J-INTECH (Journal of Information and Technology)
Terakreditasi Kemendikbud SK No. 204/E/KPT/2022
E-ISSN: 2580-720X || P-ISSN: 2303-1425

Pemanfaatan Deep Convolutional Auto-encoder untuk Mitigasi Serangan Adversarial Attack pada Citra Digital

Putu Widiarsa Kurniawan S[1*], Yosi Kristian[2], Joan Santoso[3]

[1]Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terpadu Surabaya, Indonesia
[2]Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terpadu Surabaya, Indonesia
[3]Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terpadu Surabaya, Indonesia

Informasi Artikel

Diterima: 18-05-2023
Direvisi: 30-05-2023
Diterbitkan: 30-06-2023

Kata Kunci
Informasi; serangan adversarial; citra digital; Deep Convolutional Auto-encoder; mitigasi

***Email Korespondensi:**
putu.w20@mhs.istts.ac.id

Abstrak

Serangan adversarial pada citra digital merupakan ancaman serius bagi penggunaan teknologi machine learning dalam berbagai aplikasi kehidupan sehari-hari. Teknik Fast Gradient Sign Method (FGSM) telah terbukti efektif dalam melakukan serangan pada model machine learning, termasuk pada citra digital yang terdapat dalam dataset ImageNet. Penelitian ini bertujuan untuk mengatasi permasalahan tersebut dengan memanfaatkan teknik Deep Convolutional Auto-encoder (AE) sebagai metode mitigasi serangan adversarial pada citra digital. Hasil penelitian menunjukkan bahwa serangan FGSM dapat dilakukan pada sebagian besar citra digital, namun ada beberapa citra yang lebih tahan terhadap serangan. Selain itu, teknik mitigasi AE efektif dalam mengurangi dampak dari serangan adversarial pada sebagian besar citra digital. Akurasi model serangan dan mitigasi masing-masing sebesar 14.58% dan 91.67%.

Abstract

Adversarial attacks on digital images pose a serious threat to the utilization of machine learning technology in various real-life applications. The Fast Gradient Sign Method (FGSM) technique has proven to be effective in conducting attacks on machine learning models, including digital images found in the ImageNet dataset. This research aims to address this issue by utilizing the Deep Convolutional Auto-encoder (AE) technique as a method for mitigating adversarial attacks on digital images.The results of the study demonstrate that FGSM attacks can be performed on the majority of digital images, although there are certain images that are more resilient to such attacks. Furthermore, the AE mitigation technique proves to be effective in reducing the impact of adversarial attacks on most digital images. The accuracy of the attack and mitigation models is measured at 14.58% and 91.67%, respectively.

1. Pendahuluan

Pembelajaran mesin memainkan peran penting pada beberapa aspek. Mulai dari mempelajari sebaran penyakit atau wabah pada suatu wilayah (Dyas Irvan Masruri Sugeng Widodo, and Febry Eka Purwiantono. 2021) sampai dengan tugas klasifikasi citra. Namun terdapat beberapa ancaman dalam pemanfaatan, diantaranya Fast Gradient Sign Method (FGSM) adalah salah satu dari contoh teknik serangan *adversarial attack* yang digunakan untuk mengecoh model deep learning dengan menambahkan noise pada citra digital. Serangan ini dilakukan dengan mengambil gradien dari fungsi loss terhadap input, kemudian mengalikan gradien tersebut dengan epsilon, yang kemudian ditambahkan ke citra asli untuk menghasilkan citra yang dimanipulasi (Goodfellow, Shlens, and Szegedy 2014) . Sebagai hasilnya, citra yang dimanipulasi dapat mengecoh model deep learning dan menghasilkan prediksi yang salah. Serangan FGSM memiliki potensi bahaya yang besar pada citra digital, terutama pada aplikasi-aplikasi yang memerlukan keakuratan yang tinggi, seperti pengenalan wajah dan deteksi objek pada sistem keamanan dan keselamatan. Serangan ini dapat digunakan untuk mengelabui model deep learning dan menghasilkan prediksi yang salah, yang dapat menyebabkan kesalahan dalam pengambilan keputusan dan mengancam keamanan dan keselamatan pengguna. Tujuan penelitian ini adalah untuk menguji dan mengevaluasi efektivitas penggunaan deep convolutional auto-encoder sebagai mekanisme mitigasi serangan adversarial pada citra digital. Penelitian Penelitian ini memiliki beberapa manfaat yang dapat diidentifikasi, antara lain. Memberikan kontribusi pada pengembangan mekanisme pertahanan yang lebih kuat dan efektif terhadap serangan adversarial pada model jaringan saraf konvolusional. Dengan memahami efektivitas penggunaan deep convolutional auto-encoder sebagai mekanisme mitigasi, penelitian ini dapat membantu meningkatkan keamanan dan keandalan model deep learning dalam tugas klasifikasi citra. Pemahaman Lebih Lanjut tentang Serangan Adversarial: Melalui penelitian ini, akan tercipta pemahaman yang lebih baik tentang serangan adversarial, termasuk metode serangan yang umum digunakan seperti Fast Gradient Sign Method (FGSM).

