E-ISSN : 3031-9692 Vol. 1 No. 2 2024
Institut Teknologi dan Bisnis STIKOM Bali, 22-04-2024

Pengaruh Hyperparameter Tuning pada DeepSpeech2

Pande Putu Prana Pratistha[1)], Roy Rudolf Huizen[2)], Dadang Hermawan[3)]
Program Studi Magister Sistem Informasi
Institut Teknologi dan Bisnis STIKOM Bali
Denpasar, Indonesia
e-mail: *222012002@stikom-bali.ac.id[1)], roy@stikom-bali.ac.id[2)] , dadang@stikom-bali.ac.id[3)]*

Abstrak

Penelitian ini menginvestigasi efek dari Hyperparameter Tuning pada model Automatic Speech Recognition (ASR) yang dikenal sebagai DeepSpeech2. Fokus dari penelitian adalah pada optimasi parameter spesifik seperti tipe RNN yang digunakan (LSTM dan GRU) serta jumlah layer dalam arsitektur model. Tujuan utama adalah untuk mengidentifikasi konfigurasi yang optimal yang bisa mengurangi Word Error Rate (WER) sambil mengelola kompleksitas komputasi secara efektif. Analisis komprehensif menunjukkan bahwa konfigurasi menggunakan LSTM dengan 5-layer memberikan Word Error Rate terendah, yaitu 71.40%, yang mengindikasikan superioritasnya dibandingkan dengan GRU dalam hal akurasi. Konfigurasi dengan lebih banyak layer cenderung mengarah pada overfitting, yang diindikasikan oleh peningkatan Word Error Rate. Studi ini menggunakan dataset berbahasa inggris yang bersifat terbuka, yaitu LibriSpeech. Temuan dari penelitian ini membantu untuk penerapan dari ASR, menunjukkan bahwa tuning hyperparameter yang tepat untuk mencapai kinerja optimal tanpa menambah beban komputasi yang tidak perlu.

Kata kunci: *ASR, Deepspeech2 ,HyperParameter, Deep Learning, Word Error Rate.*

1. Pendahuluan

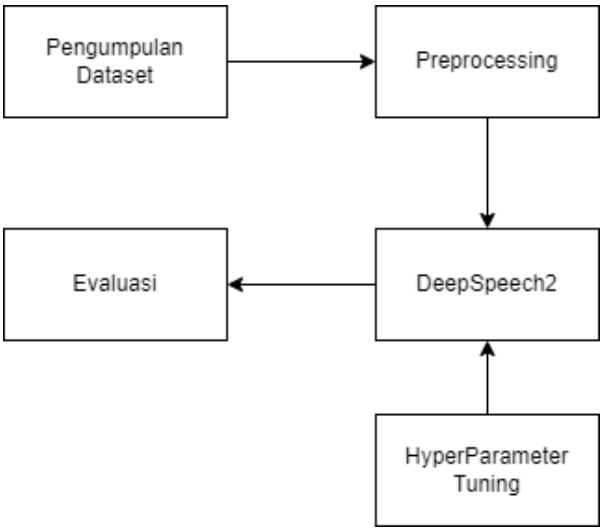
Saat ini, suara merupakan salah satu bentuk cara kita berinteraksi dengan komputer, yang dimana suara merupakan salah satu cara alami kita dalam berinteraksi[1], oleh karena kemudahannya ini suara populer digunakan sebagai alat picu perintah dalam Internet of Things, seperti *smart speaker*, *smart TV* dan sebagainya[2]. Selain sebagai alat untuk berinteraksi dengan komputer, suara juga bisa digunakan sebagai alat identifikasi diri biometrik[3], yang juga membantu dalam bidang kesehatan dalam mengoperasikan perangkat – perangkat tersebut, bahkan diimplementasikan sebagai alat untuk interaksi alat bantu untuk kaum difabel[4].

Namun demikian, *machine learning* pada suara merupakan hal yang kompleks, seperti yang ditulisan pada [5], kompleksitas ini muncul karena beberapa faktor yang saling terkait. Yang pertama, karena melibatkan penanganan data berdimensi tinggi dengan ketergantungan temporal yang signifikan, memerlukan model canggih untuk menangkap hubungan jangka panjang. Variabilitas dalam sinyal audio, yang dipengaruhi oleh kondisi perekaman yang beragam, perbedaan pembicara atau instrumen, dan kebisingan latar belakang, menambah lapisan kompleksitas lainnya. Transformasi seperti mengonversi sinyal menjadi *Mel Frequency Cepstral Coefficients* (MFCC) adalah proses penting namun kompleks yang harus mempertahankan informasi kritis. Kemudian, model *deep learning*, memerlukan dataset besar dan sumber daya komputasi yang substansial. Terakhir, penanganan non-linearitas yang melekat dan berbagai jenis kebisingan dalam sinyal audio memperumit strategi peningkatan sinyal yang efektif dan strategi pengurangan kebisingan.

