Content-Based Exploration of Archival Images Using Neural Networks

Tobi Adewoye, Xiao Han, Nick Ruest, Ian Milligan, Samantha Fritz, and Jimmy Lin
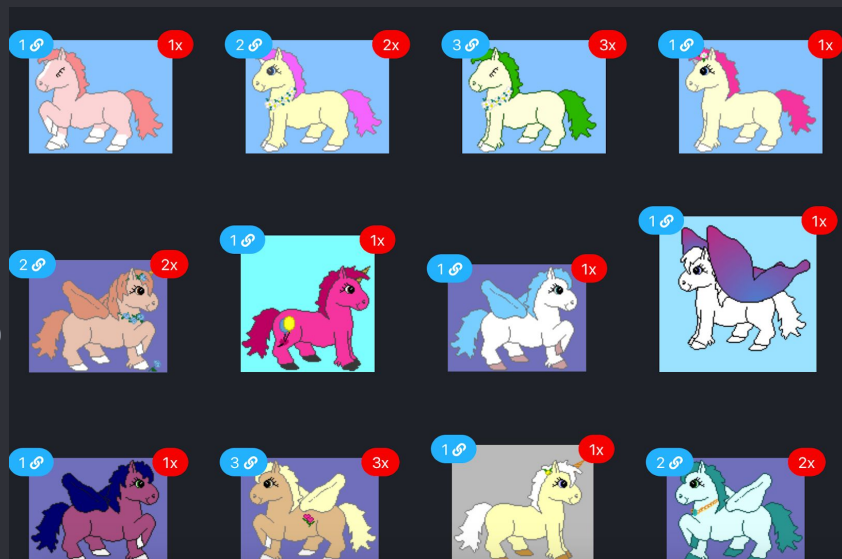
JCDL 2020 | 1-5 August 2020

Content-Based Exploration of Archival Images Using Neural Networks

Tobi Adewoye, Xiao Han, Nick Ruest, Ian Milligan, Samantha Fritz, and Jimmy Lin

@unleasharchives

What's This?

DAIRE (Deep Archival Image Retrieval Engine) is an image exploration tool based on latent representations derived from neural networks, which allows scholars to "query" using an image of interest and rapidly explore similar images within a web archive. This work represents one part of our broader effort to move away from text-centric analyses of web archives and scholarly tools that are direct reflections of methods for accessing the live web.

When the site is loaded, a random "query" image is shown in the top-left corner. Our tool produces a collage of images that are similar in content:

The icon in the top-right corner of each image shows the number of exact copies of the image in the web archive.

The icon in the top-left corner of each image shows how many web pages contain the image. Clicking on the icon brings up a list of those pages, with links to archived copies in the Internet Archive's Wayback Machine. Clicking on any of the images uses it as the query image, and in this manner our tool supports "distant viewing" of web archives.

The images in this demo are drawn from the EnchantedForest neighborhood of GeoCities (from the late 1990s to early 2000s), selected because of strong content moderation imposed at the time. However, be warned that our tool might occasionally surface inappropriate or offensive images.

The code for this demo is open source and can be found on GitHub here.

DEMO: http://daire.cs.uwaterloo.ca