2.1. Studi Literatur

Beberapa penelitian telah dilakukan untuk mempelajari potensi bahaya serangan FGSM pada citra digital. Misalnya, pada sebuah studi yang dilakukan oleh Carlini dan Wagner (Carlini and Wagner 2016) pada tahun 2016, mereka menunjukkan bahwa serangan FGSM dapat berhasil digunakan untuk mengecoh model deep learning dalam beberapa kasus, bahkan ketika model tersebut dilindungi oleh teknik pertahanan lainnya. Studi ini menunjukkan bahwa serangan FGSM memiliki potensi bahaya yang besar pada citra digital, dan perlu adanya upaya-upaya untuk melindungi model deep learning dari serangan ini. Pada sebuah studi yang dilakukan oleh (Chakraborty et al. 2018), mereka mengajukan sebuah metode pertahanan dengan menggunakan Auto-encoder untuk melindungi model deep learning dari serangan FGSM pada citra digital. Dalam studi tersebut, Auto-encoder digunakan untuk merekonstruksi citra asli dan menghilangkan noise yang ditambahkan pada citra yang dimanipulasi oleh serangan FGSM. Metode ini berhasil menunjukkan tingkat keberhasilan yang baik dalam melindungi model deep learning dari serangan FGSM pada beberapa dataset citra digital. Setiap serangan bergantung pada sedikit perubahan pada sampel masukan dengan mengoptimalkan parameter jaringan yang berbeda. Penelitian sebelumnya telah menguji efektivitas penggunaan kekurangan sebagai bentuk pertahanan terhadap gangguan serangan semacam itu. Contohnya, penelitian oleh (Marzi et al. 2018) menunjukkan bahwa membatasi kekurangan dengan memilih K elemen terbesar dalam hal magnitudo dari total N elemen terbesar pada sampel klasifikasi dalam domain wavelet berhasil mengurangi kesalahan klasifikasi yang disebabkan oleh serangan berbahaya pada Support Vector Machine (SVM). Selain itu, penelitian oleh (Bhagoji et al. 2017) menunjukkan bahwa memproyeksikan data berdimensi tinggi ke dalam subruang berdimensi lebih rendah menggunakan Principle Component Analysis (PCA) efektif dalam mengurangi keberhasilan serangan. (Gu and Rigazio 2014) menunjukkan bahwa DAE (Denoising Auto-encoder) dapat efektif digunakan untuk menghapuskan noise Gaussian yang dimasukkan ke dalam sampel masukan, tetapi penelitian tersebut tidak mengeksplorasi penggunaannya dalam skenario serangan yang berbahaya. Meskipun demikian, penelitian ini menunjukkan bahwa arsitektur auto-encoder denoising berlapis banyak dapat efektif digunakan untuk menghilangkan noise yang dimasukkan.

2. Metode Penelitian

Alur sistem mitigasi serangan adversarial attack pada citra digital menggunakan metode auto-encoder dimulai dari pengumpulan dataset citra digital yang terdiri dari citra asli yang tidak terpengaruh serangan, serta citra-citra yang telah. Preprocessing pada dataset citra, seperti normalisasi dan preprosesing lainnya yang diperlukan untuk mempersiapkan data sebagai input pada auto-encoder. Melakukan traning auto-encoder menggunakan citra asli yang tidak terpengaruh adversarial attack. Auto-encoder bertujuan untuk mempelajari representasi yang kompak dari citra asli. Menerapkan adversarial attack pada citra-citra dari dataset serangan. Pada penelitian ini adversarial attack yang digunakan adalah FGSM (Fast Gradient Sign Method). Memanfaatkan model auto-encoder yang telah dilatih untuk mendeteksi serangan adversarial pada citra. Ketika citra masukan yang mungkin terpengaruh serangan diberikan ke auto-encoder, dilakukan pengecekan kesalahan rekonstruksi antara citra masukan dan citra yang direkonstruksi oleh auto-encoder (Sahay, Mahfuz, and Gamal 2018).

