Mixtura - A Diffusion-Based Image CAPTCHA Generation Mechanism Leveraging Equidistant Latent Space Clustering

Parag Paul
CTO, Rig AI
Seattle, WA, USA
parag@rigai.co

Saurabh Sarkar
CTO, Chicory
Chicory AI, Seattle, USA
sarkar@chicory.ai

Avinash Anand
CEO, Rig AI
Seattle, WA, USA
avinash@rigai.co

Prof. Jeff Ku
Department of Computer Science
University of California, Berkeley, USA
wzk0004@auburn.edu

***Abstract*—This paper introduces a novel approach to generate resilient visual CAPTCHAs using Stable Diffusion, a state-of-the-art generative model. Drawing upon the insights from the original model, we introduce combinations in the latent space organization induced by diffusion processes, thus creating a novel methodology for generating CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) images, which can be leveraged for bot detection. Our approach leverages a controlled multiphase diffusion process to synthesize visual challenges that are inherently difficult for automated systems to decipher while remaining readily and intuitively solvable by human observers. The objective is to generate CAPTCHA images that adhere to a coherent perceptual logic, embodying features such as visual continuity, familiar symmetry, and rudimentary objects, which are fundamental to human visual processing. Simultaneously, the integrated application of Gaussian noise at multiple phases introduces an increased level of controlled indistinctness, effectively disrupting the straightforward analytic models typically employed by AI-driven image recognition systems.**

I. Introduction

CAPTCHAs (Completely Automated Public Turing Test to Tell Computers and Humans Apart) remain critical for securing web interactions against bots. Traditional text-based CAPTCHAs suffer from increased vulnerability due to advances in optical character recognition (**OCR**) and deep learning. Recent advances in deep learning have significantly altered the landscape of CAPTCHA security, revealing vulnerabilities that were previously considered robust against automated attacks. In particular, deep neural networks (**DNNs**), particularly convolutional neural networks (**CNNs**), have demonstrated remarkable efficacy in breaking various types of CAPTCHA, particularly text-based implementations.

For example, deep learning techniques, especially using **CNNs**, have effectively targeted and broken traditional text-based CAPTCHAs. Research indicates that sophisticated deep learning models have achieved high accuracy in recognizing distorted text, which has long been a primary defense mechanism for these CAPTCHAs (Siddhartha et al., 2023 [1]; Kovács & Tajti, 2023; [2](Che et al., 2021) [3]. These models utilize vast amounts of labeled training data to learn to identify characters and words even under significant distortion, thereby undermining the CAPTCHA's intended purpose of distinguishing between human and automated responses (Kwon et al., 2020 [4]; (Nouri & Rezaei, 2020) [5].

An important example of this trend was highlighted in a study conducted by Nouri and Rezaei, who reported a customized deep neural network model that achieved a cracking accuracy of 98.94% for numerical datasets and 98.31% for alphanumeric datasets (Nouri & Rezaei, 2020). This impressive performance underlines the urgent need for new CAPTCHA designs that can withstand such advances in machine learning.

To address these limitations, we explore the use of generative diffusion models, specifically Stable Diffusion, for producing dynamic, on-the-fly CAPTCHAs. Our contribution lies in applying equidistant clustering constraints to structure visual concepts in grid or simpler layouts while introducing semantic noise to hinder algorithmic exploitation. We present a two-phase generation pipeline that combines text-to-image synthesis and spatial arrangement logic, suitable for web- or game-based verification systems.

II. Methodology

Our pipeline consists of four major components:

A. Class Definition and Prompt Design:

We select multiple visual categories *(e.g., "human," tree", "busstand")* and generate image batches for each class using prompt-engineered **Stable Diffusion** queries. Prompt diversity and style modifiers *(e.g. "pixel art," "isometric")* introduce visual variance.

B. Equidistant Spatial Clustering:

This section describes two models for equidistant spatial clustering:

1) Morphing and Incremental De-noising Model : This model involves morphing various classes within images, followed by incremental de-noising using U-net architecture from diffusion models. The goal is to identify a point in the search space equidistant from all clusters. It serves faster than the grid approach; however, an image with multiple latent features, although evident to humans, challenges machine analysis.

