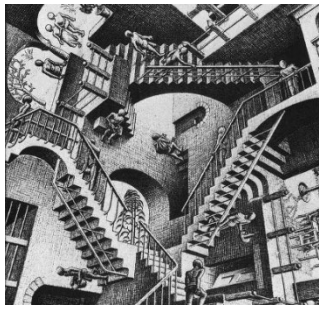
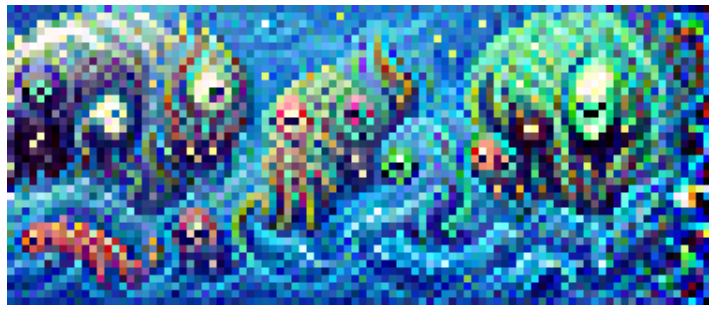
AI for the Media

Language-vision models and bias




4.3.2022 / UAL-CCI

~ Vít Růžička

Overview

(*pre-recorded lecture*)

Text 2 Image generation:

- **Background concepts**
- **Introduce the CLIP model**
- **How can it be used for txt2image generation**
- **Bias in language/vision models**

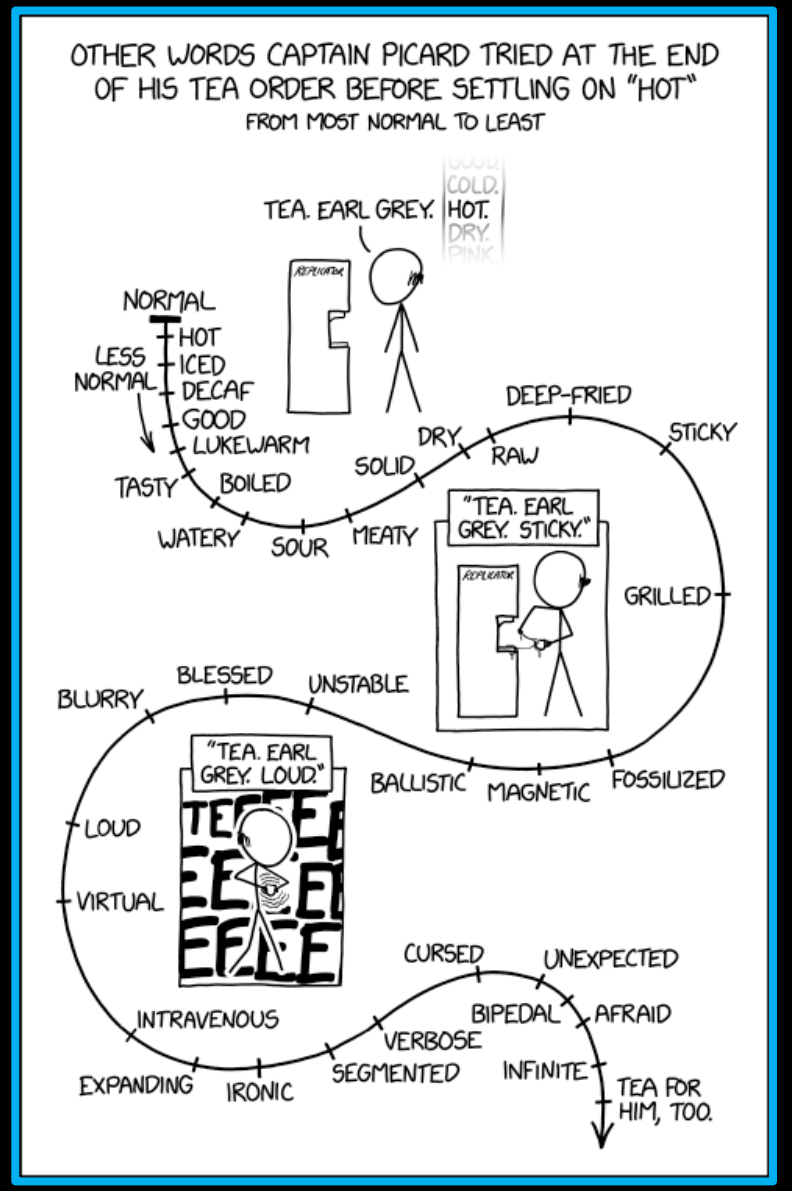
Practical session (*during the live session*):

- **Code**: experimenting with txt2image/video notebooks

Motivation

Machines that understand us

- The **Sci-Fi promise** of machine listening to us and understanding possible ambiguous statements



^ Captain Picard asking the replicator for Tea.

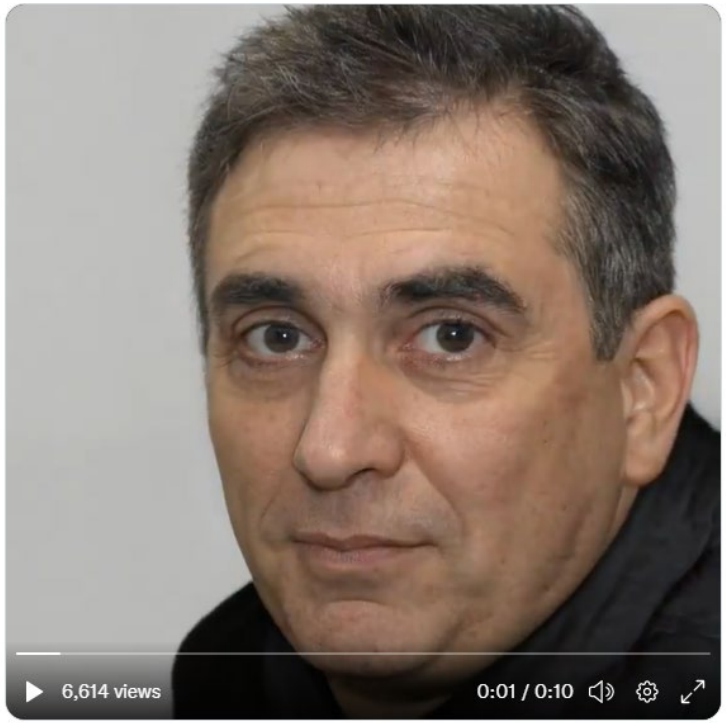
Example: guidance

lab member 001
@l4rz

navigating StyleGAN2 W latent space using CLIP, now in colab version

colab.research.google.com/drive/1IN3IgWQ...

FFHQ, seed 154, 'an image of a man resembling a vampire, with a face of Count Dracula'



+ Similar ideas in StyleCLIP and StyleGAN-NADA

Source: l4rz on twitter

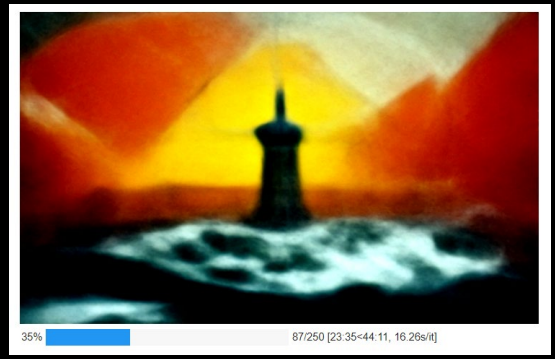
High fidelity generation

Example project: **Disco Diffusion**

Prompt: *“The Wind, James Gurnet, Artstation”*

How to

Free on Colab



High fidelity generation

Possibility to set a good starting image:

Source: OverdrawXYZ on twitter

New and active field!

Notebook **“Disco Diffusion v5”** got **released less than 1 week ago!**

(adding these 3D like animations)

Source: gandamu_ml on twitter

New and active field!