2.2. Analisa Masalah

2.2.1 Adversarial Training

Pada adversarial training, model dilatih untuk menghadapi serangan atau gangguan yang disebabkan oleh data yang dimodifikasi dengan sengaja oleh musuh atau pihak berlawanan. Misalnya, dalam pengenalan citra, model dapat dilatih dengan menyertakan contoh citra yang dimodifikasi sedemikian rupa sehingga model tersebut menjadi lebih tangguh terhadap serangan, seperti penambahan noise pada citra atau pengubahan sebagian informasi pada citra. Dalam konteks adversarial training, discriminator dan generator merupakan dua komponen yang biasanya digunakan dalam arsitektur Generative Adversarial Networks (GANs). Discriminator adalah komponen dalam GANs yang bertugas membedakan antara data asli dan data yang dihasilkan oleh generator (Goodfellow, Pouget-Abadie, et al. 2014). Tujuan utama dari discriminator adalah untuk mempelajari fungsi yang dapat membedakan antara citra asli dan citra palsu. Dalam pelatihan, discriminator diberikan data yang berisi campuran antara citra asli dan citra yang dihasilkan oleh generator. Diskriminator secara bertahap diperbarui agar dapat membedakan dengan lebih baik antara keduanya. Generator adalah komponen lain dalam GANs yang bertugas membuat data baru yang menyerupai data asli. Generator berusaha untuk menghasilkan data yang dapat menipu atau mengelabui discriminator sehingga dapat meyakinkan discriminator bahwa data yang dihasilkan adalah data asli. Dalam pelatihan, generator menerima input z (misalnya, vektor angka acak) dan menghasilkan citra palsu berdasarkan input tersebut. Tujuannya adalah untuk mempelajari distribusi data asli sehingga generator dapat menghasilkan data yang semirip mungkin dengan data asli.

2.2.2 Adversarial Attack

Adversarial attack (serangan adversarial) adalah upaya untuk menciptakan atau menemukan data input yang dimodifikasi dengan sengaja, yang disebut adversarial examples, dengan tujuan mengelabui atau menipu model yang dilatih. Adversarial attack biasanya ditargetkan pada model pembelajaran mesin, terutama jaringan saraf (neural networks), dan dapat memanipulasi hasil prediksi model. Serangan adversarial dapat dilakukan dengan berbagai metode, termasuk: Fast Gradient Sign Method (FGSM): Serangan ini melibatkan menghitung gradien model terhadap input dan kemudian menambahkan noise ke input dengan arah gradien yang berlawanan untuk menciptakan adversarial examples (Goodfellow, Shlens, et al. 2014). Iterative FGSM (I-FGSM): Metode ini adalah variasi dari FGSM di mana serangan dilakukan secara iteratif dengan memperbarui adversarial examples beberapa kali untuk meningkatkan keberhasilan serangan. Carlini and Wagner Attack: Serangan ini menggunakan pendekatan optimisasi untuk menemukan adversarial examples dengan meminimalkan jarak antara input yang dimodifikasi dan input asli, sambil mempertahankan ketidakcocokan loss function.

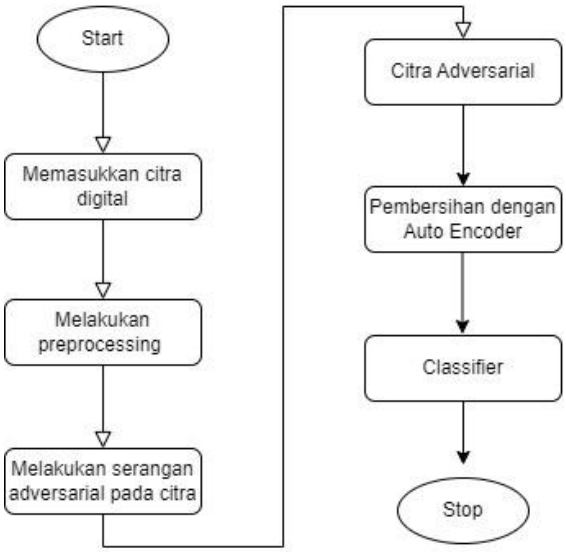
2.2.3. Auto-Encoder

Auto-encoder (AE) adalah tipe jaringan saraf tiruan yang digunakan untuk unsupervised learning. Auto-encoder terdiri dari encoder dan decoder, dengan tujuan merekonstruksi data inputnya. Encoder

mengompres data input menjadi representasi dimensi rendah yang disebut ruang laten atau kode, sedangkan decoder mencoba merekonstruksi input asli dari representasi yang telah dienkoding. Ide utama dibalik auto-encoder, model learning membaca representasi terkompresi dari data input, yang menangkap fitur-fitur paling penting. Proses ini dapat berguna untuk tugas-tugas seperti reduksi dimensi, pembelajaran fitur, pemulihan dari data yang terdistorsi, dan deteksi anomali. Auto-encoder dilatih dengan meminimalkan kesalahan rekonstruksi antara input asli dan output yang dihasilkan oleh decoder (Bank, Koenigstein, and Giryes 2020).

