MAKALAH PENELITIAN LPB SUMUT 2023

Spaced Learning

Metode *Ensemble* untuk *neural network*

Delfino Jeconiah Djaja, Fhilip Yanus
Guru Pembimbing : Sungguh Ponten Aritonang
SMAS Bangun Insan Mandiri – Medan, Sumatera Utara;
delfinojaconiahdjaja@gmail.com; CS

A. Abstrak Penelitian

Neural network merupakan algoritma *deep learning* yang terinspirasi dari cara kerja otak manusia. Namun apa yang akan terjadi bila data yang dimiliki ribut dan kompleks disisi lain *random forest* dan *adaboost* merupakan teknik *ensemble* yang sangat populer karena reliabilitas yang tinggi. Pada makalah, ini kami mengembangkan sebuah algoritma gabungan *random forest dan adaboost* dan membandingkannya dengan model-model lain dengan *dataset MNIST, Fashion MNIST, dan CIFAR10*

B. Latar Belakang Penelitian

Deep learning merupakan subbidang dari *machine learning* yang difokuskan pada pengembangan dan penerapan *neural network* yang mendalam dan kompleks guna mengolah serta menganalisis data. Misi utamanya adalah memungkinkan komputer memahami hirarki representasi data dengan pendekatan yang serupa dengan cara otak manusia mengolah informasi.

Saat ini, kami tengah mengkaji model-model deep learning dan semakin mengenal beberapa keterbatasan yang dimilikinya. Salah satu kelemahan yang teridentifikasi adalah dalam menghadapi tantangan memahami pola dalam data yang kompleks. Ketika menghadapi dataset yang rumit, model-model deep learning tampaknya harus mengatasi lebih banyak fitur dan pola yang memerlukan pemahaman. Pertanyaan yang muncul adalah: apakah pendekatan dengan hanya menggunakan satu *neural network* sudah cukup untuk mengatasi kompleksitas ini?

Meskipun menggunakan satu *neural network* tunggal bisa menjadi pendekatan yang efektif, pada beberapa kasus, kompleksitas tinggi dalam dataset mungkin menuntut penerapan strategi yang lebih canggih. Dalam konteks ini,

pilihan yang lebih maju dan rinci perlu dipertimbangkan.

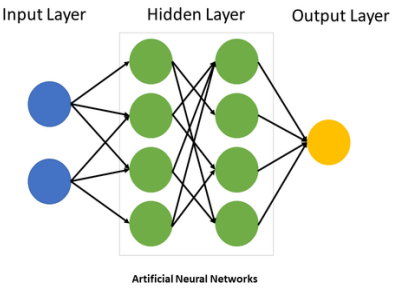
C. Perumusan Masalah

1. Bagaimana dampak kelemahan model-model *deep learning* terhadap kemampuan mereka dalam mempelajari pola pada *dataset* yang kompleks?
2. Apakah pendekatan dengan menggunakan satu *neural network* sudah cukup untuk menangani kompleksitas ini, atau apakah diperlukan pendekatan lebih lanjut yang lebih canggih?
3. Bisakah *neural network* fokus mempelajari kesalahan yang dibuat oleh *neural network* sebelumnya?

D. Studi Pustaka

a. Neural Network

Neural network merupakan model komputasi yang terinspirasi oleh cara kerja neuron dalam otak manusia. Neuron mengandung suatu *activation* yaitu suatu angka antara 0 dan 1.



gambar 1.1 *Neural network*

Struktur Neural Network terbentuk atas berikut:

i. **Input Layer** : *Input layer* merupakan lapisan yang terdapat kumpulan neuron yang *activation*-nya diinputkan sesuai data.
ii. **Hidden Layers** : *Hidden Layers* merupakan lapisan dimana neuron-neuron berinteraksi dan juga dimana *Neural Network*-nya dapat belajar.
iii. **Output Layer** : Output layer merupakan lapisan terakhir dalam model Neural Network mengeluarkan outputnya. Output tersebut dapat berbentuk *regression* dan *classification*.