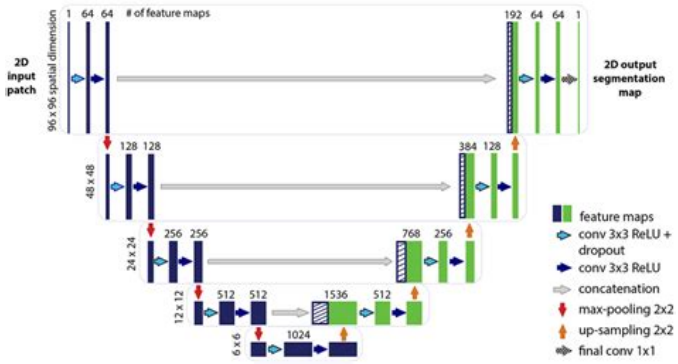

Fig. 1. The concatenation step is augmented by state of other classes at this exac stage in the noising phase

2) Concatenation based Approach: This method avoids morphing multiple classes, instead focusing on a gradual denoising approach. Multiple pipelines are created for each image class, and their intermediary stages are merged with the target-class pipeline image at the same stage. For instance, to include a human face in a CAPTCHA, the process starts with a human face, and during concatenation, latent features from other classes like animals or furniture are added during the concatenation state. This process provides additional non-original latent features. Both approaches feature parametric modifications, altering the phase of addition, repetition weight, and number of classes, with the aim of preserving latent features. Various modifications are created using different parametric combinations, such as the phase at which the addition is done, the weight given to the repetitions, and the number of classes used, so that we do not lose all the latent features.

C. Noise Injection and Distortion:

Further ambiguity can be added using post-processing. Each image is post-processed with blur, jitter, hue shift, and local pixel noise. Gaussian overlays and partially occluding shapes are optionally added to further obscure machine-learned patterns.

D. CAPTCHA Construction and Annotation:

The final grid is rendered with click-sensitive labels and metadata that indicate the true label of each cell. Human users are prompted to "select all [class] images."

III. Evaluation

We evaluated the effectiveness of our method along two axes:

Using pre-trained classifiers *(e.g. [6] CLIP, ResNet50 [7])*, we measure the precision of the top-1 in identifying the target class under varying noise levels and cluster overlap conditions. Our experiments show a 35-60% drop in machine precision with only a minor performance impact on human participants.

Fig. 2. The morphed image demonstrates how adjusting noise levels via diffusion can increase CAPTCHA complexity for machines while remaining moderately easy for humans to solve.

Usability and Human Accuracy: A group of 25 volunteers completed 60 randomized CAPTCHAs, yielding a 93% average precision and an average solve time of 4.5 seconds, validating both solvability and speed.

A. Conclusion and Future Work

This paper presents a generative, flexible, and robust method for CAPTCHA generation by combining **stable diffusion** image synthesis, equidistant clustering, and adversarial image corruption. Our approach is resistant to pre-trained image classifiers and makes them scalable across platforms. Current limitations for on-device computation will keep the image generation service as a service with more capabilities on the server end. The actual goal will be to create CAPTCHA's on the fly using technologies like WebGPU, WASM, and they could be used for on-device games for personalized CAPTCHA's. Another scope for use will be for *Proof of Humanity* systems like **Proof of Leg Work** [8] protocol, which uses puzzle solving in a way that humans can solve the puzzle in a moderately hard way, but it will be nearly impossible or computationally infeasible, fiscally difficult for machines to solve. This creates a solution where only humans can mine the **POLW** coin.

References

[1] A. Siddhartha, P. Rahul, A. A. Chaitanya, and D. Mohan, "Captcha recognition using dcnn," *Interantional Journal of Scientific Research in Engineering and Management*, 2023.

[2] Kovács and T. Tajti, "Captcha recognition using machine learning algorithms with various techniques," *Annales Mathematicae Et Informaticae*, 2023.

