B20AI013_Minor2

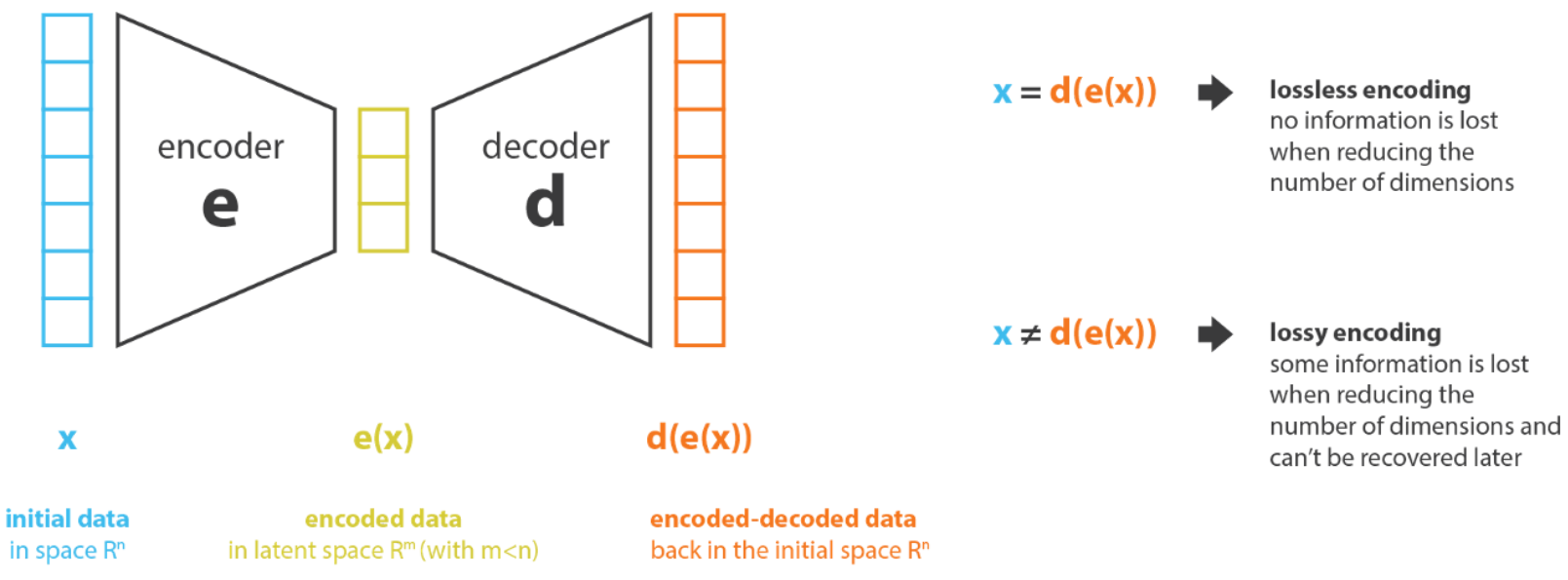
Ishaan Shrivastava - B20AI013

Course - Dependable AI

Question 1

Stable Diffusion is the latest addition to an entourage of generative diffusion models such as DALL-E and MidJourney. Unlike its competitors, Stable Diffusion has the ability to be hosted on a desktop or even a laptop equipped with a consumer grade GPU. This has made the text-to-image, inpainting and other generative capabilities of Diffusion models available to the public. Stable diffusion was created by StabilityAI, in cooperation with a large number of researchers and non-profit organizations.