2.3. Arsitektur Sistem

Seluruh aliran proses yang digunakan dalam arsitektur yang diusulkan ditunjukkan pada gambar 1. Langkah awal melakukan preprocessing, klasifikasi citra digital, membangun kumpulan data citra adversarial dengan label (supervised learning), kemudian melakukan auto encode yang memproses citra dengan menghilangkan noise atau perturbasi dan akhirnya luaran dari model adalah citra terklasifikasi dengan area penting yang disorot di dalamnya. Sebagai langkah awal dalam arsitektur, citra diambil sebagai masukan, dilakukan proses klasifikasi dan pembuatan dataset citra adversarial yang fully supervised. klasifikasi dilakukan untuk memastikan keseluruhan citra. Dataset yang terdiri dari dua citra yang diklasifikasikan berbeda (asli dan adversarial) dijadikan sebagai masukan ke auto-encoder. Auto-encoder menghilangkan noise dari citra adversarial input dan menghasilkan citra denoise sebagai output. Kemudian memproses citra yang diperoleh dari model klasifikasi citra untuk selanjutnya melakukan print hasil.



Gambar 1. Alur Sistem Mitigasi

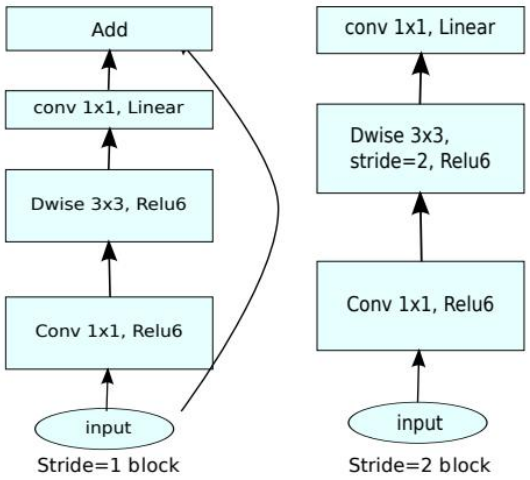
2.4. Basis Model

MobileNetV2 menggunakan blok residual terbalik yang disebut Inverted Residuals. Blok ini terdiri dari beberapa layer konvolusi yang diatur dalam urutan: konvolusi 1x1, non-linearitas (ReLU), konvolusi Depthwise 3x3, non-linearitas, dan konvolusi 1x1. Struktur ini memungkinkan model untuk memiliki kapasitas yang lebih besar dengan jumlah parameter yang lebih sedikit.

Linear Bottlenecks: Model ini juga menggunakan linear bottlenecks di antara layer-layer konvolusi. Bottlenecks terdiri dari konvolusi 1x1 yang diikuti oleh non-linearitas dan diikuti lagi oleh konvolusi 1x1. MobileNetV2 menggunakan konvolusi depthwise separable, yaitu proses konvolusi yang terpisah antara konvolusi spatialis (depthwise) dan konvolusi linier (pointwise) (Sandler et al. 2018).

Selain itu, MobileNetV2 juga memanfaatkan skala lebar (width multiplier) yang dapat disesuaikan. Dengan mengurangi skala lebar, model dapat memiliki jumlah saluran yang lebih sedikit pada setiap lapisan, yang secara signifikan mengurangi jumlah parameter yang perlu dihitung. Dengan begitu, MobileNetV2 mampu mengontrol ukuran dan kompleksitas model sesuai dengan kebutuhan, sehingga lebih hemat sumber daya.

Selanjutnya, desain efisien MobileNetV2 juga merupakan faktor penting dalam keberhasilannya sebagai model yang ringan (Howard et al. 2017). Arsitektur ini dirancang secara khusus untuk meminimalkan latensi dan penggunaan sumber daya. Teknik seperti shortcut connections juga digunakan untuk meningkatkan aliran informasi dalam model dan mengurangi beban komputasi. Dalam kombinasi dengan pendekatan depthwise separable, desain efisien MobileNetV2 memberikan keseimbangan antara kinerja yang baik dan kebutuhan sumber daya yang rendah, menjadikannya pilihan yang cocok untuk perangkat dengan sumber daya terbatas seperti perangkat mobile dan embedded.



Gambar 2. Arsitektur Mobilenetv2

Bagian utama dari MobileNet adalah blok depthwise separable convolution, yang terdiri dari konvolusi depthwise yang dilakukan secara terpisah untuk setiap saluran input, diikuti oleh konvolusi pointwise untuk menggabungkan saluran-saluran tersebut. Ini memungkinkan MobileNet untuk mencapai kinerja yang baik dengan jumlah parameter yang lebih kecil dibandingkan dengan arsitektur jaringan saraf konvolusi lainnya. Tabel 1 adalah contoh layer dari sebuah model mobilenetv2 beserta parameter.