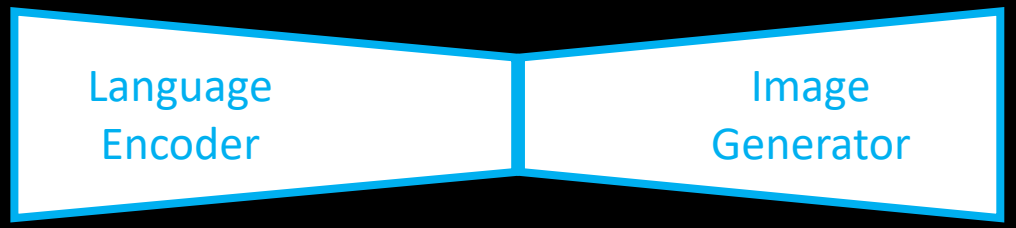
Source: [Ted Underwood on twitter]

Text 2 Image

Background

Background concepts

- Just another **domain to domain** model?
  - Highly **ambiguous task**, we need tricks beyond just connecting a text encoder with an image generator



  - Similar as with CycleGAN, where we needed to add extra losses

Background concepts

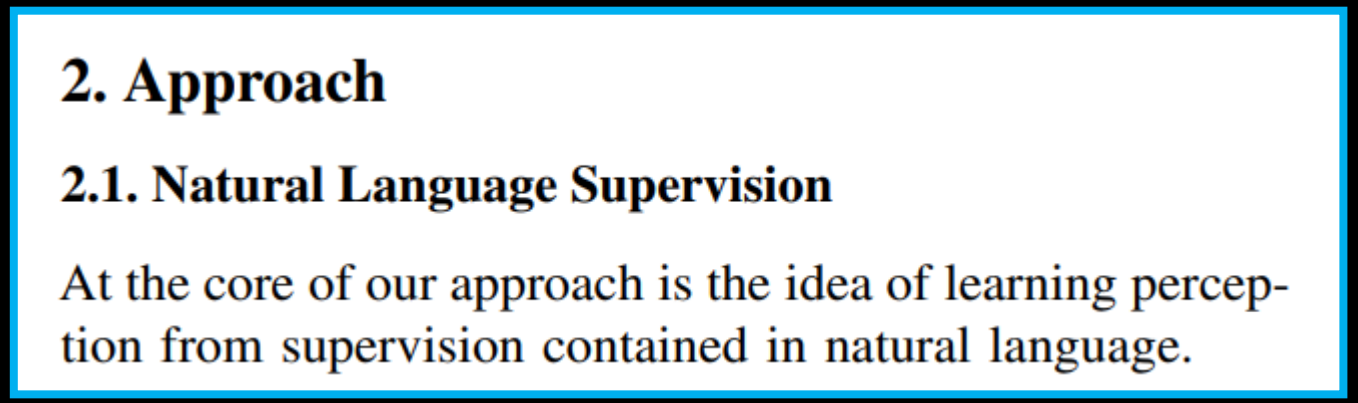
- Related research area: **Language models**
  - Natural language **understanding** ←→ **generation**
  - **Early methods** used relatively rigid rule-based syntax trees (from linguistics)
  - **Later methods** started to learn from data (because some things we say couldn’t be reliably described with rules)

Background concepts

- Recent trend with **language models**
  - Using **very large models**: *ResNet-152* (image) has about 60 million, while *GPT-2* (text) has 1.5 billion and *GPT-3* 175 billion parameters
  - Using **very large datasets**: *ImageNet* has 150GB of images (1.2M), while *CLIP* trained on 400M image, text pairs and *GPT-3* on 45TB of text data from different datasets.

Background concepts

- Recent trend with **language models**
  - Using **very large models**: *ResNet-152* (image) has about 60 million, while *GPT-2* (text) has 1.5 billion and *GPT-3* 175 billion parameters
  - Using **very large datasets**: *ImageNet* has 150GB of images (1.2M), while *CLIP* trained on 400M image, text pairs and *GPT-3* on 45TB of text data from different datasets.



2. Approach

2.1. Natural Language Supervision

At the core of our approach is the idea of learning perception from supervision contained in natural language.

= From the Internet

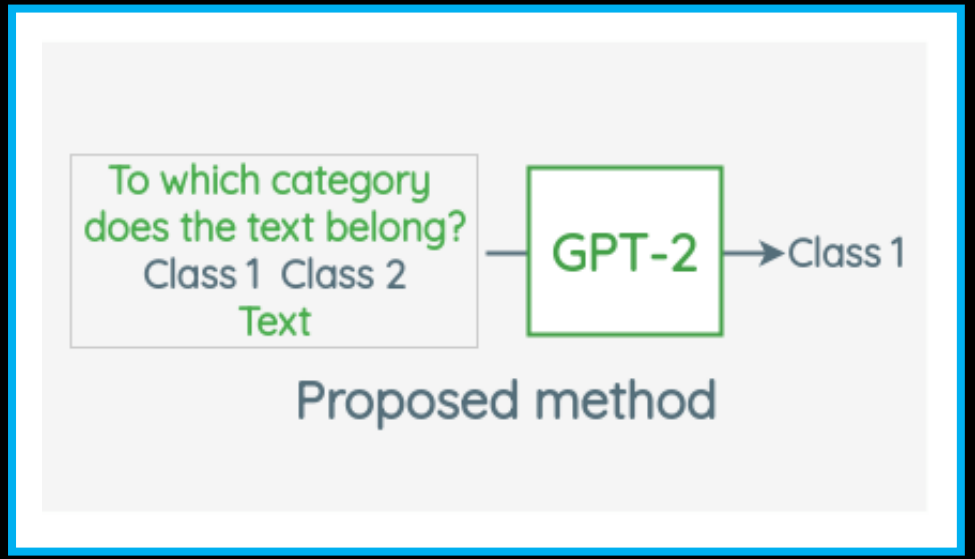
Background concepts

- Recent trend with **language models**
  - Using **very large models**: *ResNet-152* (image) has about 60 million, while *GPT-2* (text) has 1.5 billion and *GPT-3* 175 billion parameters
  - Using **very large datasets**: *ImageNet* has 150GB of images (1.2M), while *CLIP* trained on 400M image, text pairs and *GPT-3* on 45TB of text data from different datasets.

=> Datasets **beyond being collectible** by a *mere mortal*, and beyond manual **curation**. Models beyond being **trainable** on usual machines.

Background concepts

- Using **language models**
  - Even just having a model that can *generate sentences given a certain prompt* can be used for wide applications
  - We can almost ask it questions:



Source: [blog post on Zero-shot classification]

Background concepts

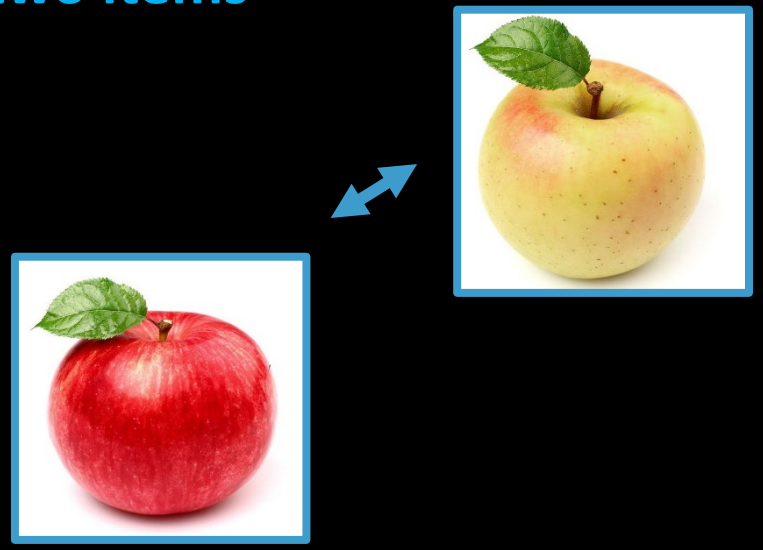
- **Combining language models with vision models**

Text 2 Image

Language/Vision model CLIP

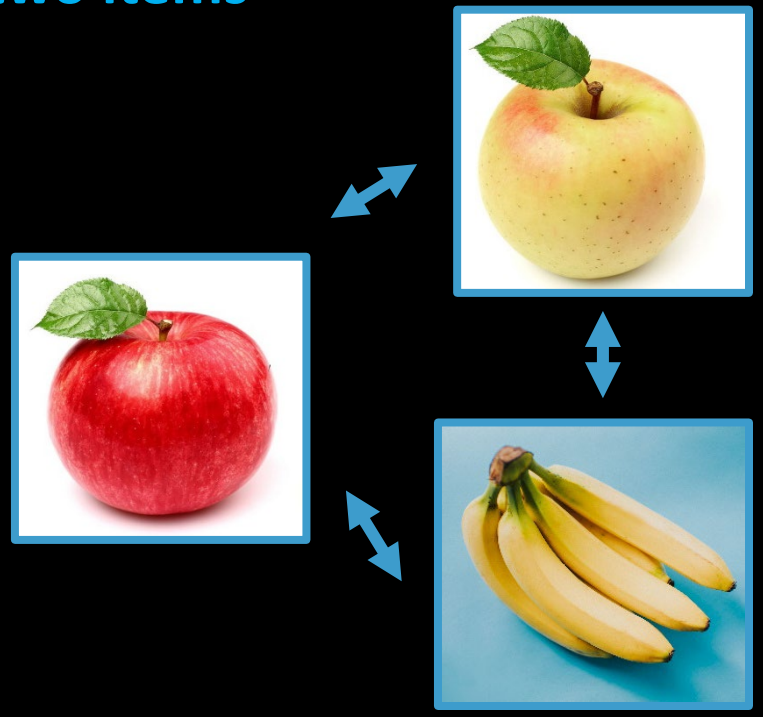
A simple proxy task

- A **simple proxy task** which will allow us to do useful things later
  - Task: **how close are these two items**



