implemented?
    - She focused on finetuning

Action Items:

- Read through Renee's code particularly her training script and the format of her Katana Jobs.
- Start to integrate a similar structure into my own code
- Read the meta-paramater information and workout which parameters would also be significant for the downstream task of dialect identification.
- Understand what implantation methods used by Renee can be used in the implementation of an Arabic dialect identifier.

Week 4 Supervisor Meeting

Attendees: Dr Beena Ahmed, Madeline Younes
Date: 23/05/22

Progress Update:

- Completed Pre-processing scripts;
- Further testing on audio worked well in some scenarios but distorted others. Planning on training without them first.
- Currently debugging basic wav2vec implementation, memory allocation issues. Unsure if this due to how batching is performed.
- Decided to preform initial training will be done on the smaller dev dataset, as it can be downloaded onto Katana.

Action Items:

- Continue to work on debugging training code.

Week 5 Supervisor Meeting

Attendees: Dr Beena Ahmed, Madeline Younes
Date: 30/05/22

Agenda:

- Currently debugging training script encountering various errors.

Action Items:

- Find bench marking code

Week 7 Supervisor Meeting

Attendees: Dr Beena Ahmed, Madeline Younes
Date: 11/06/22

Progress Check-in:

Added additional modules to the model

- Added data collector
- Added a mean linear pooling layer (should be a similar process when I want to add a downstream model)

Still debugging, current bugs are:

- Issues with memory allocation
- Attempts to mitigate it:
    - Increased batch size
    - Truncated audio clips to 1s
    - Switched the audio reader library to audiotorch which is meant to have more efficient storage in encoding.
    - Switched str labels to *torch.long*
- Passing through batching returned the wrong input type
    - Haven't fixed (turned off batching for now)
- Also there's an error within training loop in where there's the wrong Convolutional tensor dimensions

```
My input nn.Conv2d: [batch_size, channels, height, width]
```

`Expected input nn.Conv1d: [batch_size, channels, seq_len]`

- Proposed Solution: attempt to flatten the layer

Questions:

- Is there anyone else other than Renee familiar with the hugging face training system/library that I can ask questions to?
- Any further information about trip to Dubber?

Action Items:

- Email Ian from Dubber on who to get support from within his team
- If still need support after talking to Dubber engineers contact phd student: *Antoni Dimitriadis antoni.dimitriadis@unsw.edu.au*
- Email Taryn about further details about Dubber trip.

Week 9 Supervisor Meeting

Attendees: Dr Beena Ahmed, Madeline Younes

Date: 28/06/22

Progress Check-in:

- What I learnt at Dubber
- Greater understanding of the base pretrained networks
- Particularly which layers need to be frozen and which need to be fine tuned
- Completely refactored code to use custom datasets & dataloader classes to facilitate batching
- Have been able to finetune wav2vec2-base on small amount of data
- What I have been doing this week
- Remove code redundancies
- Attempted to generate a classification report encountered bugs

Questions:

- What documentation should I submit next week?

Action Items:

- Add printout for total accuracy for each Epoch
- Remove code redundancies
- Submit job to katana with entire dev dataset for training and entire test dataset for testing
- Start tuning metaparamaters

- *Spend the rest of today trying to fix classification report (which will generate a printout of the confusion matrix)*
- Submit Meeting minute logs and executive summary which summarises the work completed over the term.