PENGGUNAAN MODEL TRANSFORMER PADA *AUDIOVISUAL SPEECH RECOGNITION* UNTUK BAHASA INDONESIA

TESIS

Karya tulis sebagai salah satu syarat
untuk memperoleh gelar Magister dari
Institut Teknologi Bandung

Oleh

GUGY LUCKY KHAMDANI
NIM: 23517041
(Program Studi Magister Informatika)

PROGRAM STUDI MAGISTER INFORMATIKA

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA

INSTITUT TEKNOLOGI BANDUNG

Mei 2019

JUDUL TESIS: PENGGUNAAN MODEL TRANSFORMER PADA *AUDIOVISUAL SPEECH RECOGNITION* UNTUK BAHASA INDONESIA

Oleh
Gugy Lucky Khamdani
NIM: 23517041
(Program Studi Magister Informatika)
Institut Teknologi Bandung

Menyetujui

Pembimbing

tanggal 21 Maret 2019.

Pembimbing I,	Pembimbing II
Dr. Dessi Puji Lestari, ST., M.Eng.	Nugraha Priya Utama, S.T., M.A., Ph.D.
NIP. 197912012012122005	NIP. 118110074

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

Bab IV Eksperimen dan Evaluasi

Bab ini menjelaskan tahapan implementasi pembangunan model, konfigurasi eksperimen, dan evaluasi kinerja model berdasarkan hasil eksperimen yang dilakukan.

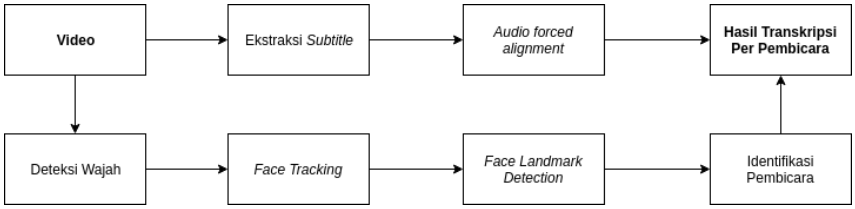
IV.1 Tujuan Eksperimen

Eksperimen dalam penelitian ini dilakukan dalam penentuan model bahasa dan parameter model, dan juga untuk mengukur kinerja model ASR, VSR, dan AVSR yang dibangun. Kinerja yang diukur adalah hasil pembangkitan kalimat yang ditranskripsikan dari suara dan gambar pada video. Kinerja tersebut diukur dengan menggunakan metrik *word error rate* (WER). Kualitas pembangkitan transkripsi dinilai baik jika memilki WER yang rendah.

IV.2 Pembangunan Model

Berdasarkan hasil eksperimen yang dilakukan, dibangun model ASR, VSR, dan AVSR yang sesuai untuk mengenali ucapan dalam bahasa Indonesia.

IV.2.1 Persiapan dan Pembentukan Transkripsi

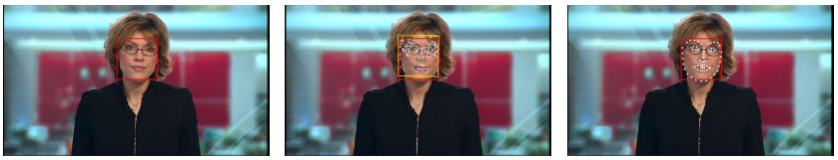


Gambar IV.1: Alur proses persiapan transkripsi.

Untuk melatih model diperlukan transkripsi dengan cap waktu yang selaras dengan suara yang diucapkan. Beberapa video di YouTube sudah ada yang menyediakan *subtitle* dan juga sudah selaras waktunya dengan kata-kata yang diucapkan di video. Akan tetapi subtitle tersebut hanya memiliki cap waktu dalam satuan kalimat, tidak kata per kata, sehingga perlu dilakukan penyelarasan menggunakan Penn Phonetics Lab Forced Aligner (dibangun berdasarkan kakas sumber terbuka HTK *Toolbox*). Teks yang akan diselaraskan merupakan teks dari *subtitle* yang kemudian dipetakan menjadi cara pelafalannya dengan kamus pelafalan yang sudah dibuat dengan menggunakan kakas Corpus Management Tools. Kakas penyelaras tersebut menggunakan algoritma Viterbi untuk menghitung *maximum likelihood alignment* antara

audio (yang dimodelkan menggunakan fitur PLP) dengan teks.

IV.2.2 Persiapan Korpus Video



Gambar IV.2: **Kiri:** Pendeteksian wajah (*bounding box* merah). **Tengah:** *tracking* wajah menggunakan fitur KLT (*bounding box* kuning). **Kanan:** Pendeteksian *landmark* wajah.