Tabel 1. Layer dan Parameter Mobilenetv2

No. Layer	Tipe Layer	Input Size	Output Size	Output Depth
1	Convolutional (3x3, 32)	H x W x 3	H/2 x W/2 x 32	32
2	Bottleneck (1x1, 16)	H/2 x W/2 x 32	H/2 x W/2 x 16	16
3	Inverted Residual (3x3, 24, 1)	H/2 x W/2 x 16	H/4 x W/4 x 24	24
4	Inverted Residual (3x3, 24, 1)	H/4 x W/4 x 24	H/8 x W/8 x 24	24
5	Inverted Residual (3x3, 32, 1)	H/8 x W/8 x 24	H/8 x W/8 x 32	32
6	Inverted Residual (3x3, 32, 1)	H/8 x W/8 x 32	H/16 x W/16 x 32	32
7	Inverted Residual (3x3, 32, 1)	H/16 x W/16 x 32	H/16 x W/16 x 32	32
8	Inverted Residual (3x3, 64, 2)	H/16 x W/16 x 32	H/32 x W/32 x 64	64
9	Inverted Residual (3x3, 64, 1)	H/32 x W/32 x 64	H/32 x W/32 x 64	64
10	Inverted Residual (3x3, 64, 1)	H/32 x W/32 x 64	H/32 x W/32 x 64	64
11	Inverted Residual (3x3, 64, 1)	H/32 x W/32 x 64	H/32 x W/32 x 64	64
12	Inverted Residual (3x3, 96, 1)	H/32 x W/32 x 64	H/32 x W/32 x 96	96
13	Inverted Residual (3x3, 96, 1)	H/32 x W/32 x 96	H/32 x W/32 x 96	96

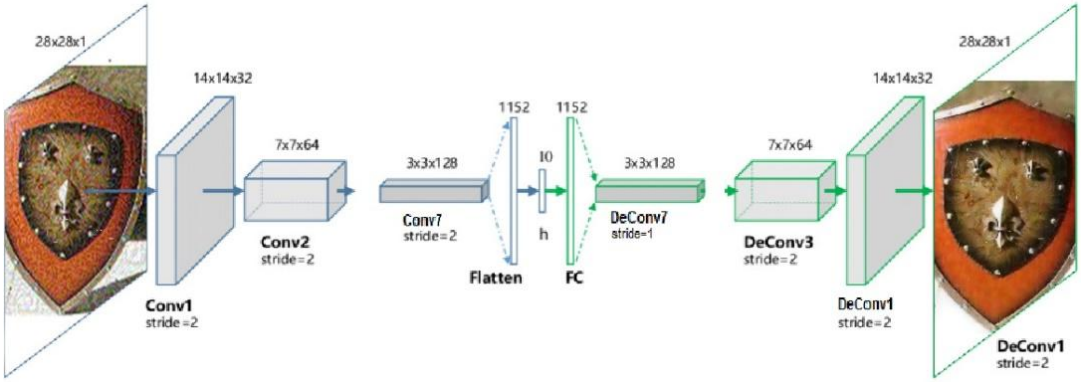
14	Inverted Residual (3x3, 96, 1)	H/32 x W/32 x 96	H/32 x W/32 x 96	96
15	Inverted Residual (3x3, 160, 2)	H/32 x W/		

2.5. Dataset Imagenet

Penelitian ini akan menggunakan bobot pretrained dari dataset ImageNet. Dengan menggunakan bobot pretrained dari dataset ImageNet (Chrabaszcz, Loshchilov, and Hutter 2017) , penelitian ini memanfaatkan pengetahuan dan representasi fitur yang telah dipelajari oleh jaringan saraf dalam tugas klasifikasi gambar dalam skala besar. Bobot pretrained ini memberikan dasar yang kuat untuk parameter awal model, memungkinkannya menggunakan fitur-fitur yang telah dipelajari dan potensial meningkatkan performa pada tugas yang dituju. Memanfaatkan bobot pretrained dari ImageNet dapat menghemat sumber daya komputasi dan waktu yang signifikan karena model tidak perlu dilatih dari awal (Russakovsky et al. 2014) . Sebagai gantinya, model dapat diperbaiki atau dilatih lebih lanjut pada dataset atau tugas yang spesifik, yang mungkin memiliki data latih terbatas atau fokus pada domain yang berbeda. Dengan memanfaatkan pendekatan transfer learning dengan bobot pretrained ImageNet, penelitian ini bertujuan untuk mendapatkan manfaat dari kemampuan generalisasi model pretrained dan meningkatkan performanya pada tugas atau dataset yang sedang diteliti. Pendekatan ini memungkinkan model belajar representasi yang bermakna dari dataset yang dituju dengan lebih efektif dan potensial mencapai akurasi dan ketahanan yang lebih tinggi dalam klasifikasi atau analisis citra digital.

2.6. Pelatihan Model Auto-Encoder

Bangun arsitektur terdiri dari 10 layer encoding convolution dan 4-layer decoding convolution pada gambar 3, menggunakan loss function MSE, algoritma optimisasi yang digunakan adalah Adam optimizer. Pelatihan dilakukan secara iteratif dengan mengulangi langkah optimisasi pada data latih hingga mencapai kondisi konvergensi. Luaran berupa citra digital yang bersih dari *noise* diharapkan bisa menjadi citra digital masukan yang lebih baik bagi *classifier*.