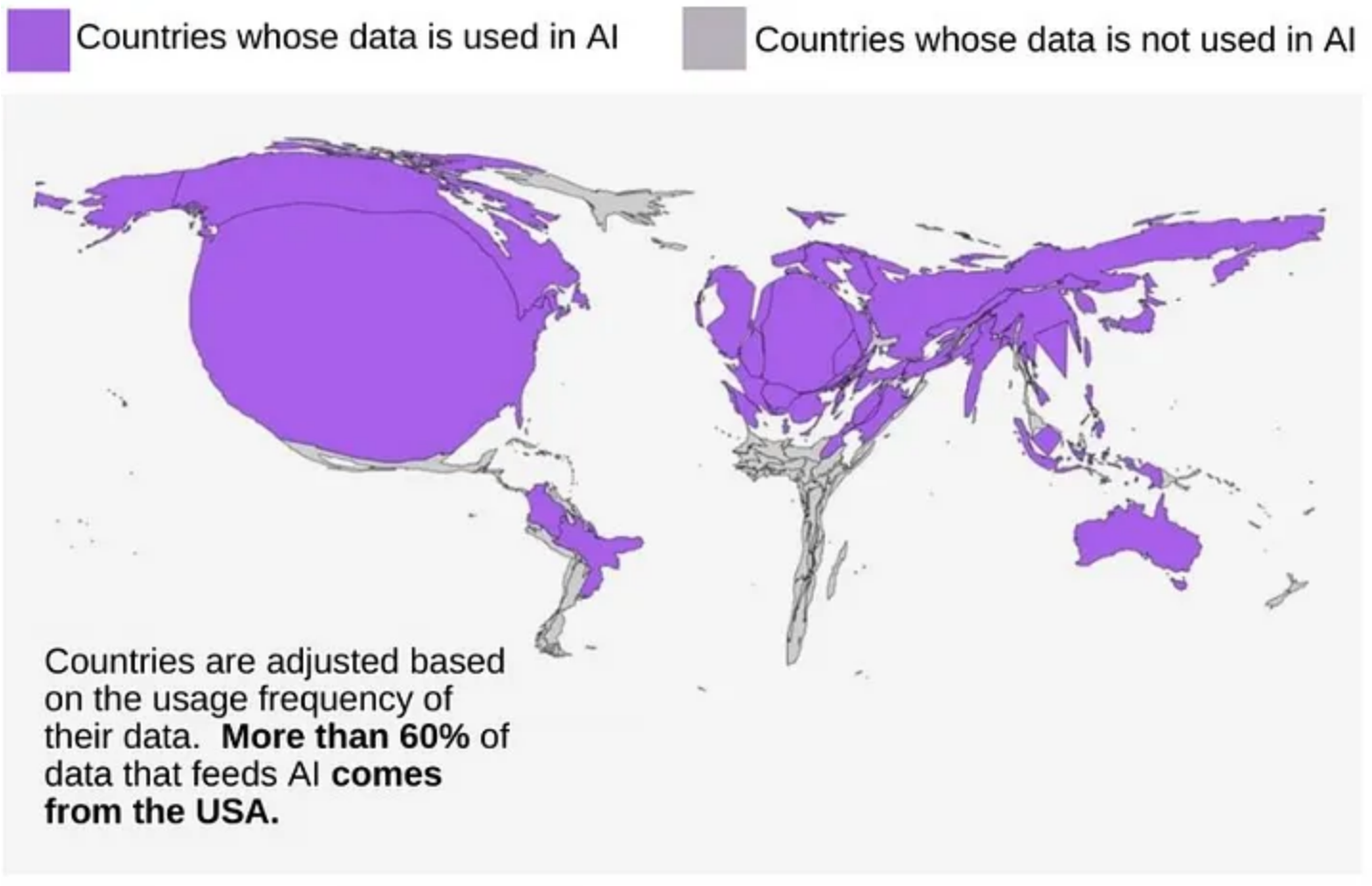
Stable diffusion works by transferring the diffusion process from the larger, bigger image-space to the smaller latent-space, which allows the model to be much smaller and be hosted on consumer-grade hardware accessible to anyone. The schematic is given below:



Stable diffusion is embroiled in controversies due to its abilities to near-replicate and extrapolate artworks that were a part of its training set, without the express permission or consent of the original creator. This controversry points at a larger issue than art itself, which is the issue of representation in AI.

According to the model card on github, it mentions: *"The model was trained on a large-scale dataset LAION-5B which contains adult material and is not fit for product use without additional safety mechanisms and considerations."*

Below is a schematic diagram of the regional representation of images use in the dataset. With this, let us explore some specific ways that I found in which Stable diffusion is biased.



1. If a character is asked to be generated, it has a tendency to be female.

   It usually takes an overwhelming amount of effort in order to get stable diffusion to generate males. This is representative of an important underlying societal bias where females are more likely to post images of themselves online [1]. This leads to a representation bias in the dataset which leads to the model generations preferring the generation of females over males even when explicitly stated otherwise.