Corpus video yang telah dikumpulkan dideteksi wajah-wajah yang terdapat pada setiap framenya menggunakan metode berbasis HOG. Setelah wajah berhasil dideteksi, masing-masing wajah tersebut di*tracking* menggunakan KLT *tracker*, yang berguna juga dalam mengurangi *false positive* pada saat tahap pendeteksian wajah. Kemudian dari wajah yang sudah terdeteksi tersebut dideteksi *landmark*nya, yang kemudian *landmark* tersebut bisa digunakan untuk menentukan posisi mulut pada wajah. Untuk menentukan siapa yang sedang berbicara pada video, ditentukan dengan cara menghitung jarak ternormalisasi antara bibir atas dan bibir bawah dari setiap wajah yang terdeteksi, dan juga dihitung frekuensi buka tutup dari bibirnya.

IV.2.3 Ekstraksi Fitur

Terdapat dua ekstraksi fitur yang digunakan dalam penelitian ini, yaitu MFCC yang digunakan untuk mengektrasi fitur audio, dan CNN yang digunakan untuk mengektraksi fitur frame pada video. Konfigurasi parameter dari MFCC menggunakan 36-MFCC berdasarkan penelitian Yuwan (2018) dan untuk konfigurasi parameter CNN menggunakan konfigurasi pada penelitian Chung dkk. (2017).

IV.2.4 Eksperimen Pemodelan Sekuens

Pada eksperimen pemodelan sekuens, sekuens akan memodelkan dari tiga jenis masukan, yaitu masukan akustik saja, masukan visual saja, dan masukan gabungan akustik dan visual. Akan tetapi arsitektur model dari sekuensnya itu sendiri tetap sama, sehingga parameter-parameter yang bisa diuji pun jumlahnya akan tetap sama. Pemodelan sekuens yang digunakan adalah model transformer dan diimplementasikan dengan menggunakan TensorFlow dan PyTorch, yang disediakan oleh

Dai dkk. (2019) pada laman GitHub [1].

Sebelum pemodelan sekuens dilakukan, perlu dilakukan proses ekstraksi fitur seperti yang sudah dijelaskan sebelumnya. Setelah itu, model transformer dibentuk dengan menggunakan perintah

```
bash run_enwik8_base.sh train --work_dir PATH_TO_WORK_DIR
```

untuk melakukan proses pelatihan dan menggunakan perintah

```
bash run_enwik8_base.sh eval --work_dir PATH_TO_WORK_DIR
```

untuk melakukan proses evaluasi. `PATH_TO_WORK_DIR` adalah direktori tempat hasil pemodelan disimpan. Selain itu ada juga opsi-opsi tambahan yang diuji seperti

- `--batch_chuck` untuk menukar performa kecepatan pelatihan dengan memori yang digunakan. Untuk `batch_chuck>1`, program akan membagi setiap data latih menjadi `batch_chuck` bagian dan melakukan pelatihan pada setiap *batch* secara berurutan, dan gradien yang terkumpul akan dibagi dengan jumlah `batch_chuck`.
- `--div_val` untuk mengurangi dimensi *embedding*.
- `--fp16` dan `--dynamic-loss-scale` untuk menjalankan pelatihan dengan menggunakan mode pseudo-fp16 dan *dynamic loss scaling*. Untuk penggunaan `--fp16` perlu dilakukan pemasangan *package* `apex`[2] terlebih dahulu.
- `mem_len=0` untuk melakukan pelatihan tanpa menggunakan mekanisme rekurens.
- `attn_type=2` untuk melakukan pelatihan dengan model transformer standar tanpa menggunakan *positional encoding* relatif.

IV.3 Skenario Eksperimen

Proses pengujian dibagi menjadi tiga tahap. Pengujian pertama dilakukan terhadap pembangkitan transkripsi menggunakan model akustik. Pengujian kedua dilakukan terhadap pembangkitan transkripsi menggunakan model visual. Terakhir, pengujian

[1] `https://github.com/kimiyoung/transformer-xl`
[2] `https://github.com/NVIDIA/apex/`

ketiga dilakukan terhadap pembangkitan transkripsi menggunakan model akustik dan model visual.

IV.4 Hasil Eksperimen dan Evaluasi

Upabab ini membahas tentang hasil eksperimen pemodelan sekuens sesuai dengan implementasi pada upabab IV.2

IV.4.1 Parameter Model ASR

IV.4.2 Parameter Model VSR

IV.4.3 Parameter Model AVSR

Daftar Pustaka

Chung, J. S., A. Senior, O. Vinyals, dan A. Zisserman (2017). "Lip Reading Sentences in the Wild". *IEEE Conference on Computer Vision and Pattern Recognition*.

Dai, Z., Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, dan R. Salakhutdinov (2019). "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context". arXiv: 1901.02860.

Yuwan, R. (2018). "Pemodelan Akustik Berbasis Deep Neural Network Pada Sistem Pengenal Ucapan Spontan Bahasa Indonesia Memanfaatkan Active Learning". PhD thesis. Institut Teknologi Bandung, p. 51.