Security of Text to Icon/Image conversions

Zobaida Alssadi, Marius Silaghi
Department of Computer Science
Florida Institute of Technology
Melbourne, FL, USA
zalssadi2021@my.fit.edu, msilaghi@fit.edu

Abstract

Using images and icons to represent news or narratives has grown in popularity. Still, one critical problem is that they are not equivalent to language, making them vulnerable to adversary attacks. This study examines the impact of image-poisoning attacks based on polysematic words and of image attacks based on cultural differences when converting text to images. Such attacks can lead to the loss of important information and create confusion and incorrect interpretations of the intended meaning, misinforming the general public. This study specifically focuses on possible effects in a news and story context. This study highlights the significance of taking security considerations into account when image-based attacks are relevant and urges the development of strategies to ensure that information is conveyed through images and icons in a culturally appropriate and accurate manner, as well as to prevent image tampering and the spread of false information by hackers.

Introduction

While the automatic generation of image-based representation of texts becomes more common, we highlight new challenges and attacks on their applications involving cultural differences and poisoned image selection. Images/icons have gained popularity due to their visual appeal, attention-grabbing qualities, communication efficiency, and adaptability to different mediums. Technology has made it easier to create, share, access, and accept images/icons, leading to increased use in computer applications (Hemenway 1982).

Natural language processing and computer vision have both focused on the conversions of text to images/icons. These techniques train and evaluate models using human-annotated images/icons where the amount and quality of the training data are generally correlated to the achieved level of performance (Hossain et al. 2021).

There are, however, limitations and challenges associated with this where image-poisoning attacks based on polysematic words can create visually related but semantically incorrect icons or images. On the other hand, image attacks based on cultural differences could cause misconceptions and confusion, and these differences may alter how images/icons are understood.

In light of recent developments in Chatbot and BARD models, security vulnerabilities have become a growing concern. This paper explores the challenges of text-to-image/icon generation, including adversarial attacks and cultural differences that affect the interpretation of images/icons. By gaining a deeper understanding of these challenges, our research can contribute to the identification of effective defenses and tactics for enhancing the accuracy and appropriateness of visual communication using images and icons. This can have significant implications for developing more reliable and robust text-to-image/icon generation techniques and help advance the field of natural language processing and computer vision.

Related work

DALLE 2 is an OpenAI program that uses deep learning architecture to convert natural language inputs into realistic images. It has been connected to the comprehension of human language and the invention of a language that is not immediately intuitive to humans (Leivada, Murphy, and Marcus 2022).

In this study, we use DALLE 2’s image-generation capability to demonstrate some of its potentials. An example output is shown in Figure 1 for the text:

"Einstein paints while Newton solves puzzles. Einstein is wearing a lab coat. Newton is wearing a t-shirt, pants, and a cap."

Figure 1: DALLE 2 generated image based on the given input text

Midjourney is an open-access program that uses text questions to generate visuals by comparing internet images to user-submitted requests (King 2022). The same text script
has been used to illustrate a sample of this tool’s potential and capability and Figure 2 shows the result.

![Figure 2: MidJourney generated image based on the given input text](image)

Text2Icon uses AI to generate icon-based visualizations of text, analyzing words, activities, and objects to create a series of icons that visually depict the text’s essential concepts and plot points.

Figure 3 shows the generation of the following text into icons using Text2Icons.

**Breaking News: The price of Bitcoin crashed after Russian President Vladimir Putin announced a military operation in the Donbas region of Ukraine.**

![Figure 3: Text2Icon generated image based on the given input text](image)

**Background**

The main topics of previous research and publications have been the expressiveness of images/icons and their capability to clearly convey complicated ideas to a broad audience. Text2Icons is an AI method that converts text descriptions into icons making it simpler to comprehend and recall complex information. The system determines the best icons to reflect the meaning of the text input using NLP and machine learning methods (Valente 2021).

This tool converts text into icons using natural language processing (NLP) and ML algorithms to analyze the inputs and then extracts the features to comprehend the text’s content and choose the suitable icons to convey the story using a specified icon database. ML algorithms are then trained to determine which icons correspond to specific text inputs. The resulting narrative or news is then shown as a series of icons that convey the text input’s intended meaning. The system intends to make text-based stories or news more accessible and understandable by presenting them visually. The study also discusses the possible advantages of utilizing icons to communicate stories, including better knowledge retention, simpler communication, and more compelling material (Valente 2021).

Deep learning-based NLP systems are vulnerable to adversarial attacks, which can cause false or misleading results (Marulli, Verde, and Campanile 2021). Image poisoning attacks are attacks on machine learning models that alter training data to misclassify a sample or class. There are two types of image poisoning attacks: triggerless and backdoor attacks.