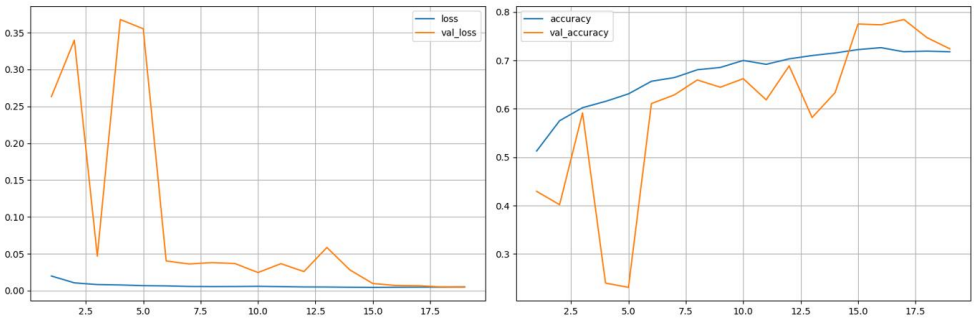


Gambar 3. Arsitektur Auto-encoder yang Ditraining

Menghitung gradien dari fungsi kerugian terhadap parameter w, yang diberikan oleh $\frac{\partial \mathrm{L}}{\partial \mathrm{w}}$. Gradien ini menunjukkan arah dan besar perubahan yang harus dilakukan pada setiap parameter untuk meminimalkan *loss function*. Kemudian menghitung momentum eksponensial bergerak rata-rata dari gradien *m* dan momentum eksponensial bergerak rata-rata dari gradien kuadrat *v*, dengan menggunakan (1):

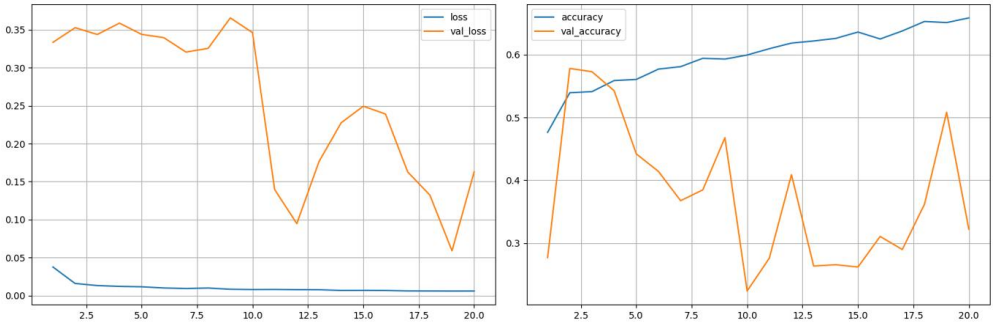
$$w \leftarrow w - \alpha \frac{m}{\sqrt{\bar{v}}+\epsilon}\ m \leftarrow \beta 1 m + (1-\beta 1)\frac{\partial L}{\partial w}\ v \leftarrow \beta 2 v + (1-\beta 2)(\frac{\partial L}{\partial w})2 \quad (1)$$

di mana $\beta 1$ dan $\beta 2$ adalah faktor penurunan momentum, dan diatur menjadi 0,9 dan 0,999, secara berturut-turut. Dimana α adalah learning rate, dan ϵ adalah sebuah bilangan kecil untuk mencegah pembagian dengan nol. Training model dilakukan dengan memperhatikan parameter Tabel 2.



Gambar 4 Evaluasi Model Nilai (A)

Perbedaan utama antara kedua set parameter adalah learning rate yang berbeda. Pada nilai parameter (A) mampu menghasilkan akurasi yang relatif stabil pada epoch terakhir



Gambar 5 Evaluasi Model Nilai (B)

Penggunaan nilai parameter (B) memang mempercepat proses training namun performa baik itu akurasi dan *loss value* dari model kurang konsisten pada awal dan akhir epoch training.