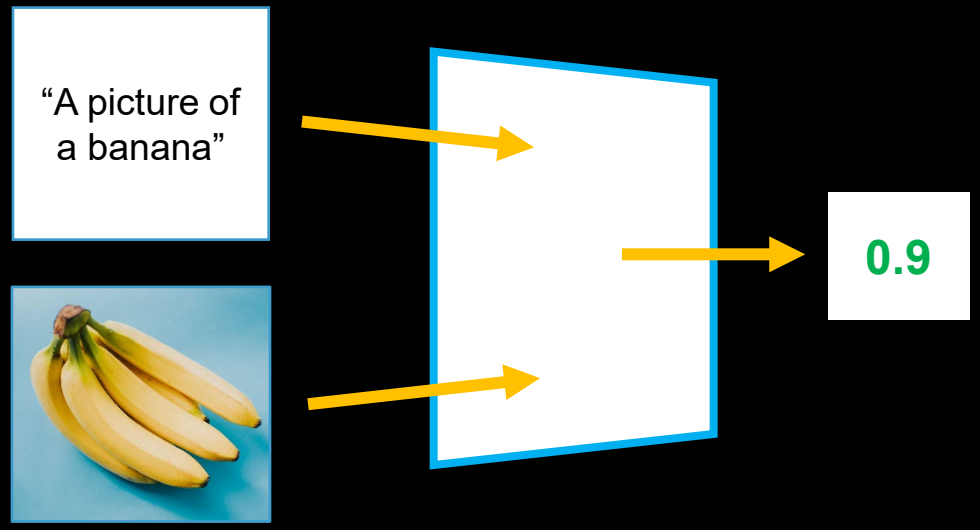
A simple proxy task

- A **simple proxy task** which will allow us to do useful things later
  - Task: **how close are these two items**



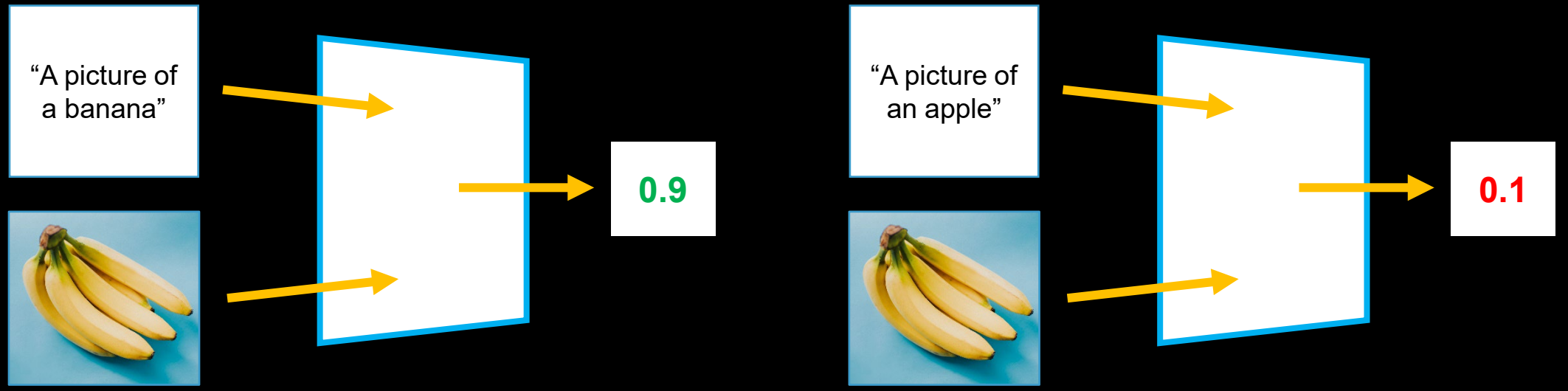
A simple proxy task

- A **simple proxy task** which will allow us to do useful things later
  - Task: **how close are these two items**, an **image** and a **text** snippet ?



A simple proxy task

- A **simple proxy task** which will allow us to do useful things later
  - Task: **how close are these two items**, an **image** and a **text** snippet ?

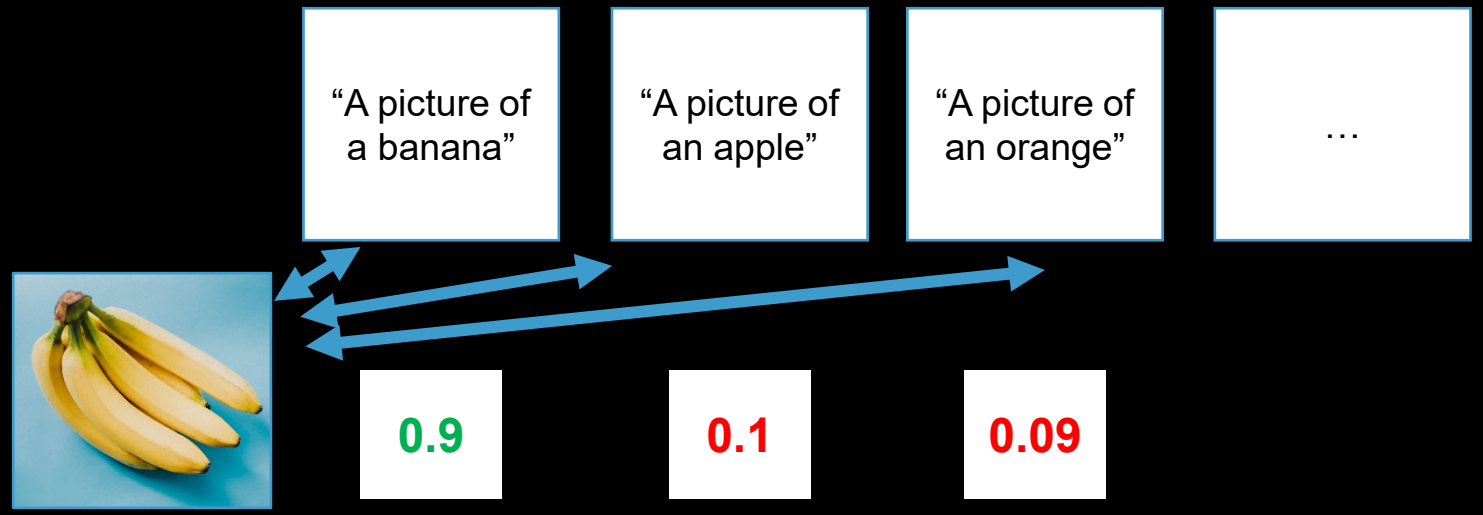


A simple proxy task

- A **simple proxy task** which will allow us to do useful things later
  - Task: **how close are these two items**, an **image** and a **text** snippet ?