1. **Triggerless attacks:** in techniques like Feature Collision (FC) and Convex Polytope (CP), poison samples are produced by introducing minor alterations to the base pictures to make their feature representations resemble those of the target image.

2. **Backdoor attacks:** like Class-Conditional Label-Based Poisoning (CLBD) and High-Transfer Backdoor Poisoning (HTBD), call for the development of poison examples with a concealed trigger, like a particular patch or pattern, that causes the model to categorize the target image incorrectly when the trigger is present (Schwarzschild et al. 2021).

Adversarial attacks on images using deep neural networks (DNNs) threaten the security and reliability of DNN-based systems due to their sensitivity to even minute input changes. Image adversarial attacks are techniques used to gradually change an image while keeping its visual appearance, leading to a machine learning model incorrectly identifying the modified image.

1- **Perturbation-based attacks:** These attacks poison the original image somewhat, usually by adding noise, in an effort to deceive the model into classifying it incorrectly (Prakash et al. 2018).

2- **Reconstruction-based attacks:** These attacks involve reconstructing the image from scratch to cause the model to misclassify it. Examples include the Carlini-Wagner attack (Carlini and Wagner 2017).

3- **Evasion attacks:** By taking advantage of weaknesses in the model’s design or training set, these attacks aim to get the model to identify an image incorrectly (Carlini and Wagner 2017; Prakash et al. 2018).

The expressiveness of icons in expressing meaning and information has been assessed in earlier studies and papers. The study (dos Reis et al. 2016) looked at variables that can affect an icon’s expressiveness, including cultural background, user familiarity, level of abstraction, usage of color and animation, and the inclusion of text labels. Moreover, it has been estimated that it is difficult to produce icons that are both internationally recognizable and culturally neutral (Holtz, Nocun, and Hansen 2011).

**Formal Problem Statement**

Converting text to icons and images can be vulnerable to two types of attacks: image-poisoning attacks based on polysemantic words and image attacks based on cultural differences. Image poisoning can alter icon selection and produce
false or damaging images, while cultural differences can lead to misinterpretations and misunderstandings of icons and images, which can spread false information and deceive the public.

Here we look at how these adversarial attacks, which may lead to losing important information or details, can have an impact. This can result in misunderstanding or incorrect interpretations of the intended meaning. The study specifically focuses on the possible effects in a news and story context based on an individual’s background, culture, and educational level, where it is possible for erroneous information to propagate or for the general public to be misinformed.

The study highlights the importance of protecting against hackers tampering with images and disseminating false information to ensure accurate and suitable information exchange.

Evaluation

**Image-poisoning attacks based on polysemantic words:**
An attacker can manipulate the actions of AI algorithms that aid in converting text to images or icons by employing poisoning techniques. For instance, they may include a “backdoor” that would enable them to activate particular, unwanted behaviors when certain trigger words or phrases appeared in the input text. Changing the model’s architecture or weight parameters can also affect AI output negatively.

For example, if the attacker wants the target image to be identified as the sun but it is actually a picture of the moon, they could alter the training data by adding a minor pattern (like a sticker or a particular arrangement of pixels) to the training set’s images of the sun. After training with the tainted data set, the model will wrongly identify the target image as the sun.

Such an attack is risky because it can result in inaccurate predictions and it’s difficult to detect whether a model has been attacked. This could lead to serious security and privacy issues such as the spread of false information and the theft of private data.

Attacks on text-to-icon or image-generation models may produce visually identical but semantically inaccurate results, particularly in news contexts where disseminating false or misleading information may have far-reaching effects.

Example: Let us consider a news report on a new digital currency that is shown as a golden coin. "Breaking News: A new digital currency crushes Competitors!" In this case, an attacker could poison the picture database with a golden coin’s image made to appear damaged or otherwise broken. That may be in contrast to a reality where this digital currency is secure and trustworthy. This could give the impression that the coin is unreliable or prone to failure.

Investing in a new digital currency based on inaccurate information can negatively affect its reputation and market value.

Another example: Breaking News: Scientists Make Breakthrough in Cancer Research with New Treatment, where the breakthrough is represented by an image of a rocket launching into space.

**Image attacks based on cultural differences** can happen when the meaning of an image or icon is culturally specific and may not be universally understood. For example, an image representing a gesture or symbol may have a different meaning or connotation in a different cultural context.

Attacker could take advantage of this attack by intentionally using images that are culturally inappropriate, offensive, or misleading to spread misinformation or create confusion among the target audience.