[3] A. Che, Y. Liu, H. Xiao, H. Wang, K. Zhang, and H. Dai, "Augmented data selector to initiate text-based captcha attack," *Security and Communication Networks*, 2021.

[4] H. Kwon, H. Yoon, and K.-W. Park, "Robust captcha image generation enhanced with adversarial example methods," *Ieice Transactions on Information and Systems*, 2020.

[5] Z. Nouri and M. Rezaei, "Deep-captcha: A deep learning based captcha solver for vulnerability assessment," 2020.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[8] P. W.-s. K. Parag Paul, Avinash Anand, "Proof of leg work," http://www.polw.cc, accessed: 2024-10-01.

***Index Terms*—Latent diffusion model, mixture of Gaussian noise,**

CAPTCHA

Mixtura - A Diffusion-Based Image CAPTCHA Generation Mechanism Leveraging Equidistant Latent Space Clustering

A modern security solution

Parag Paul, Saurabh Sarkar, Avinash Anand, Prof. Wei-shinn Ku

Collaboration between RigAI, POLW, Chicory AI

Parag Paul, Avinash Anand, Seattle
Washington
Email: parag@rigai.co
Saurabh Sarkar, Seattle
Washington

Email: sarkar@chicory.ai

Abstract

This paper introduces a novel approach to generate resilient visual CAPTCHAs using Stable Diffusion, a state-of-the-art generative model. Drawing upon the insights from the original model, we introduce combinations in the latent space organization induced by diffusion processes, thus creating a novel methodology for generating CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) images, which can be leveraged for bot detection. Our approach leverages a controlled multiphase diffusion process to synthesize visual challenges that are inherently difficult for automated systems to decipher while remaining readily and intuitively solvable by human observers. The objective is to generate CAPTCHA images that adhere to a coherent perceptual logic, embodying features such as visual continuity, familiar symmetry, and rudimentary objects, which are fundamental to human visual processing. Simultaneously, the integrated application of Gaussian noise at multiple phases introduces an increased level of controlled indistinctness, effectively disrupting the straightforward analytic models typically employed by AI-driven image recognition systems.

Introduction

CAPTCHAs (Completely Automated Public Turing Test to Tell Computers and Humans Apart) remain critical for securing web interactions against bots. Traditional text-based CAPTCHAs suffer from increased vulnerability due to advances in optical character recognition (**OCR**) and deep learning. Recent advances in deep learning have significantly altered the landscape of CAPTCHA security, revealing vulnerabilities that were previously considered robust against automated attacks. In particular, deep neural networks (**DNNs**), particularly convolutional neural networks (**CNNs**), have demonstrated remarkable efficacy in breaking various types of CAPTCHA, particularly text-based implementations.

For example, deep learning techniques, especially using **CNNs**, have effectively targeted and broken traditional text-based CAPTCHAs. Research indicates that sophisticated deep learning models have achieved high accuracy in recognizing distorted text, which has long been a primary defense mechanism for these CAPTCHAs (Siddhartha et al., 2023[**?**]; Kovács & Tajti, 2023; [**?**](Che et al., 2021)[**?**]. These models utilize vast amounts of labeled training data to learn to identify characters and words even under significant distortion, thereby undermining the CAPTCHA's intended purpose of distinguishing between human and automated responses (Kwon et al., 2020[**?**]; (Nouri & Rezaei, 2020)[**?**].

An important example of this trend was highlighted in a study conducted by Nouri and Rezaei, who reported a customized deep neural network model that achieved a cracking accuracy of 98.94% for numerical datasets and 98.31% for alphanumeric datasets (Nouri & Rezaei, 2020). This impressive performance underlines the urgent need for new CAPTCHA designs that can withstand such advances in machine learning.

To address these limitations, we explore the use of generative diffusion models, specifically Stable Diffusion, for producing dynamic, on-the-fly CAPTCHAs. Our contribution lies in applying equidistant clustering constraints to structure visual concepts in grid or simpler layouts while introducing semantic noise to hinder algorithmic exploitation. We present a two-phase generation pipeline that combines text-to-image synthesis and spatial arrangement logic, suitable for web- or game-based verification systems.