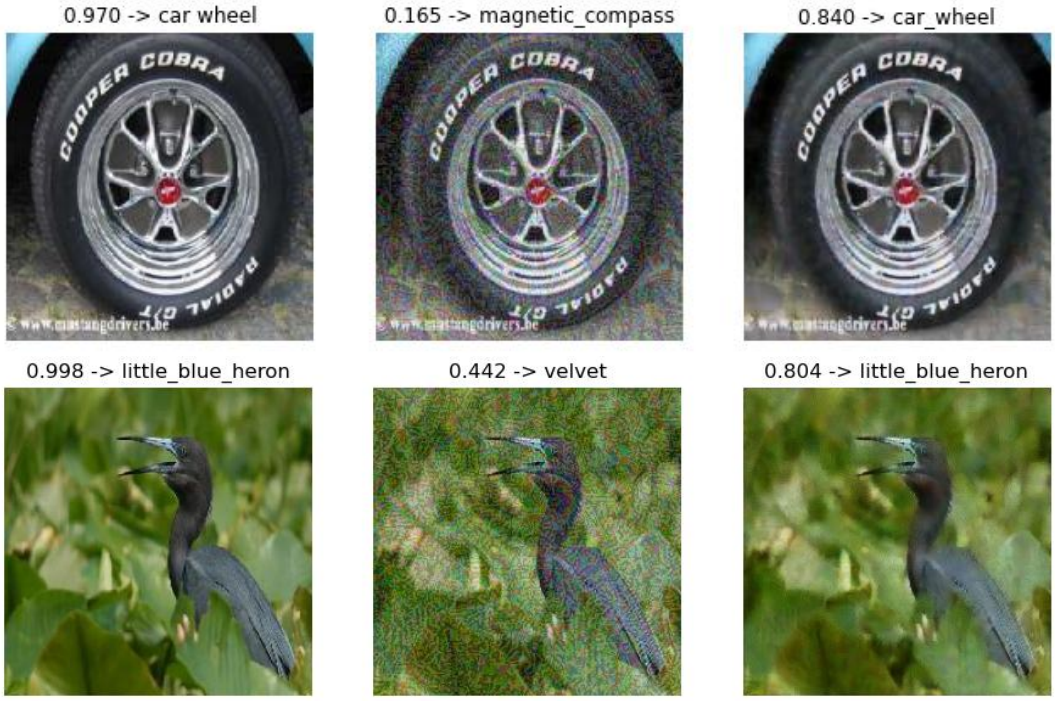
Tabel 2. Parameter Training dari Model Auto-Encoder

No.	Parameter	Nilai (A)	Nilai (B)
1	Sizes of AE layers	8, 5	8, 5
2	Activation Function AE	Sigmoid	Sigmoid
3	Optimizer	Adam	Adam
4	Loss Function	MSE	MSE
5	Learning Rate	0.002	0.003
6	Number of Epoch	20	20
7	Batch size when training AE	25	25

Proses training dilakukan dengan mengevaluasi hasil seperti akurasi dan loss value, dapat dilihat pada gambar 4 dan gambar 5, sehingga parameter hasil (A) digunakan pada proses testing selanjutnya.

3. Hasil dan Pembahasan

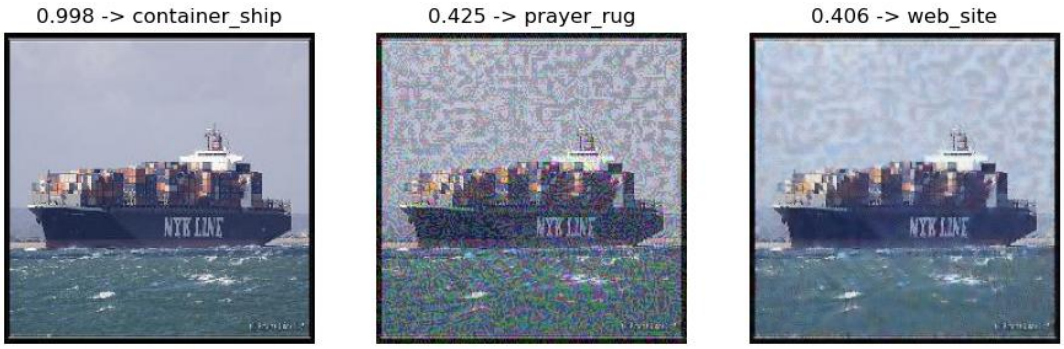
Ujicoba dilakukan dengan menyerang citra digital menggunakan metode FGSM pada model pembelajaran mesin Mobilenet dan melakukan mitigasi dengan menggunakan auto-encoder. Metode mitigasi ini dilakukan dengan merekonstruksi citra asli dari citra yang telah diserang, dengan tujuan menghilangkan noise dan meningkatkan keakuratan hasil klasifikasi. Hasil eksperimen menunjukkan bahwa metode mitigasi dengan auto-encoder dapat meningkatkan hasil klasifikasi yang lebih akurat. Training dan testing menggunakan *golden ratio* yaitu 80:20 untuk data training dan testing. Jumlah dataset 239 citra digital dari repositori https://github.com/EliSchwartz/imagenet-sample-images sehingga jumlah data training: 191 dan data testing: 48



Gambar 6. Dari kiri: Citra Asli, Citra Adversarial, Citra AE: Serangan Berhasil Mitigasi Berhasil

Gambar 4 menunjukkan bagaimana hasil deteksi awal sebuah citra digital original memiliki nilai confidence diatas 90 % citra tersebut diserang FGSM mengalami penurunan nilai confidence sampai dengan 10 – 40% serta classifier salah mengenali target class dari citra tersebut. Citra kemudian dilakukan mitigasi dengan auto-encoder, membersihkan noise membuat classifier kembali memperoleh nilai confidence yang baik > 80%.