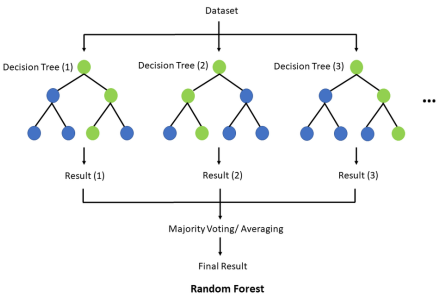
Variasi Neural Network :

1. CNN(Dalam penelitian ini kami menggunakan CNN arsitektur *Tiny VGG*)
2. RNN

3. Transformer
4. LSTM
5. GRU

b. Random Forest

Random forest merupakan penggabungan dari beberapa *decision tree* dan setiap *decision tree* dilatih dengan *bootstrapped* data yang berbeda-beda dan hasil *random forest* diambil dari voting atau rata-rata dari hasil yang dikeluarkan oleh masing-masing *decision tree* dengan tujuan untuk mengurangi *overfitting* dan meningkatkan variasi antar pohon mengidentifikasi fitur-fitur penting dalam dataset.



gambar 1.2 *Random Forest*

c. Adaboost

Adaboost, atau *Adaptive Boosting,* adalah sebuah algoritma *machine learning* yang digunakan untuk meningkatkan kinerja model prediksi dengan menggabungkan beberapa model lemah menjadi satu model kuat dengan memberikan bobot kepada kesalahan-kesalahan yang dibuat agar dapat lebih fokus dalam mempelajari kesalahan. *Weak learners* ini bisa berupa model sederhana yang memiliki performa yang sedikit di atas kemampuan tebakan acak.

d. Penggabungan

Dalam upaya mengoptimalisasi kemampuan model deep learning dalam mengidentifikasi pola dalam data, pendekatan yang diadopsi adalah mengintegrasikan penerapan metode *random forest* dan *adaboost* dengan jaringan *neural network* yang dikenal sebagai salah satu paradigma terkini dalam *machine learning.*

Metode *random forest* telah terbukti sebagai algoritma yang sangat efektif dalam pengolahan data. Pendekatan ini melibatkan penggunaan sejumlah pohon keputusan independen yang kemudian hasil prediksi mereka digabungkan untuk menghasilkan hasil akhir yang lebih konsisten dan akurat. Random Forest terbukti efisien dalam mengatasi data berdimensi tinggi dan kompleks. Untuk meningkatkan efektivitas dari *Random Forest*

kami menambah fungsi *splitting* dan pengacakan (*randomization*) data, yang mampu meningkatkan variasi dalam pelatihan model dan mengurangi potensi overfitting.

Di sisi lain, jaringan *neural network* mengadopsi konsep struktur dan fungsi jaringan saraf biologis dalam pembelajarannya. Kemampuan jaringan *neural network* dalam menangkap pola-pola kompleks dalam data didukung oleh rangkaian lapisan interkoneksi, yang disebut arsitektur deep learning. Namun, perlu diingat bahwa jaringan *neural network* kadang-kadang cenderung mengalami overfitting dan membutuhkan jumlah data yang besar untuk pelatihan yang optimal.

Kami juga menambahkan fungsi-fungsi dari *adaboost*. Seperti fungsi yang memungkinkan *neural network* untuk lebih fokus mempelajari kesalahan yang dibuat *neural network* sebelumnya dan fungsi pengambilan hasil akhir yang lebih bias pada *neural network* dengan akurasi yang lebih tinggi

Melalui integrasi metode Random Forest dengan jaringan *neural network*, tujuan yang diharapkan adalah menggabungkan potensi keduanya secara sinergis. Metode Random Forest mampu digunakan untuk mengidentifikasi fitur-fitur yang paling relevan dari dataset yang kompleks dan bervariasi. Di sisi lain, jaringan *neural network* dapat memperdalam pemahaman tentang pola-pola yang lebih kompleks dan abstrak dalam data, sehingga meningkatkan daya prediksi dan generalisasi model.

Tindakan *splitting* dan pengacakan data dalam implementasi Random Forest menambah dimensi penting dalam pendekatan ini, dengan memberikan variasi yang diperlukan untuk mengatasi kompleksitas data dan meminimalkan risiko overfitting.

Dengan demikian, integrasi yang matang antara metode Random Forest dan jaringan *neural network* dapat membuka potensi baru dalam kemampuan model dalam menghadapi tantangan analisis data yang semakin rumit dan dinamis.

Pendekatan ini memiliki potensi untuk menghasilkan model yang lebih tangguh dan adaptif terhadap data dengan komprehensif, serta mengurangi risiko overfitting yang mungkin terjadi pada model *neural network* yang murni. Dengan demikian, integrasi antara Random Forest dan jaringan *neural network* dapat memperluas kapabilitas model deep learning dalam menghadapi tantangan analisis data yang semakin kompleks dan beragam.