**New Study Finds Clockwise Running May Boost Brain Function**

In various cultures, the clockwise symbol can signify various things, from prosperity and good luck in Hinduism to the Buddha’s footsteps in Buddhism. This emphasizes how
important it is to consider cultural differences when developing icons or symbols for communication (Phillips 2022).

Figure 7: Text2Icon generated image based on the given input text.

Another example is shown in Figure 7 when converting the text “Millions of Muslims travel to the Kaaba in Mecca every year to perform the Hajj pilgrimage.” to icons using Text2Icons. Showing these icons to people from non-Muslim cultures, it’s possible that their comprehension of the icons will be restricted owing to their lack of knowledge of the cultural and religious importance of the Hajj pilgrimage and the Kaaba in Mecca. In addition, the icons would fail to express the religious or cultural context of the Hajj pilgrimage, making it impossible for anyone outside of the Muslim community to understand the meaning behind the icons completely.

Accordingly, cultural and background differences must be considered when converting text to icons and pictures, as ignoring them can lead to misunderstandings, misinterpretation, and security problems. Designers and developers must consider these differences when creating icons, as an icon that symbolizes safety in one culture could signify danger in another.

Discussion

DALLES DALLES was given the prompt, “Einstein paints while Newton solves puzzles. Einstein is wearing a lab coat. Newton is wearing a t-shirt, pants, and a cap”.

The analysis attempts to demonstrate DALLE 2’s capacity to understand natural language prompts and produce inventive synthetic images matching the input text. A critical detail that may not be conveyed through a simple image or icon is that Newton is wearing a cap, a t-shirt, and pants while Einstein is wearing a lab coat. This may cause misinterpretations of the narrative and the lack of necessary context.

Moreover, cultural, political, and individual biases can also influence the human understanding of the images. For example, people with a knowledgeable background might associate these two famous physicists with intelligence, seriousness, and science. Thus, the story’s context is affected, as this may turn into a completely different interpretation of the story. Furthermore, personal biases can influence an outfit’s perception, such as a lab coat representing knowledge and education, while a t-shirt or jeans representing casualness.

MidJourney has been used to generate an image based on the same input text, it produces a unique artistic representation of Einstein and Newton wearing white coats and appearing as if they are talking where one of them sits at the desk with papers and an ink box in front of him, while the other is standing.

That might be attractive visually. However, as a general AI image production tool, Midjourney might be unable to catch all of the text’s finer points and nuances. This means some important information may be left out, leading to a less accurate representation.

Text2Icons uses natural language processing to identify important keywords and concepts in the input text, then searches a database for relevant icons and arranges them to represent the text visually.

In Figure 3, The car crushed icon was created to express “Crush” from the text prompt. This icon can indicate that a car accident rather than a market decline caused the crash, which might allow false information to propagate. As it is more challenging to confirm the integrity of an icon-based message, using icons instead of text can make it simpler for malicious actors to disseminate false information or tricks which can harm the audience by causing them to make poor investment decisions.

Conclusions and Outlook

AI tools like Text2Icons, converting text into icons/images, present a promising approach to making complex information more attractive (Valente 2021). However, some restrictions and difficulties come with this. One significant issue is these algorithms are vulnerable to adversarial attacks, including image-poisoning attacks based on polysemantic words and image attacks based on cultural differences.

Image-poisoning attacks manipulate images or subtly alter words to create multiple meanings or interpretations when translated into images or icons.

For example, depending on the situation, an attacker might replace another word for “bank,” switching between “bank of a river” or “bank account”, having a different meaning when used in an icon or image. The AI tool could fail to distinguish between the meanings and might create a picture or icon that suggests a meaning that was not intended, causing misunderstanding or disinformation.

On the other hand, image attacks based on cultural differences essentially use cultural nuances and variations in how things are seen. Attackers can use this vulnerability to alter the meaning of images to communicate a different message. Additionally, an attacker could use images or icons that are unfamiliar or not commonly used by the intended audience.

For example, a positive or neutral icon in one culture may have a harmful or offensive meaning in another. Additionally, an attacker could use images or icons that are unfamiliar or not commonly used by the intended audience, which could cause confusion and misinterpretation.

Hence, image-based attacks, whether due to polysemantic words or cultural differences, can have significant consequences such as the loss of critical information, confusion, and affecting public opinion, decision-making, and trust.

A social interview experiment and survey are being conducted to gain a deeper understanding of the impact of image attacks on individuals from diverse backgrounds. Its goal is to identify effective defenses and tactics for enhancing visual communication using images and icons.
References