💡 Prompt: “female in her 20s studying, spectacles, white tshirt, black trousers, on desk, monitor desktop”

Prompt: “someone reading a book, garden, light clothes,airy, blonde, slim”

2. If a university student is asked to be generated, they are always happy in the photo.

   Stable diffusion somehow never generates an unhappy college student. As if nobody is ever sad in college!!

Prompt: “University student holding a laptop”

Prompt: “University students in a library in a line”

3. Stable diffusion is heavily biased towards more recent generations. For example, prompting it to generate “Young Jimmy Carter 1978” will generate a fairly old-looking picture of him.

Question 2

Paper Review - Visualization of Deep Reinforcement Learning using Grad-CAM: How AI Plays Atari Games? - IEEE 2019

Reinforcement learning agents, DRLs in particular, suffer from the problem of a lack of explainability of outcomes. This is a major concern since we cannot trust an agent if it cannot explain its actions.

A3C (Asynchronous Advantage Actor-Critic) is a deep reinforcement learning algorithm that utilizes a critic network that learns the value function of the environment, and multiple parallelly-running actor networks which learns the policy of the agent.

In order to help visualize the spatial information flowing through the network better, this paper proposes modifying the architecture by adding pooling layers after every convolution. Then, Grad-CAM is implemented by computing the gradients uptil the last convolutional layer in order to visualize the importance of the various different pixels on the screen on the class-heatmaps thus produced.

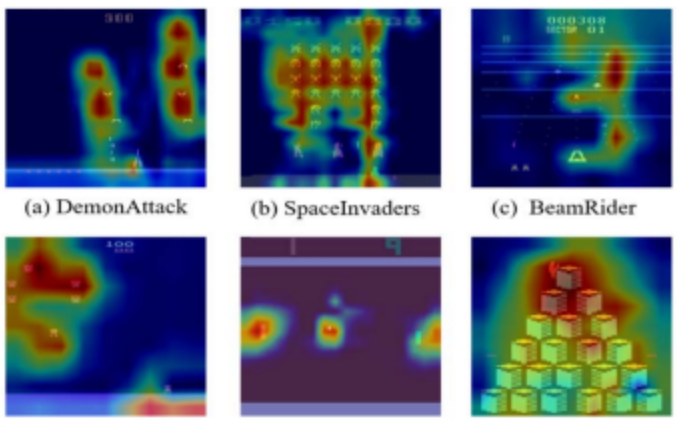


Fig. 3. Grad-CAM results for various Atari games

Thus, this paper demonstrates the usefulness of explainable, interpretable visualizations of a network by showing the distribution of attention of an agent across the input map, and showing that the agent learns to focus on different parts of the map depending on the game being played.

Problem

However, this paper fails to address the issue of bias in reinforcement learning. Consider for example, an agent playing the ubiquitous game, PacMan. Let us assume that we learnt an agent using reinforcement learning and that it seems to be performing very well at its task. We also visualize the attention and the decision making process of the agent by using the technique proposed in this paper. We observe that the attention is distributed on top of the powerups in the bottom right corner. Satisfied by what we see, we consider the model validated and continue with our deployment.
However, there is almost always a distributional shift between the training and the testing conditions. It just so happens that the agent encounters an iteration of the game where the powerup is located at the centre, however it keeps seeking the powerup in the bottom left due to its bias in learning that rewards are located at that corner of the map (incorrectly), as opposed to learning what we want it to learn, which is to identify the powerup correctly in a robust manner.

This is a very simple example but it aims at a significant problem with this paper: The Grad-CAM visualizations can mislead the observer into thinking that the model has learnt an optimizer whereas it simply learnt a few heuristics in its training that may not generalise in the deployment environment.

In order to fix this issue, I propose the adversarial training of an RL agent such that the adversary tries to maximise its objective by fooling the agent by generating environments that the agent does not yet generalise across, forcing it to learn what fools it. This forms a very robust training paradigm that should be free from the bias discussed above. All in all, It is important to note that Grad-CAM, although very useful as a tool for increasing the explainability of deep networks, should be used carefully and only when absolutely needed.

References

[1] Gender and image sharing on Facebook, Twitter, Instagram, Snapchat and WhatsApp in the UK: Hobbying alone or filtering for friends?