BAB 1

PENDAHULUAN

1.1. Latar Belakang Masalah

Deep learning adalah salah satu algoritma dalam pembelajaran mesin yang banyak di gunakan dan juga efektif. Dalam beberapa tahun terakhir, jaringan saraf tiruan yang mendalam (termasuk yang berulang) telah memenangkan banyak kontes dalam pengenalan pola dan pembelajaran mesin (Schmidhuber, 2015).

Deep Convolutional Neural Networks (CNN) telah menjadi metode yang paling menjanjikan untuk pengenalan objek, berulang kali menunjukkan hasil pemecahan rekor untuk klasifikasi gambar dan deteksi objek dalam beberapa tahun terakhir (Gong et al., 2014). Arsitektur CNN terbukti efektif untuk tugas-tugas penglihatan komputer (Krizhevsky et al., 2012). Jaringan konvolusional yang mendalam telah menghasilkan terobosan dalam pemrosesan gambar, video, ucapan dan audio, sedangkan jaring berulang telah menyoroti data sekuensial seperti teks dan ucapan (LeCun et al, 2015) .

Jaringan saraf tiruan yang intensif komputasi dan intensif memori sulit untuk digunakan pada *embeded system* dengan sumber daya perangkat keras yang terbatas atau perangkat mobile (Han et al, 2015). Kedalaman dari jaringan saraf tiruan merupakan komponen yang sangat penting untuk mendapatkan kinerja yang bagus (Simonyan & Zisserman, 2014).

Dengan meningkatnya jumlah kedalaman maka meningkat juga jumlah parameter yang ada di dalamnya, contohnya VGGNet mempunyai 138 juta

parameter (>500 MB). Maka dari itu model deep learning yang sangat memakan sumber daya *hardware* sulit di gunakan dalam perangkat *mobile* dan juga *embeded system* yang mempunyai sumber daya perangkat keras terbatas.

Dengan menghapus bobot yang tidak penting dari jaringan, beberapa perbaikan dapat diharapkan: generalisasi yang lebih baik, lebih sedikit contoh pelatihan yang diperlukan, dan peningkatan kecepatan belajar (LeCun et al, 1990). Kompresi model di perlukan untuk mengurangi ukuran dari *deep neural network* sekecil mungkin tanpa mengurangi akurasi dari model tersebut secara signifikan agar dapat di gunakan pada *embeded system* atau pun perangkat *mobile*. Contohnya pada aplikasi Android jika model *deep learning* dengan ukuran yang besar di implementasikan dalam pengembangan aplikasi maka ukuran memori aplikasi tersebut juga akan terlalu besar untuk di distribusikan di playstore atau di media lain, sedangkan jika model di letakkan pada server dan melakukan komunikasi dengan aplikasi melalui *webservice* maka aka ada jeda waktu untuk mendapatkan hasil dan membebani server untuk menangani banyak *request*. Karena itu maka di perlukan metode untuk mengkompres ukuran model *deep learning* agar bisa di implementasikan di perangkat mobile secara lokal dan tidak perlu internet untuk melakukan akses sehingga hasil dapat dilihat secara langsung.

1.2. Rumusan Masalah

Berdasarkan latar belakang di atas maka dapat di ambil rumusan masalah yaitu bagaimana melakukan kompresi pada model *deep neural network* sehingga memiliki ukuran yang lebih kecil tanpa mengurangi akurasi model secara signifikan.

1.3. Ruang Lingkup

Untuk memperjelas masalah yang akan dibahas dan agar tidak terjadi pembahasan yang meluas atau menyimpang, maka perlu kiranya dibuat suatu batasan masalah. Adapun ruang lingkum permasalahan yang akan dibahas adalah sebagai berikut:

a. Model *deep learning* yang digunakan adalah model AlexNet dari *torchvision models* sebagai objek kompresi.
b. *Pruning* model dan analisis tingkat *sparsity* tiap parameter dalam model serta pengaruh pruning terhadap ukuran memori dan akurasi model.
c. Metode *vector quantization* digunakan setelah *pruning* pada tiap parameter dalam model AlexNet.
d. Model setelah dilakukan *pruning* dan *v*ector *quantization* di kompres menggunakan zipfile dari python.
e. Data untuk percobaan berupa gambar anjing dan kucing berjumlah 2000 untuk training dan 800 untuk data validasi.

1.4. Tujuan Penelitian

Penelitian ini diharapkan dapat memberikan pengetahuan tentang bagaimana melakukan kompresi model deep learning dengan menggunakan metode *pruning* dan *vector quantization* sehingga menghasilkan model berukuran kecil dan tidak mengurangi akurasi secara signifikan.

1.5. Manfaat Penelitian

a. Bagi developer, penelitian ini dapat membantu dalam pengembangan model *convolutional neural network*, utamanya untuk kompresi model menggunakan metode *pruning* dan *vector quantization*.
b. Bagi peneliti lain, penelitian ini dapat dijadikan bahan referensi untuk penelitian kompresi model *deep learning*, utamanya yang berkaitan dengan metode *pruning* dan *vector quantization*.

1.6. Sistematika Penulisan

Berikut merupakan sistematika penulisan skripsi yang akan di buat :

BAB I. PENDAHULUAN

Pada bab ini berisi tentang penjelasan mengenai latar belakang masalah, rumusan masalah, ruang lingkup, tujuan penelitian, manfaat penelitian dan sistematika penulisan.

BAB II. TINJAUAN PUSTAKA DAN DASAR TEORI

Pada bab ini berisi tentang pembahasan sumber pustaka yang digunakan sebagai pedoman perancangan penelitian dan penjelasan yang berhubungan dengan penelitian yang digunakan sebagai landasan dalam penelitian.

BAB III. METODE PENELITIAN

Pada bab ini berisi tentang analisis kebutuhan, bahan/data, peralatan dan perancangan sistem yang akan digunakan.

BAB IV. IMPLEMENTASI DAN PEMBAHASAN

Pada bab ini menguraikan tentang pembuatan aplikasi yang merupakan implementasi dari hasil analisa dan perancangan, pengujian sistem dan kesimpulan.

BAB V. PENUTUP

Pada bab ini berisi kesimpulan yang dihasilkan dari pembahasan penerapan sistem dan saran-saran guna pengembangan sistem yang telah dibuat.