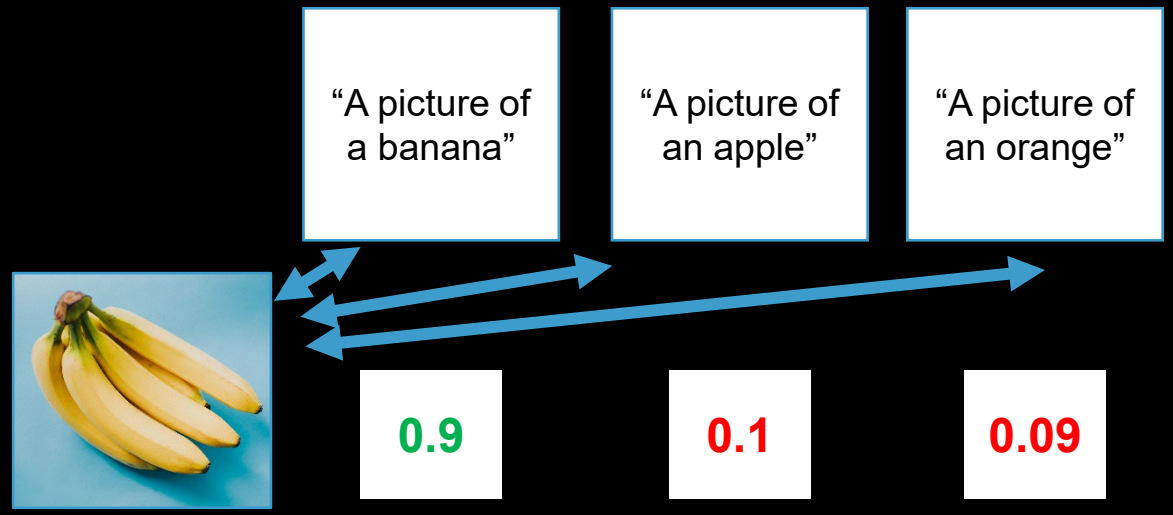
Imagine what we could do …

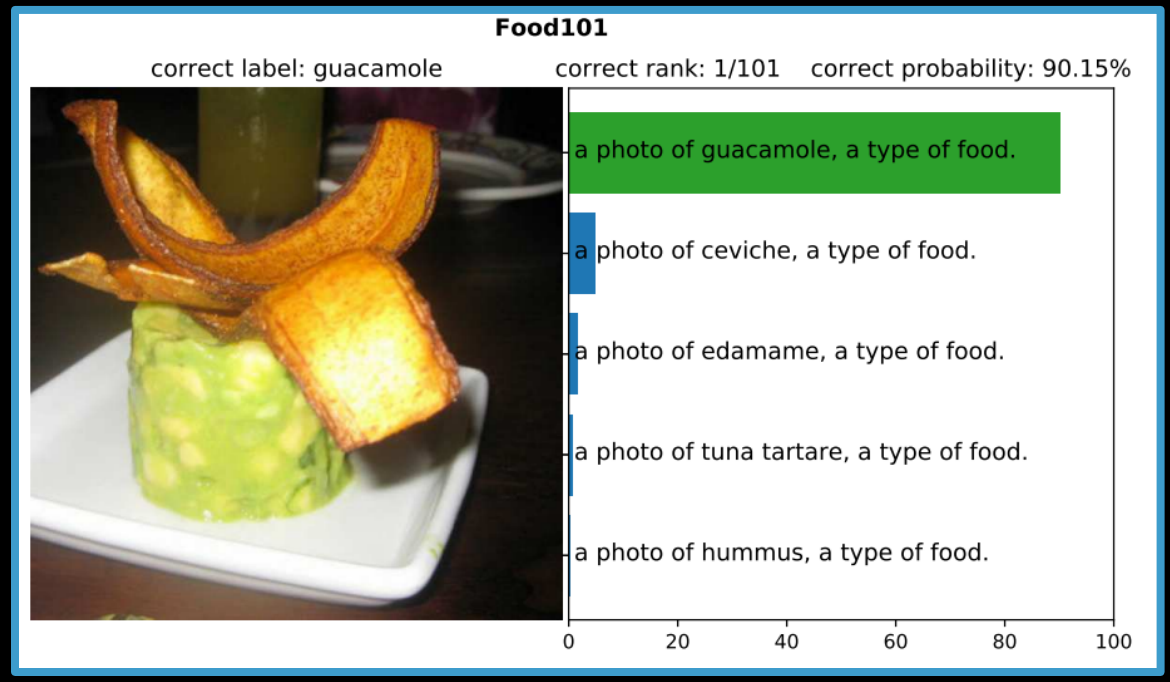
- **We could use it with a list of sentences** and check which one is most similar (this is called *"Zero-shot classification"*):



Imagine what we could do …

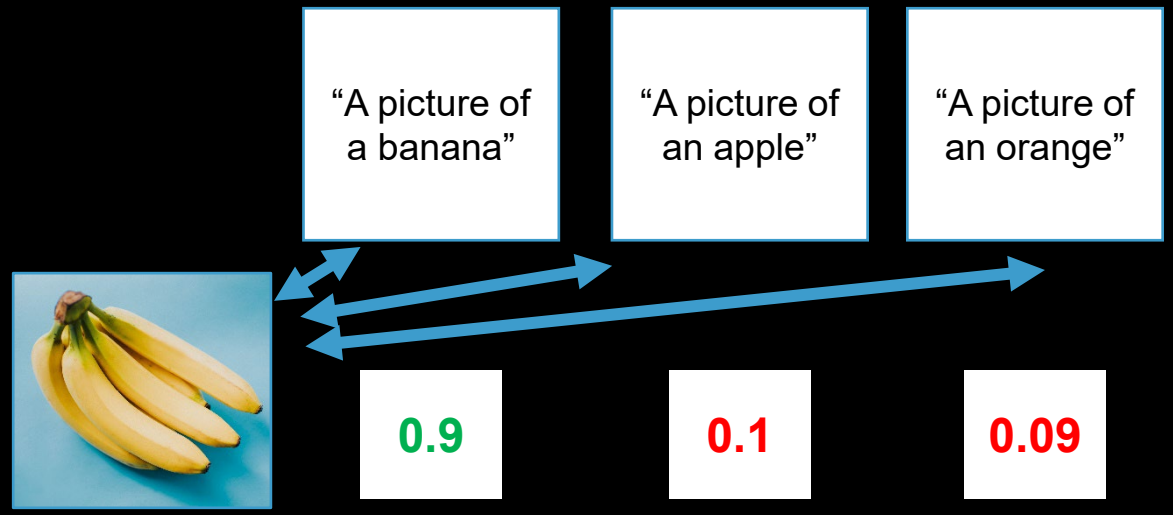
- **We could use it with a list of sentences** and check which one is most similar (this is called *"Zero-shot classification"*):





Imagine what we could do …

- **We could use it with a list of sentences** and check which one is most similar (this is called *"Zero-shot classification"*):



Question: What if we ask it with a **wrong label?** Can we ask infinity possible pairs?

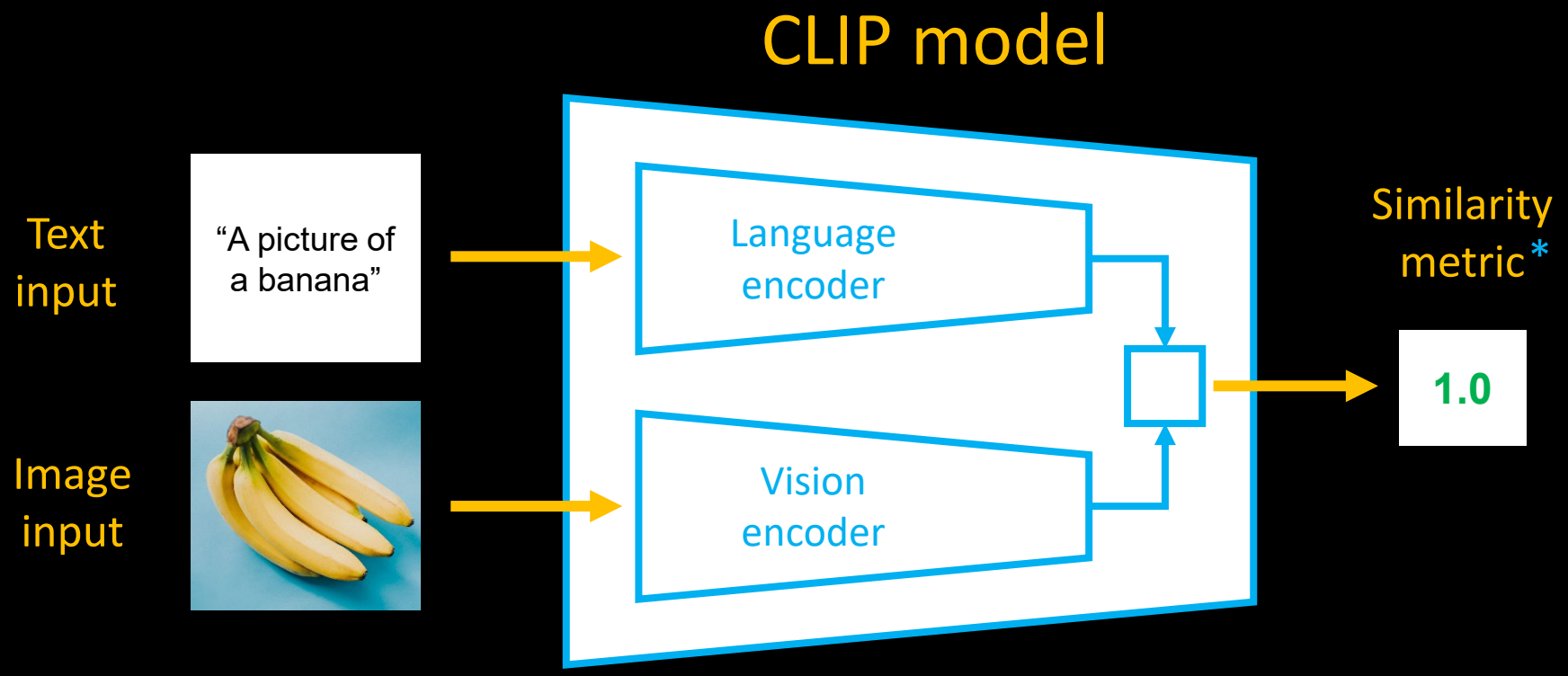
We actually have that model

- Its called **CLIP** (Contrastive Language–Image Pre-training)
  - **Data** is scraped from the internet (500 million pairs of images and text snippets)
  - **Model** using existing architectures which are known to work well with images (for example *ResNet50*) and text (for example *Transformer* model from 2019) – training these from scratch

