Implementasi *Adversarial Perturbations* untuk Pengamanan Gambar Digital dari Plagiasi AI

Annel Rashka Perdana 18220026 (*Author*)
Program Studi Sistem dan Teknologi Informasi
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung
E-mail: Annel Rashka Perdana

***Abstract*—Di era digital ini, melindungi gambar digital atau karya seni dari plagiasi oleh kecerdasan buatan (AI) menjadi tantangan yang semakin mendesak. Dengan perkembangan pesat teknologi AI, kemampuan untuk memplagiasi dan memanipulasi gambar digital secara otomatis telah mencapai tingkat yang mengkhawatirkan. Salah satu pendekatan yang menjanjikan untuk mengatasi masalah ini adalah penggunaan a*dversarial perturbations*, yaitu teknik untuk menambahkan gangguan kecil namun signifikan ke dalam gambar asli. Gangguan ini mengubah interpretasi gambar oleh model AI tanpa mengubah penampilan visualnya bagi mata manusia. Penelitian ini fokus pada implementasi model adversarial perturbations yang sudah ada untuk mengamankan gambar digital dari plagiasi AI. Pada penelitian ini, dievaluasi efektivitas teknik ini dalam melindungi gambar digital dan diukur dampaknya terhadap kualitas visual gambar asli. Hasil penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam melindungi karya seni digital dari ancaman plagiasi oleh AI serta menyediakan panduan praktis untuk implementasi teknik adversarial perturbations dalam berbagai aplikasi di dunia nyata.**

Keywords— adversarial perturbations, keamanan gambar digital, plagiasi AI, perlindungan karya seni, kecerdasan buatan

I. Pendahuluan

Pengamanan gambar digital atau karya seni dari plagiasi oleh kecerdasan buatan (AI) merupakan tantangan yang semakin mendesak di era digital ini. Dengan perkembangan pesat teknologi AI, kemampuan untuk menduplikasi dan memanipulasi gambar digital secara otomatis telah mencapai tingkat yang mengkhawatirkan. Salah satu pendekatan yang menjanjikan untuk mengatasi masalah ini adalah penggunaan adversarial perturbations, yaitu teknik untuk menambahkan gangguan kecil namun signifikan ke dalam gambar asli, sehingga mengubah interpretasi gambar tersebut oleh model AI tanpa mengubah penampilannya secara signifikan bagi mata manusia[1].

Adversarial perturbations memanfaatkan kelemahan dalam model AI, khususnya dalam cara model ini memproses dan mengenali pola dalam data visual. Studi-studi sebelumnya telah menunjukkan bahwa dengan gangguan yang sangat kecil dan tidak terlihat oleh mata manusia, model AI dapat dibuat untuk salah mengenali atau bahkan gagal mengidentifikasi gambar tersebut secara akurat[2]. Hal ini memberikan peluang untuk melindungi karya seni digital dari penggunaan tanpa izin atau plagiasi oleh sistem AI yang semakin canggih.

Penelitian ini berfokus pada implementasi model adversarial perturbations yang sudah ada untuk mengamankan gambar digital dari plagiasi AI. Pada penelitian ini, tidak dilakukan pengembangan model baru, tetapi memanfaatkan teknik yang telah terbukti efektif dalam literatur untuk diaplikasikan pada kasus penggunaan yang spesifik ini. Model *adversarial perturbations* yang digunakan mencakup metode *Universal Adversarial Perturbations* (UAP) dan metode Fast *Gradient Sign Method* (FGSM) yang telah terbukti efektif dalam berbagai konteks[3][4].

Dengan menerapkan teknik-teknik ini, penelitian ini berupaya untuk mengevaluasi seberapa efektif mereka dalam melindungi gambar digital dari plagiasi AI. Selain itu, penelitian ini juga akan mengukur dampak dari gangguan yang ditambahkan terhadap kualitas visual gambar asli, untuk memastikan bahwa gambar tetap dapat diterima oleh pengguna manusia tanpa mengurangi nilai estetika atau fungsionalnya.

II. Related Work

A. Universal Adversarial Perturbations (UAP)

Universal Adversarial Perturbations (UAP) diperkenalkan oleh Moosavi-Dezfooli et al. yang menunjukkan bahwa adalah mungkin untuk menghasilkan gangguan universal yang dapat menipu berbagai model AI secara konsisten [1]. Teknik ini telah digunakan sebagai dasar dalam banyak penelitian berikutnya karena efektivitasnya dalam berbagai konteks aplikasi, termasuk pengamanan gambar digital. UAP memungkinkan pembuatan perturbasi yang efektif untuk berbagai gambar dengan gangguan yang sama, memberikan solusi praktis untuk melindungi sejumlah besar gambar secara efisien.