E. Metodologi Penelitian

a. Data preparation

Data yang kami memilih untuk diuji dengan model kami adalah data data yang bersifat klasifikasi. Data-data tersebut adalah :

i. Dataset MNIST

Dataset digit yang terdiri dari gambar tulis tangan berupa angka-angka. Atribut gambar dari kumpulan data direpresentasikan sebagai matriks 8x8 dari nilai skala abu-abu untuk setiap gambar.

ii. Fashion MNIST

Dataset pakaian yang terdiri dari gambar tipe-tipe pakaian seperti kaos, sepatu, dll.

iii. CIFAR-10

Dataset yang terdiri atas 10 *class* yang masing-masing mengandung 6000 gambar. Ada 5000 gambar pelatihan dan 1000 gambar pengujian per *class*. Contoh *class* pesawat, mobil, burung, dll.

b. Variabel dan perumusan

i. Variabel Independen

Splitting dan Pengacakan Data: Penggunaan teknik *splitting* dan pengacakan (randomization) pada data sebagai variabel kontrol dalam penerapan metode Random Forest.

Kombinasi Metode: Integritas dan kombinasi yang dioptimalkan antara implementasi metode Random Forest, *neural network* dan *adaboost.*

Pembobotan Kesalahan: Penggunaan teknik *error weighting* sebagai variabel kontrol dalam penerapan metode *adaboost*

ii. Variabel Dependen

Performa Identifikasi Pola: Akurasi, presisi, dan recall dalam mengenali pola-pola kompleks dalam data menjadi indikator utama kinerja model.

Kemampuan Generalisasi: Kemampuan model untuk menggeneralisasi dan memprediksi dengan baik pada data yang belum dilihat sebelumnya.

iii. Variabel Kontrol

Jumlah *Neural Network*: Jumlah *Neural network* yang membentuk algoritma *spaced learning* sebagai variabel kontrol untuk mengukur pengaruh variasi ini terhadap performa model.

jumlah data per *training*:jumlah data yang digunakan dalam 1 kali *training neural network*

Pembobotan kesalahan:seberapa banyak kesalahan di naikan.

iv. Variabel Moderasi

Kombinasi Optimal: Interaksi antara pengesplitan data, pengacakan, dan konfigurasi jumlah *neural network* untuk mengevaluasi kombinasi optimal yang menghasilkan peningkatan kinerja yang paling signifikan.

v. Variabel Eksternal

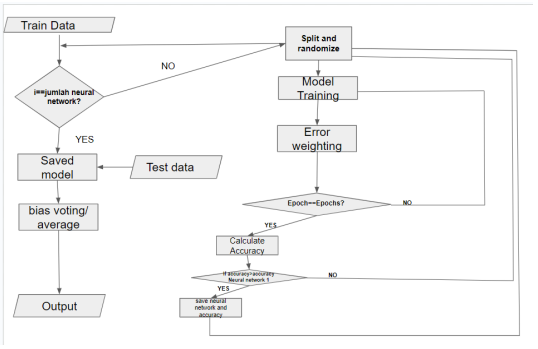
Jumlah Data Pelatihan: Jumlah data pelatihan yang digunakan dalam penelitian sebagai faktor eksternal yang berpotensi mempengaruhi hasil.

Dengan mengidentifikasi dan menganalisis variabel-variabel di atas, penelitian ini akan menggali dampak integrasi antara metode Random Forest dan jaringan *neural network* dalam meningkatkan kemampuan model dalam mengidentifikasi pola dalam data dengan lebih akurat dan efektif.

c. Cara perbandingan dengan model lain

Untuk membandingkan model kita dengan yang lain kami melatih model kami menggunakan *dataset MNIST, Fashion-MNIST, CIFAR 10* dan kemudian membandingkannya dengan performa dan akurasi dari model-model lain.

F. Hasil dan Pembahasan



gambar 2.1 *Flowchart spaced learning*

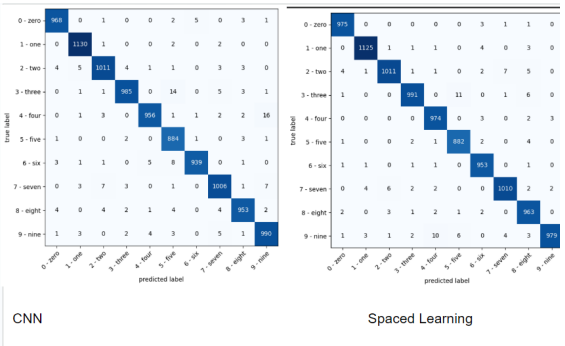
Gambar Diatas merupakan *flowchart* dan *pseudocode* dari algoritma hasil telitian kami