Oleh karena itu, dalam rangka meningkatkan efektivitas proses machine learning yang kompleks pada suara, langkah yang bisa diambil adalah melakukan *HyperParameter Tuning*[6]. Proses *HyperParameter Tuning* adalah proses penyesuaian dan modifikasi dari parameter-parameter yang ada dalam algoritma *machine learning*[7]. Dengan mengatur dan mengoptimalkan parameter ini, diharapkan kita dapat menemukan kombinasi nilai parameter yang paling optimal., diharapkan mendapatkan nilai parameter yang paling optimal pada pemrosesan DeepSpeech2 ini.

2. Metode Penelitian

Untuk meneliti pengaruh *hyperparameter tuning* pada penelitian ini, adapun alur yang dilakukan seperti yang ditunjukan pada Gambar 1.



Gambar 1. Alur Penelitian

2.1 Spesifikasi Perangkat Keras dan Lunak

Untuk melakukan penelitian ini, adapun beberapa perangkat keras dan lunak yang digunakan pada Tabel *1*. Spesifikasi Perangkat

Tabel 1. Spesifikasi Perangkat

Tipe Perangkat	Spesifikasi
CPU	Ryzen 5 5600 (6 Core)
RAM	16 GB
GPU	NVIDIA 1060 6GB
Penyimpanan	SSD 250GB
Sistem Operasi	Ubuntu 22.04 LTS
Framework	Pytorch & Torchaudio

2.2 Pengumpulan Dataset

Pada penelitian ini, akan menggunakan dataset yang bersifat terbuka untuk publik, yaitu LibriSpeech (https://www.openslr.org/12/) yang menggunakan data yang berjumlah 100 jam berbahasa inggris, data ini disediakan oleh Vassil Panayotov dengan of Daniel Povey.

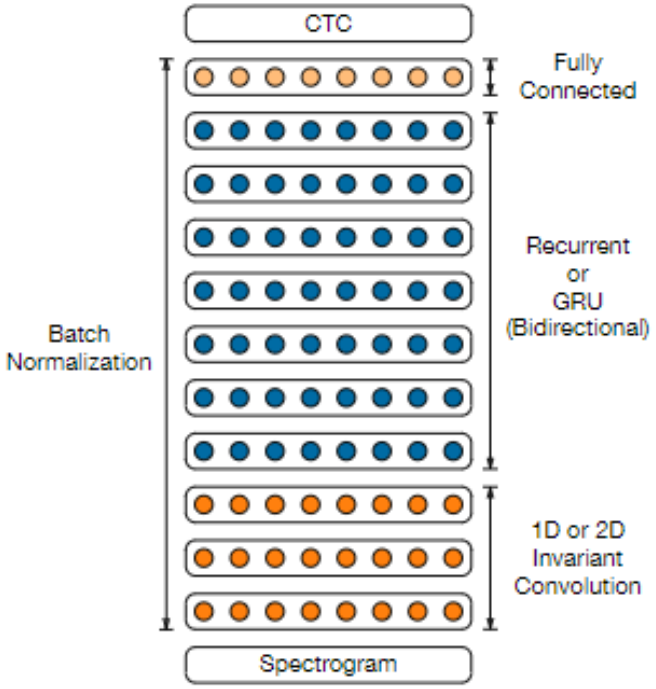
2.3 Preprocessing

Preprocessing di penelitian ini mencakup seperti, menkonversi file asli audio yang awalnya berasal dari librispeech adalah bertipe flac ke wav, kemudian juga konversi bitrate dari dataset tersebut ke 16 bit yang dimana, 16 bit adalah kualitas standar CD[8].

2.4 *DeepSpeech2*

DeepSpeech2 merupakan *Automatic Speech Recognition* (ASR) yang berguna untuk merjemahkan suara yang diucapkan ke sebuah teks. *DeepSpeech2* ini dikembangkan oleh Baidu Research dan Silicon Valley AI Lab merupakan pengembangan lebih lanjut daripada *DeepSpeech* yang sebelumnya dikembangkan oleh mozilla, yang dimana arsitekturnya menggunakan *Recurrent Neural Network* [9]. kemudian untuk pelatihannya, terdapat *Batch* Normalization[10], dan juga *Connectionist Temporal Classification* (CTC) untuk menyelaraskan ucapan masukan dengan keluaran teksnya tanpa memerlukan penyelarasan yang telah ditentukan sebelumnya antara kata-kata yang diucapkan dan yang ditulis. kemudian dengan *Mel Frequency Cepstral Coefficients* (MFCC) yang digunakan sebagai teknik ekstraksi

fitur yang bisasa digunakan dalam sistem untuk pengenalan suara dan identifikasi pembicara [11]. Arsitektur dari DeepSpeech2 di ilustrasikan seperti pada Gambar *2*. Arsitektur DeepSpeech2[9]