Gambar 7. Dari kiri: Citra Asli, Citra Adversarial, Citra AE: Serangan gagal Mitigasi Berhasil



Gambar 8. Dari kiri: Citra Asli, Citra Adversarial, Citra AE: Serangan berhasil Mitigasi Gagal

Namun, perlu diingat bahwa meskipun teknik mitigasi berhasil mengurangi dampak serangan, masih terdapat sejumlah citra yang gagal untuk diklasifikasi oleh *classifier* seperti pada gambar 6. Hal ini menunjukkan adanya kompleksitas dalam mengatasi serangan adversarial secara menyeluruh.

Tabel 3. Hasil yang diperoleh pada penelitian

Parameter	Nilai
Citra yang berhasil diserang sebanyak	41
Citra yang gagal diserang sebanyak	7
Mitigasi berhasil sebanyak	44
Mitigasi gagal sebanyak	4
Akurasi Terhadap Citra Asli	92.64%
Akurasi Terhadap Citra Adversarial Attack	14.58 %
Akurasi Terhadap Citra Dimitigasi	91.67%

4. Kesimpulan

Melalui pengujian dan evaluasi, ditemukan bahwa akurasi model serangan (Papernot et al. 2016) mencapai 14.58 %, yang menunjukkan kemampuan model dalam mengenali citra yang telah diberi serangan. Sementara itu, akurasi model mitigasi mencapai 91.67%. Tabel 3 menunjukkan efektivitas teknik mitigasi dengan Auto-encoder dalam meningkatkan ketahanan model terhadap serangan adversarial pada citra digital. Adversarial attack memiliki konsekuensi yang penting dalam keamanan model machine learning. Adversarial examples yang dihasilkan dapat menyebabkan model yang dilatih memberikan prediksi yang salah atau dipengaruhi dengan sangat mudah oleh perubahan kecil pada input. Oleh karena itu, perlu dikembangkan pemanfaatan Auto-encoder sebagai metode pertahanan memiliki akurasi sebesar 91.67% untuk melindungi model dari serangan adversarial. Penggunaan metode Auto-Encoder sebagai teknik mitigasi berhasil mengurangi dampak serangan pada sebagian besar citra digital dalam dataset ImageNet. Namun, penelitian lebih lanjut diperlukan untuk mengembangkan teknik mitigasi yang lebih kuat dan dapat mengatasi citra yang masih rentan terhadap serangan.

5. Referensi

Bank, Dor, Noam Koenigstein, and Raja Giryes. 2020. “Autoencoders.”

Bhagoji, Arjun Nitin, Daniel Cullina, Chawin Sitawarin, and Prateek Mittal. 2017. “Enhancing Robustness of Machine Learning Systems via Data Transformations.”

Carlini, Nicholas, and David Wagner. 2016. "Towards Evaluating the Robustness of Neural Networks." *Proceedings - IEEE Symposium on Security and Privacy* 39–57. doi: 10.48550/arxiv.1608.04644.

Chakraborty, Anirban, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. "Adversarial Attacks and Defences: A Survey."

Chrabaszcz, Patryk, Ilya Loshchilov, and Frank Hutter. 2017. "A Downsampled Variant of ImageNet as an Alternative to the CIFAR Datasets."

Dyas Irvan Masruri Sugeng Widodo, and Febry Eka Purwiantono. 2021. "Implementation Of k-Means For Information Systems For The Spread Of Epidemic Diseases In Kota Malang". Vol 9 No 02 (2021): J-Intech : Journal of Information and Technology. https://doi.org/10.32664/j-intech.v9i02.638

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Networks."

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2014. "Explaining and Harnessing Adversarial Examples."

Gu, Shixiang, and Luca Rigazio. 2014. "Towards Deep Neural Network Architectures Robust to Adversarial Examples."

Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications."

Marzi, Zhinus, Soorya Gopalakrishnan, Upamanyu Madhow, and Ramtin Pedarsani. 2018. "Sparsity-Based Defense against Adversarial Attacks on Linear Classifiers."

Papernot, Nicolas, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, Rujun Long, and Patrick McDaniel. 2016. "Technical Report on the CleverHans v2.1.0 Adversarial Examples Library."

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. "ImageNet Large Scale Visual Recognition Challenge."

Sahay, Rajeev, Rehana Mahfuz, and Aly El Gamal. 2018. "Combatting Adversarial Attacks through Denoising and Dimensionality Reduction: A Cascaded Autoencoder Approach."

Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. "MobileNetV2: Inverted Residuals and Linear Bottlenecks."