B. Adversarial Attack pada Deep Learning

Akhtar dan Mian melakukan survei komprehensif tentang ancaman serangan adversarial pada deep learning di bidang visi komputer [2]. Mereka menyoroti berbagai metode serangan dan pertahanan, serta memberikan wawasan tentang kelemahan model AI saat ini terhadap serangan adversarial. Survei ini memberikan dasar penting bagi penelitian ini dalam memilih metode adversarial perturbations yang tepat. Selain

itu, mereka juga mengidentifikasi teknik-teknik defensif yang dapat digunakan untuk melindungi model AI dari serangan ini.

C. Fast Gradient Sign Method (FGSM)

Goodfellow, Shlens, dan Szegedy memperkenalkan metode Fast Gradient Sign Method (FGSM), yang merupakan salah satu teknik serangan adversarial paling terkenal dan banyak digunakan [4]. Metode ini menggunakan gradien dari fungsi kerugian model AI untuk menghasilkan gangguan yang dapat menipu model dengan efisien. FGSM telah diterapkan dalam penelitian ini sebagai salah satu metode dasar untuk menghasilkan adversarial perturbations. FGSM dikenal karena kesederhanaannya dan efektivitasnya dalam menghasilkan gangguan yang dapat menipu model AI dengan perubahan kecil pada input.

III. Methodology

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

D. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

E. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive.”
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter,” not “webers/m2.” Spell units when they appear in text: “...a few henries,” not “...a few H.”

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors)*.

- Use a zero before decimal points: “0.25,” not “.25.” Use “cm3,” not “cc.” (*bullet list*)

F. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1),” not “Eq. (1)” or “equation (1),” except at the beginning of a sentence: “Equation (1) is ...”

G. Some Common Mistakes

- The word “data” is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o.”
- In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an “inset,” not an “insert.” The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
- Do not use the word “essentially” to mean “approximately” or “effectively.”
- In your paper title, if the words “that uses” can accurately replace the word using, capitalize the “u”; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones “affect” and “effect,” “complement” and

"compliment," "discreet" and "discrete," "principal" and "principle."

- Do not confuse "imply" and "infer."
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al."
- The abbreviation "i.e." means "that is," and the abbreviation "e.g." means "for example."

An excellent style manual for science writers is [7].

IV. Experiments and Results

V. Conclusion

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

H. Authors and Affiliations

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

1) For author/s of only one affiliation (Heading 3): To change the default, adjust the template as follows.

a) Selection (Heading 4): Highlight all author and affiliation lines.

b) Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select "1 Column" from the selection palette.

c) Deletion: Delete the author and affiliation lines for the second affiliation.

2) For author/s of more than two affiliations: To change the default, adjust the template as follows.

d) Selection: Highlight all author and affiliation lines.

e) Change number of columns: Select the "Columns" icon from the MS Word Standard toolbar and then select "1 Column" from the selection palette.

f) Highlight author and affiliation lines of affiliation 1 and copy this selection.

g) Formatting: Insert one hard return immediately after the last character of the last affiliation line. Then paste down the copy of affiliation 1. Repeat as necessary for each additional affiliation.

h) Reassign number of columns: Place your cursor to the right of the last character of the last affiliation line of an even numbered affiliation (e.g., if there are five affiliations, place your cursor at end of fourth affiliation). Drag the cursor up to highlight all of the above author and affiliation lines. Go to Column icon and select "2 Columns". If you have an odd number of affiliations, the final affiliation will be centered on the page; all previous will be in two columns.

I. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include ACKNOWLEDGMENTS and REFERENCES, and for these, the correct style to use is "Heading 5." Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract," will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named "Heading 1," "Heading 2," "Heading 3," and "Heading 4" are prescribed.

J. Figures and Tables

3) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1," even at the beginning of a sentence.

TABLE I. Table Styles

Table Head	Table Column Head		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

a. Sample of a Table footnote. *(Table footnote)*

b.

Fig. 1. Example of a figure caption. *(figure caption)*

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization," or "Magnetization, M," not just "M." If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization (A (m(1)," not just "A/m." Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)," not "Temperature/K."

Video Link at Youtube *(Heading 5)*

References

[1] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," arXiv preprint arXiv:1610.08401, 2017, doi: 10.48550/arXiv.1610.08401.

[2] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," IEEE Access, vol. 6, pp. 14410-14430, 2018.

[3] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," arXiv preprint arXiv:1607.02533, 2016.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 12 Juni 2024

Nama dan NIM