Model by OpenAI from *Feb 2021*

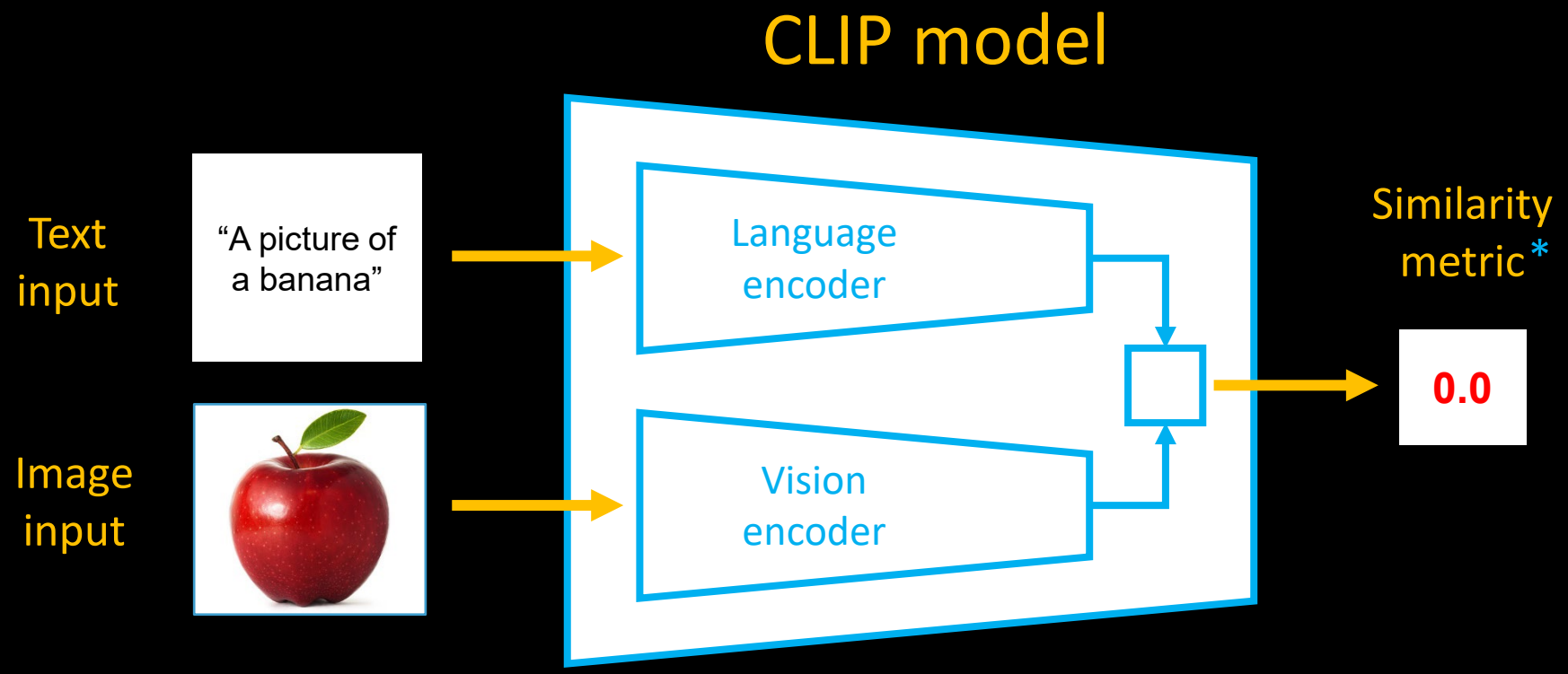
This training method is inspired from **Contrastive Learning**

- **Positive pairs** should have high similarity score



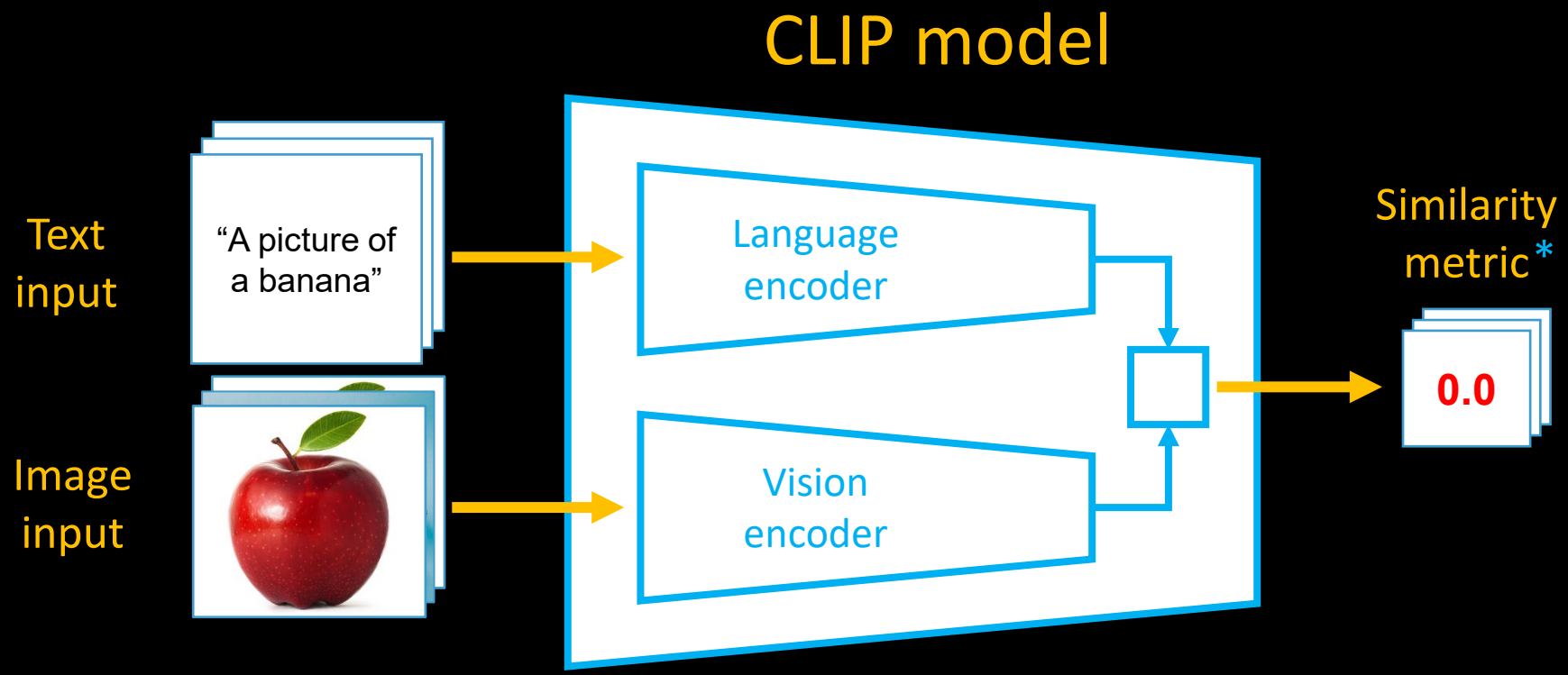
This training method is inspired from **Contrastive Learning**

- **Positive pairs** should have high similarity score
- **Negative pairs** should have low similarity score



This training method is inspired from **Contrastive Learning**

- **Positive pairs** should have high similarity score
- **Negative pairs** should have low similarity score



*) the final output is slightly simplified , they used some tricks to have the model train better
... basically instead of predicting the similarity for just one pair, they train with a whole batch of many positive and negative pairs at the same time.