A. Perbandingan model

Model	*mnist*	*fashion mnist*	*cifar 10*
spaced learning	98%	90%	70%
CNN	98%	90%	56%
random forest	94%	87%	49%

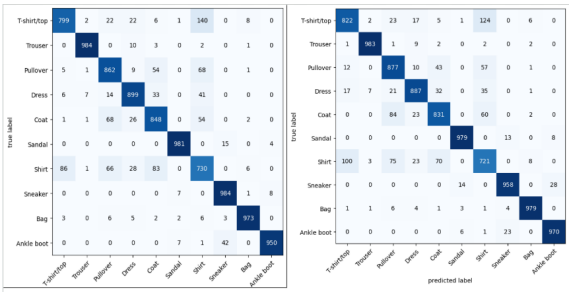
gambar 2.3 Tabel perbandingan model

Setelah melakukan *training* dengan model *CNN, spaced learning* dan *random forest* kami mendapatkan bahwa *cnn* dan *space learning* memiliki akurasi yang tertinggi dan sama(98%).



gambar 2.4 *confusion matrix* CNN dan *space learning* (dataset MNIST)

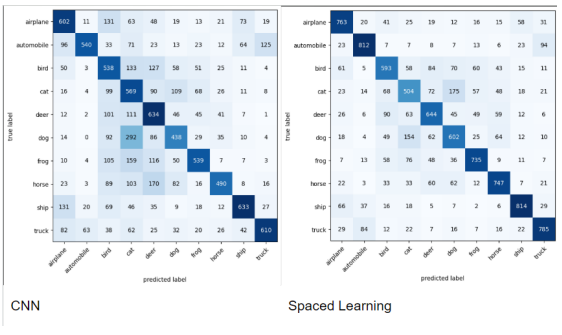
Untuk *dataset Fashion MNIST spaced learning* dan CNN memiliki akurasi yang sama(90%) *spaced learning* menggunakan lebih dari 1 *neural network.*



gambar 2.5 *confusion matrix* CNN dan *spaced learning*(*Fashion MNIST*)

Namun berdasarkan *confusion matrix* dari gambar 2.5 *spaced learning* lebih baik dalam membedakan kelas-kelas yang terlihat mirip seperti *t-shirt/top, pullover, coat, dress, coat dan shirt*.

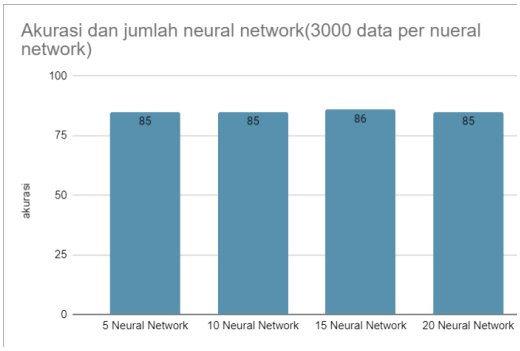
Untuk dataset *CIFAR-10 Spaced learning* berhasil meningkatkan kemampuan *CNN* dalam mempelajari pola pada data yang rumit



gambar 2.6 *confusion matrix CNN dan Space Learning (dataset CIFAR-10*)

B. Efek konfigurasi *space learning*

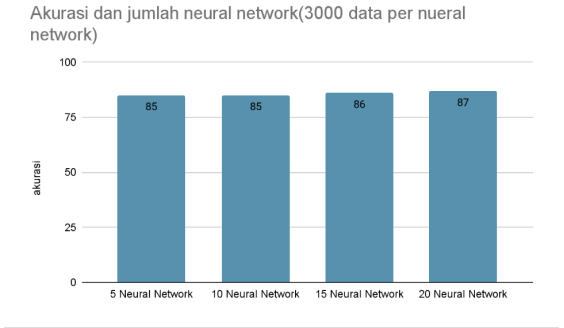
spaced learning memiliki beberapa *hyperparameter* seperti jumlah *neural network,* sebanyak apa kesalahan diberatkan, jumlah data dan jumlah *training loop.* Namun berdasarkan hasil percobaan kami menemukan bahwa jumlah *neural network* dan jumlah data sangat berpengaruh terhadap performa model.



gambar 3.1 akurasi dan jumlah *neural network* sebelum perubahan peraturan penyimpanan