Gambar 2. Arsitektur DeepSpeech2[9]

2.4 *HyperParameter Tuning*

HyperParameter Tuning merupakan proses pada *machine learning* dimana mengoptimalkan kinerja model dengan sistematis menguji kombinasi parameter yang berbeda untuk menentukan mana yang menghasilkan hasil terbaik pada dataset tertentu. Adapun hal – hal yang di konfigurasikan pada langkah ini sesuai pada Tabel *2*. HyperParameter Tuning. Pada penelitian ini, terdapat 2 tipe RNN yang digunakan sebagai *HyperParameter* yaitu, *Long Short-Term Memory* (LSTM) yaitu jenis arsitektur RNN dimana biasa digunakan untuk data yang berurutan. LSTM di desain untuk mengatasi permasalahan *vanishing gradient problem* yang terdapat pada RNN tradisional. dan *Gated Reccurent Unit* (GRU) adalah versi streamline dari RNN, yang di desain menangkap *dependency* lebih efisien pada deret data tanpa kompleksitas yang lebih tinggi daripada model tradisional RNN [12].

Tabel 2. *HyperParameter Tuning*

Nama Parameter	Kondisi
Type RNN	LSTM, GRU
Layer RNN	5, 7, 9

Adapun beberapa parameter yang bersifat *fixed* pada penelitian ini yaitu, epoch dengan nilai 3, batch size dengan nilai 8, rnn hidden layer dengan nilai 1024.

2.5 *Evaluasi*

Kemudian pada langkah evaluasi, penelitian ini akan menggunakan WER (*Word Error Rate*) dimana pada [13] merupakan salah satu metode evaluasi pada Automatic Speech Recognition yang paling umum untuk dilakukan. Dengan WER ini membantu penelitian ini untuk membandingkan hasil dari *HyperParameter Tuning* yang dilakukan, adapun rumus dari WER adalah :

$$N = S + D + H \quad (1)$$

Dimana, S adalah jumlah penggantian. D adalah jumlah penghapusan, dan H adalah total jumlah ketepatan, yaitu kata-kata yang ditranskripsi dengan benar. Sehingga WER sendiri memilik rumus

$$WER = \frac{S+I+D}{N} \quad (2)$$

Dimana N adalah jumlah kata diucapkan dalam transkrip aslinya dan I adalah jumlah penyisipan.

3. Hasil dan Pembahasan

Tabel 3. Hasil Penelitian

HyperParameter	WER	Total Running Time
GRU, 5 Layer	75.50%	339 Menit
GRU, 7 Layer	98.10%	438 Menit
GRU, 9 Laye	113%	537 Menit
LSTM, 5 Layer	71.40%	403 Menit
LSTM, 7 Layer	108.0%	531 Menit
LSTM, 9 Layer	*Do Not Finish*	

Dalam mengevaluasi nilai Word Error Rate (WER), nilai yang lebih rendah merupakan lebih baik seperti yang di tunjukkan pada paper [14], yang dimana berdasarkan data pada Tabel *3*. Hasil Penelitian , bisa dilihat nilai terbaik dimiliki pada LSTM dengan 5 Layer RNN yang memiliki nilai WER 71.40%, dan jika diamati penambahan layer pada baik pada GRU maupun LSTM membuat terjadinya degradasi performa, yang kemungkinan terjadinya overfitting pada kasus ini, yang dimana GRU pada 5 layer memiliki nilai WER 75.50% dan ketika layernya 9, nilai WER meningkat menjadi 113% yang terdapat degradasi sebesar 37.50% begitu pula pada LSTM yang memiliki nilai 71.40 dengan 5 layer, degradasi menjadi 108% pada 7 layers, menunjukan penurunan sebesar 36.6%, selain itu meningkatnya layer meningkatkan waktu berjalannya, baik pada GRU, maupun pada LSTM.