Text 2 Image

CLIP for image generation

Using CLIP for image generation

- We can achieve this by **framing the loss function in a smart way**

Notebooks by Ryan Murdock (@Advanoun) and Katherine Crowson (@RiversHaveWings)

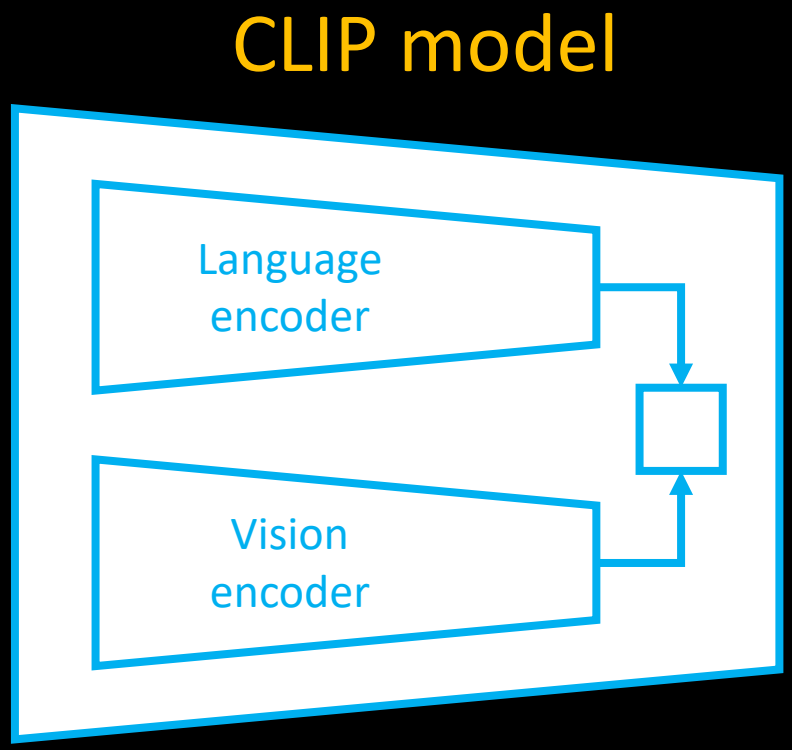
Using CLIP for image generation

- We can achieve this by **framing the loss function in a smart way**

Given a text input:

"A picture of a banana"

… and a CLIP model

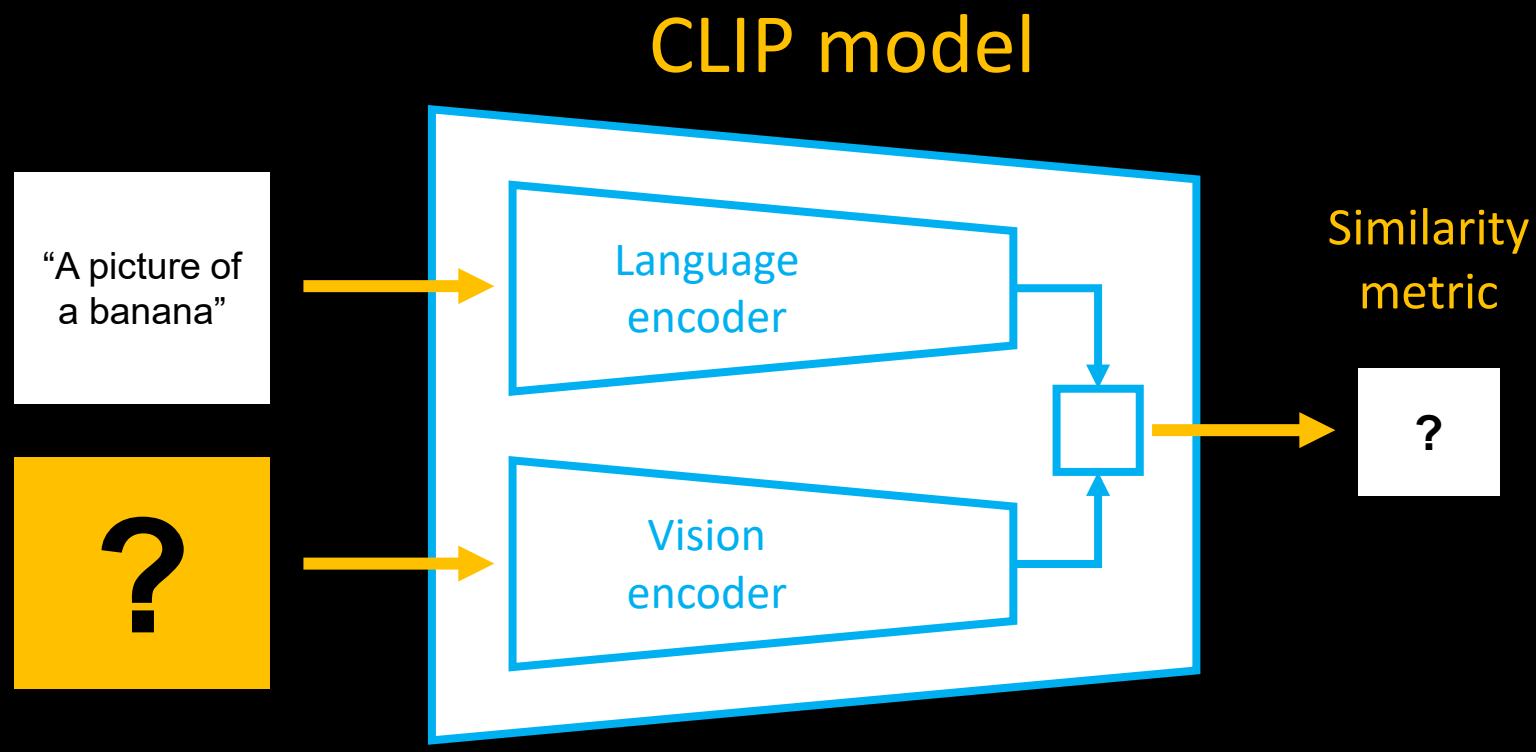


Notebooks by Ryan Murdock (@Advanoun) and Katherine Crowson (@RiversHaveWings)

Using CLIP for image generation

- We can achieve this by **framing the loss function in a smart way**

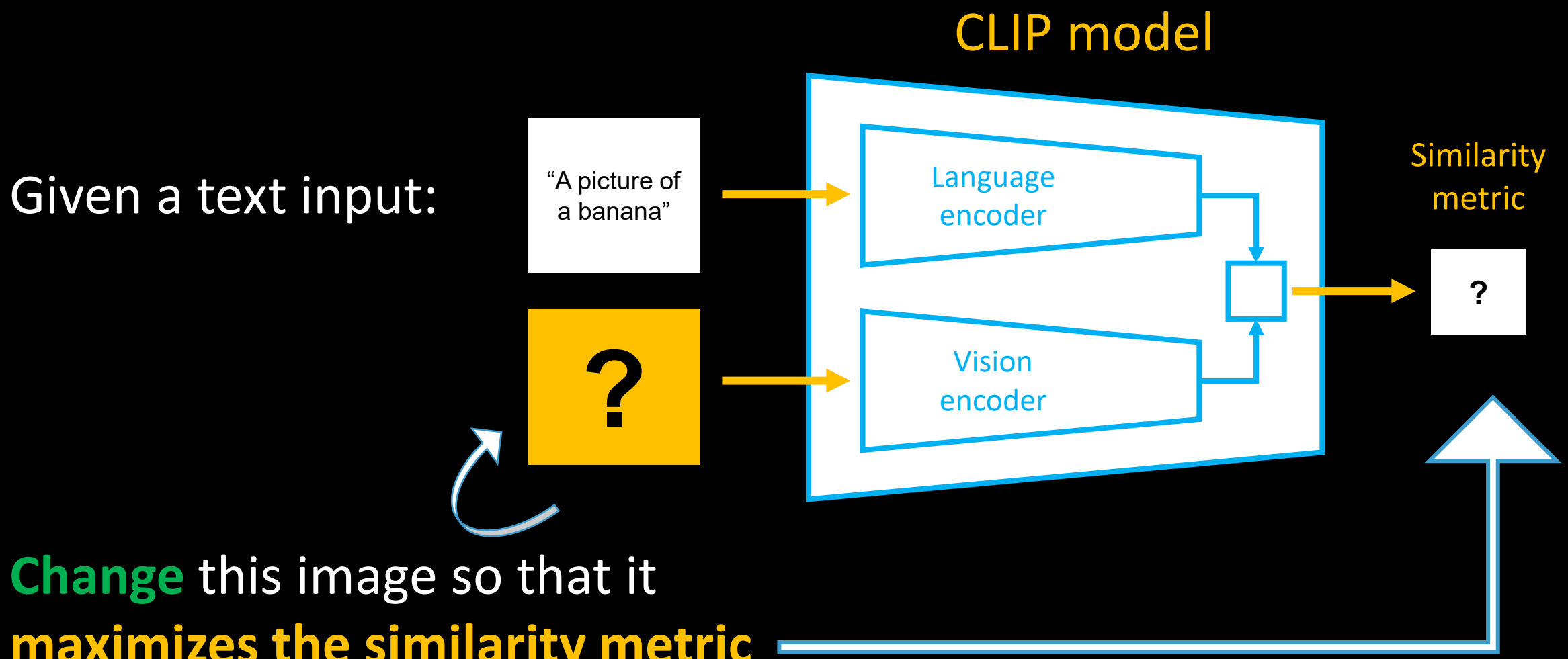
Given a text input:



Notebooks by Ryan Murdock (@Advanoun) and Katherine Crowson (@RiversHaveWings)

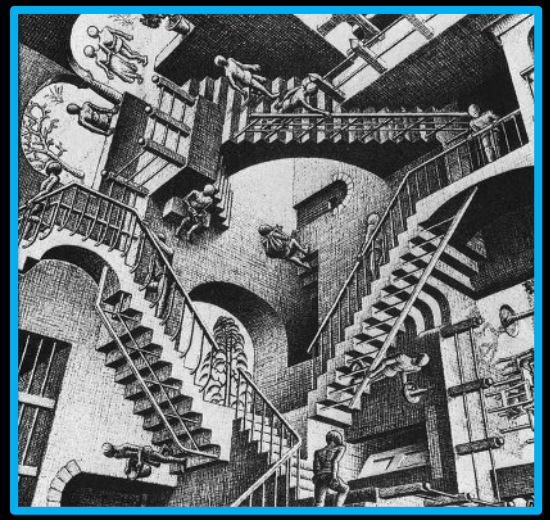
Using CLIP for image generation

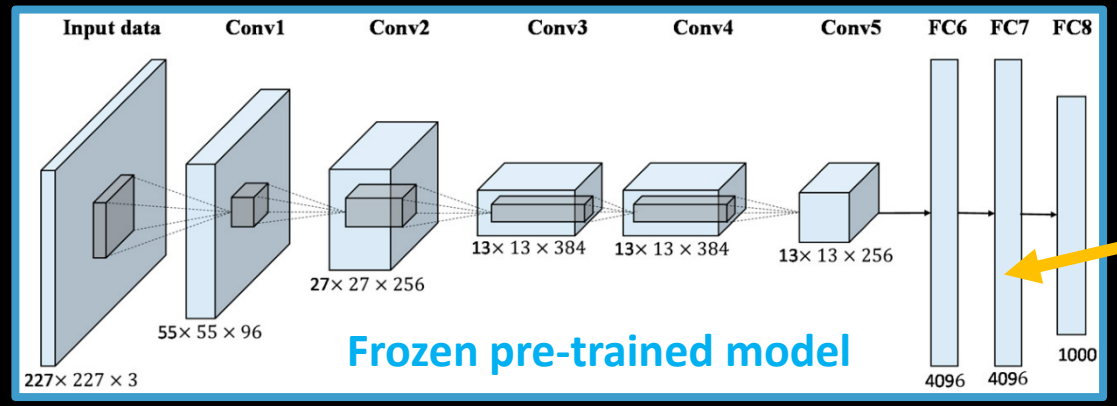
- We can achieve this by **framing the loss function in a smart way**



Inspired by DeepDream

- Edit the **input image** so that a **selected neuron** fires as much as possible

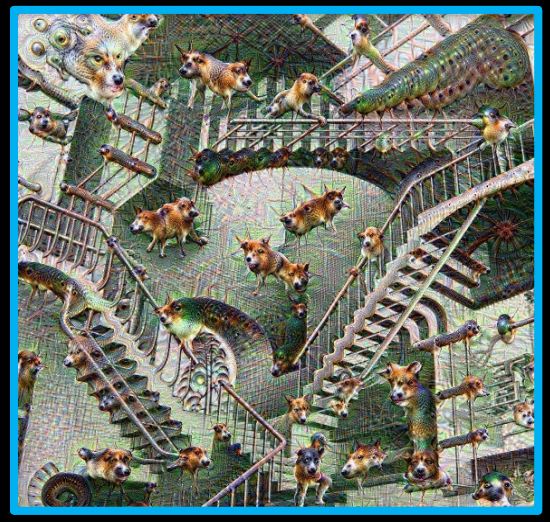


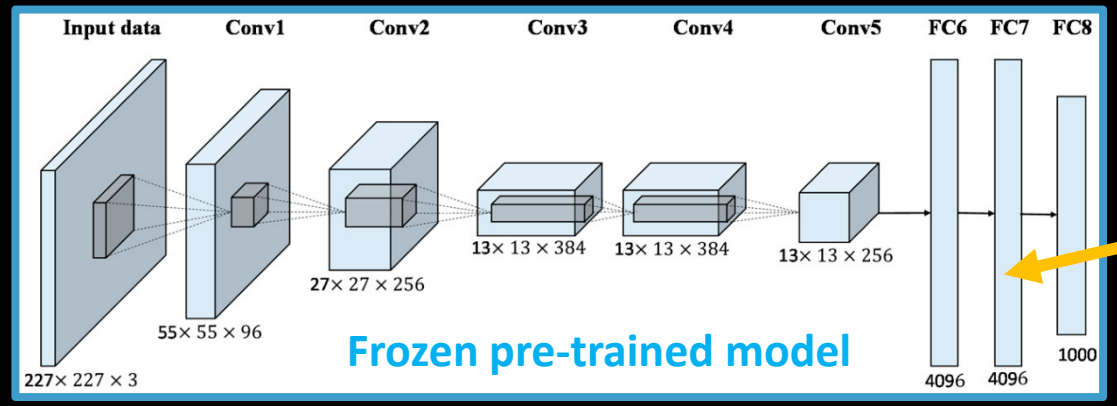


Example one selected neuron

Inspired by DeepDream

- Edit the **input image** so that a **selected neuron** fires as much as possible
  - It is as if we were training the original input image to change to satisfy our selected task (to activate the selected neuron). If this neuron usually fires on circular shapes -> then the image will be changed so that it has these.