Methodology

Our pipeline consists of four major components:

0.1 Class Definition and Prompt Design:

We select multiple visual categories *(e.g., "human," tree", "busstand")* and generate image batches for each class using prompt-engineered **Stable Diffusion** queries. Prompt diversity and style modifiers *(e.g. "pixel art," "isometric")* introduce visual variance.

0.2 Equidistant Spatial Clustering:

This section describes two models for equidistant spatial clustering:

0.2.1 Morphing and Incremental De-noising Model

This model involves morphing various classes within images, followed by incremental de-noising using U-net architecture from diffusion models. The goal is to identify a point in the search space equidistant from all clusters. It serves faster than the grid approach; however, an image with multiple latent features, although evident to humans, challenges machine analysis.

0.2.2 Concatenation based Approach

This method avoids morphing multiple classes, instead focusing on a gradual denoising approach. Multiple pipelines are created for each image class, and their intermediary stages are merged with the target-class pipeline image at the same stage. For instance, to include a human face in a CAPTCHA, the process starts with a human face, and during concatenation, latent features from other classes like animals or furniture are added during the concatenation state. This process provides additional non-original latent features. Both approaches feature parametric modifications, altering the phase of addition, repetition weight, and number of classes, with the aim of preserving latent features. Various modifications are created using different parametric combinations, such as the phase at which the addition is done, the weight given to the repetitions, and the number of classes used, so that we do not lose all the latent features.

Figure 1: Override default concatenation path

0.3 Noise Injection and Distortion:

Further ambiguity can be added using post-processing. Each image is post-processed with blur, jitter, hue shift, and local pixel noise. Gaussian overlays and partially occluding shapes are optionally added to further obscure machine-learned patterns.

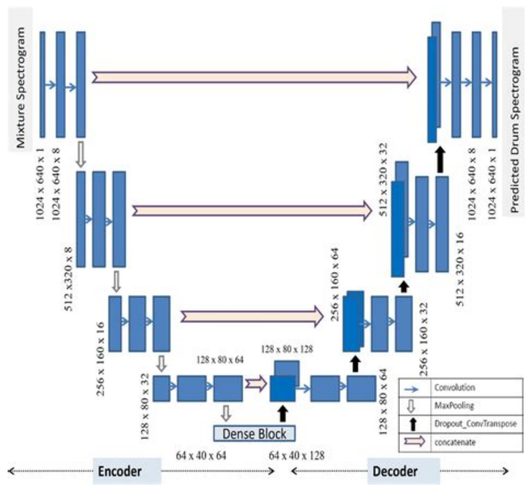

Figure 2: Override default concatenation path

We evaluated the effectiveness of our method along two axes:

Using pre-trained classifiers *(e.g. [**?**] CLIP, ResNet50[**?**])*, we measure the precision of the top-1 in identifying the target class under varying noise levels and cluster overlap conditions. Our experiments show a 35-60% drop in machine precision with only a minor performance impact on human participants.

Usability and Human Accuracy: A group of 25 volunteers completed 60 randomized CAPTCHAs, yielding a 93% average precision and an average solve time of 4.5 seconds, validating both solvability and speed.

Conclusions

This paper presents a generative, flexible, and robust method for CAPTCHA generation by combining **stable diffusion** image synthesis, equidistant clustering, and adversarial image corruption. Our approach is resistant to pre-trained image classifiers and makes them scalable across platforms. Current limitations for on-device computation will keep the image generation service as a service with more capabilities on the server end. The actual goal will be to create CAPTCHA's on the fly using technologies like WebGPU, WASM, and they could be used for on-device games for personalized CAPTCHA's. Another scope for use will be for *Proof of Humanity* systems like **Proof of Leg Work**[**?**] protocol, which uses puzzle solving in a way that humans can solve the puzzle in a moderately hard way, but it will be nearly impossible or computationally infeasible, fiscally difficult for machines to solve. This creates a solution where only humans can mine the **POLW** coin.