Kemudian meskipun LSTM memiliki nilai WER lebih baik (71.40%) dibandingkan GRU (75.50%) yang sebesar 4.10%, GRU memiliki waktu proses yang lebih cepat sebesar 64 menit dibandingkan LSTM. Sehingga jika dibutuhkan kecepatan, maka GRU terlihat lebih unggul, dimana jika akurasi diperlukan, LSTM lebih unggul, ini sesuai dengan penelitian [12] yang mengatakan bahwa GRU kompleksitas lebih rendah daripada LSTM. Kemudian karena terjadinya *Out of Memory* dari video memory (VRAM) dari GPU ketika kondisi parameter LSTM 9 Layer, yang menunjukkan keterbatasan dari perangkat yang digunakan, yang kurang bisa mengakomodir kondisi *running* tersebut.

4. Kesimpulan

Dari hasil penelitian bisa dikatakan bahwa lebih tingginya layer RNN belum tentu membuat akurasi dari DeepSpeech2 menjadi lebih baik, malah menjadi overfitting dalam kondisi saat ini, yang dimana jumlah layer 5 lebih baik, dimana selain lebih rendah dalam waktu prosesnya, nilai WERnya juga lebih baik dibandingkan layer yang lebih banyak. Kemudian mengikatkan jumlah layer, juga meningkatkan penggunaan memory pada sistem. Pada tipe RNN terjadi tradeoff jika dibandingkan kedua tipe RNN tersebut (GRU dan LSTM) GRU memiliki waktu proses yang lebih rendah, namun LSTM memiliki akurasi yang lebih tinggi dibandingkan GRU, sehingga jika permasalahannya pada akurasi, direkomendasikan menggunakan LSTM, dan jika waktu adalah titik berat dalam sebuah penelitian, maka GRU patut digunakan.

Daftar Pustaka

[1] A. Tahseen Ali, H. S. Abdullah, and M. N. Fadhil, "Voice recognition system using machine learning techniques," *Mater Today Proc*, 2022, doi: 10.1016/j.matpr.2021.04.075.

[2] A. H. Ruslan, A. Z. Jusoh, A. L. Asnawi, M. D. R. Othman, and N. I. Abdul Razak, "Development of multilanguage voice control for smart home with IoT," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, May 2021. doi: 10.1088/1742-6596/1921/1/012069.

[3] A. Jamilu Ibrahim and U. Abubakar Jauro, "Bio-metric Encryption of Data Using Voice Recognition," *Automation, Control and Intelligent Systems*, vol. 9, no. 3, p. 89, 2021, doi: 10.11648/j.acis.20210903.12.

[4] M. A. K. Al Shabibi and S. M. Kesavan, "IoT Based Smart Wheelchair for Disabled People," in *2021 International Conference on System, Computation, Automation and Networking, ICSCAN 2021*, Institute of Electrical and Electronics Engineers Inc., Jul. 2021. doi: 10.1109/ICSCAN53069.2021.9526427.

[5] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Y. Chang, and T. Sainath, "Deep Learning for Audio Signal Processing," *IEEE Journal on Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, May 2019, doi: 10.1109/JSTSP.2019.2908700.

[6] L. Li *et al.*, "A SYSTEM FOR MASSIVELY PARALLEL HYPERPARAMETER TUNING," 2020.

[7] R. Bardenet, M. Brendel, B. Kégl, M. Sebag, and S. Fr, "Collaborative hyperparameter tuning," 2013.

[8] E. Brad Meyer, A. Member, and D. R. Moran, "Audibility of a CD-Standard A/D/A Loop Inserted into High-Resolution Audio Playback*." [Online]. Available: www.bostonaudiosociety.org/media.

[9] D. Amodei *et al.*, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.02595

[10] A.-M. Avram, V. Pais, and D. Tufiș, "TOWARDS A ROMANIAN END-TO-END AUTOMATIC SPEECH RECOGNITION BASED ON DEEPSPEECH2." [Online]. Available: http://www.racai.ro/p/reterom/

[11] S. Gupta, J. Jaafar, W. F. wan Ahmad, and A. Bansal, "Feature Extraction Using Mfcc," *Signal Image Process*, vol. 4, no. 4, pp. 101–108, Aug. 2013, doi: 10.5121/sipij.2013.4408.

[12] S. Ashraf Zargar, "Introduction to Sequence Learning Models: RNN, LSTM, GRU", doi: 10.13140/RG.2.2.36370.99522.

[13] M. A. Hassan, A. Rehmat, M. U. Ghani Khan, and M. H. Yousaf, "Improvement in Automatic Speech Recognition of South Asian Accent Using Transfer Learning of DeepSpeech2," *Math Probl Eng*, vol. 2022, 2022, doi: 10.1155/2022/6825555.

[14] Y. Deng, M. Mahajan, and A. Acero, "Estimating Speech Recognition Error Rate without Acoustic Test Data."