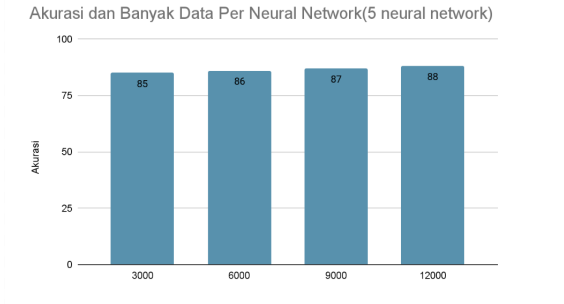
Hasil percobaan kami menunjukkan bahwa terlalu banyak n*eural network* merugikan akurasi. Setelah penelitian lebih lanjut, kami menyadari banyak neural network memiliki akurasi sama atau lebih rendah dari yang pertama. Penggunaan semua *neural network* ini menyebabkan ketidakkonsistenan dalam pengambilan keputusan.

Untuk mengatasi masalah ini, kami melakukan perubahan pada aturan penyimpanan *neural network* agar akurasinya lebih baik daripada yang pertama.



gambar 3.2 akurasi dan neural network setelah perubahan peraturan penyimpanan

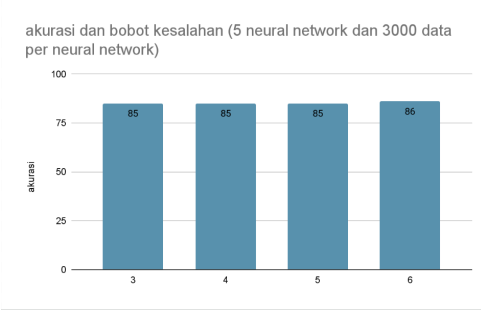
Perubahan peraturan penyimpanan berhasil memperbaiki *spaced learning*.



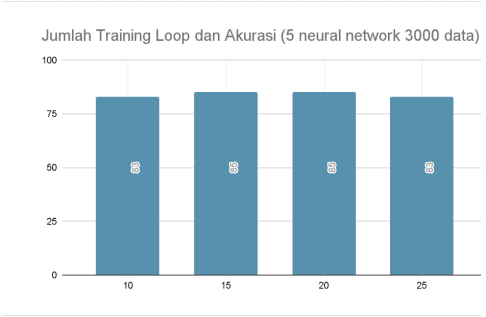
gambar 3.3 jumlah data dan akurasi

Semakin banyak data yang bisa dipelajari *spaced learning* semakin mahir dalam memprediksi. Hasil pengamatan kami menunjukan bahwa *hyperparameter* lainya tidak bisa

memberi dampak yang sangat besar sendirian



gambar 3.4 bobot kesalahan dan akurasi



gambar 3.5 *training loop* dan akurasi

G. Kesimpulan

Kesimpulan dari penelitian ini adalah *neural network* bisa menjadi pendekatan yang efektif untuk data-data yang sederhana.Namun ketika menghadapi data yang kompleks *Spaced learning* menjadi *teknik ensemble* yang efektif untuk digunakan oleh *neural network* yang dapat meningkatkan akurasi dan mengurangi kebingungan membedakan data yang terlihat sama.

H. Daftar Pustaka

Sanderson, Grant. 2017.” Neural Networks”,https://www.3blue1brown.com/topics/neural-networks ,diakses pada 29 Juli 2023 pukul 9.27.

TseKiChun. 2021. “Random Forest Explainer.png”,https://commons.wikimedia.org/wiki/File:Random_forest_explain.png,diakses pada 29 Juli 2023 pukul 13.15.

Breiman, Leo dan Adele Cutler,.”Random Forest”https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm,

diakses pada 29 Juli 2023 pukul 15.26.

Sharma, Neha, Vibhor Jain, Anju Mishra,”An Analysis Of Convolutional Neural Networks For Image Classification”,https://www.sciencedirect.com/science/article/pii/S1877050918309335,diakses pada 30 Juli 2023 pukul 9.06.

Schapire, Robert E,”Explaining Adaboost”,http://rob.schapire.net/papers/explaining-adaboost.pdf,diakses pada 3 agustus 2023 pukul 13.47.

“MNIST Database”,https://en.wikipedia.org/wiki/MNIST_database,diakses pada 10 agustus 2023 pukul 17.27.

“Fashion MNIST Dataset”https://datasets.activeloop.ai/docs/ml/datasets/fashion-mnist-dataset/,diakses pada 11 agustus 2023 pukul 15.27.

“CIFAR 10 Dataset|Paper WIth Code”https://paperswithcode.com/dataset/cifar-10,diakses pada 18 agustus 2023 pukul 15.27.