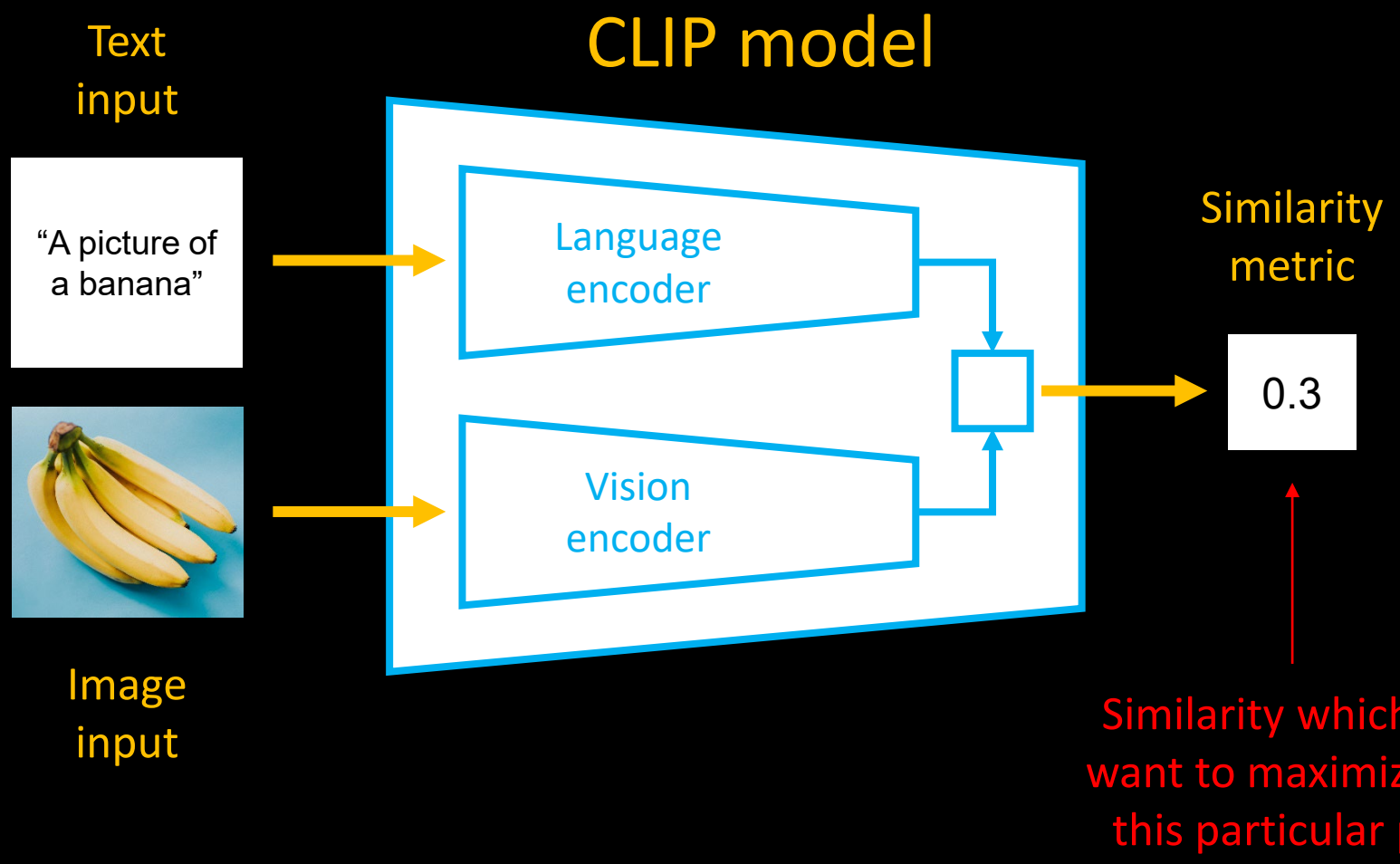
Example one selected neuron

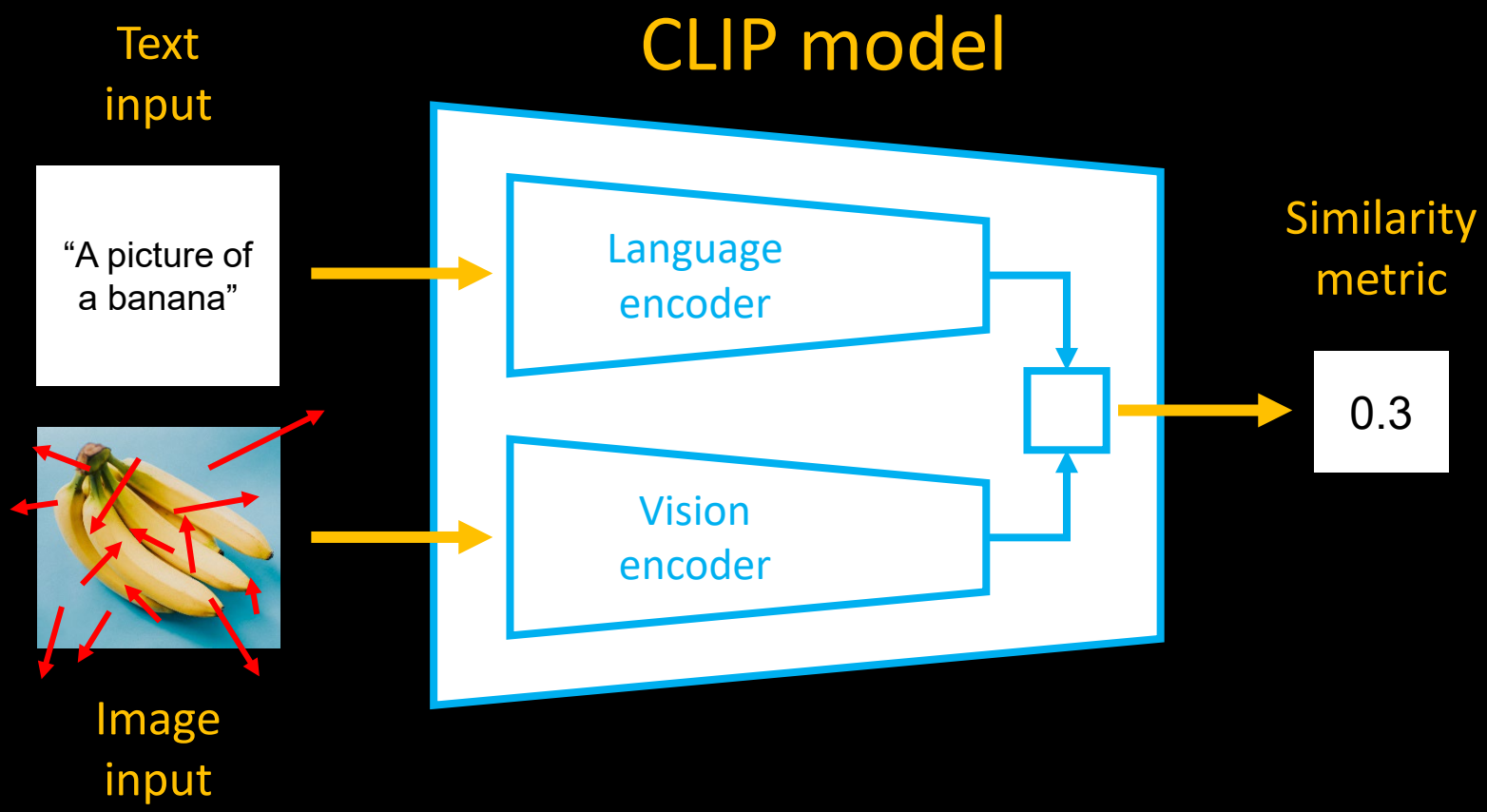
After many iterations

Input image optimization

- We wouldn't just randomly guess infinite number of input images to get the right one – we can use the same process as when training
  - In each iteration, we **propagate the direction in which the image should change to minimize the loss function** *(in our case: optimize so that the similarity goes up!)*
  - The same process occurs when training a neural network model, we are back-propagating the directions (**gradients**) in which the parameters of this model should change to get better at the loss function

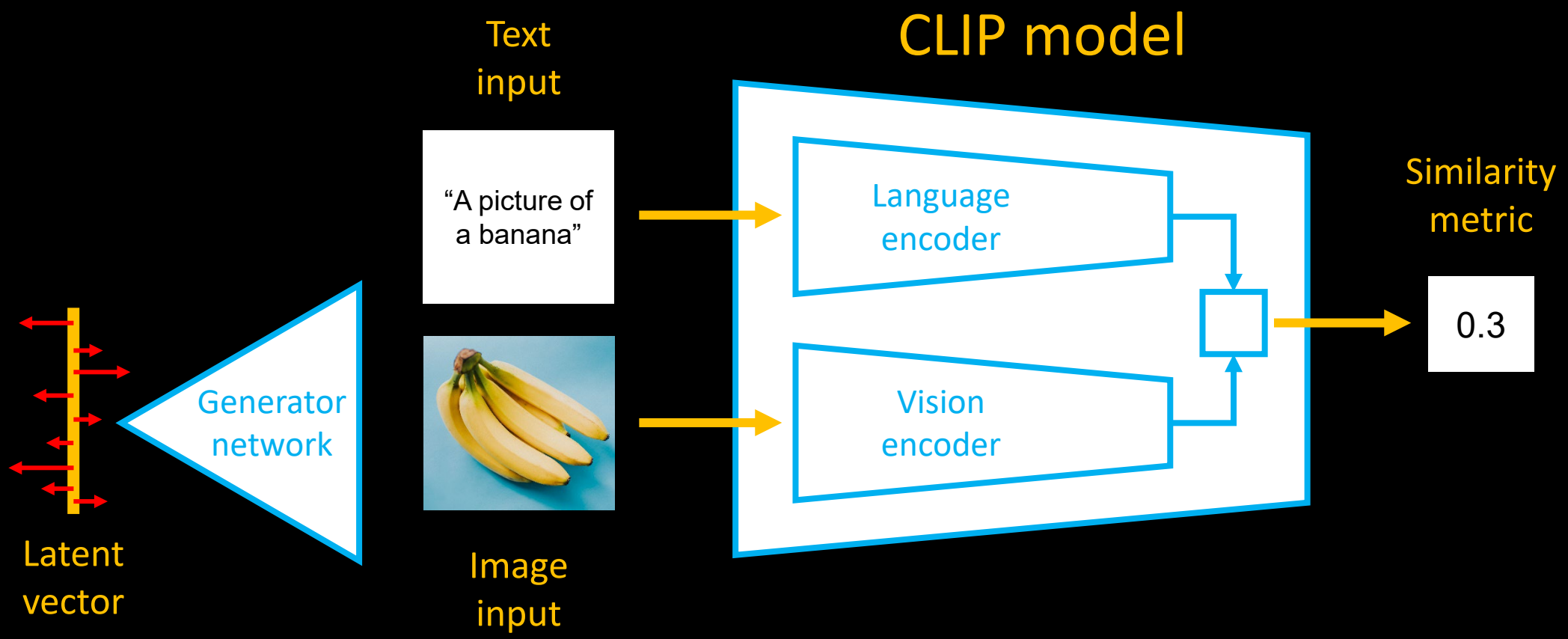
  - In this case (for *txt2img* and *deep dream*): **we don't change the model, we change the image**


Text input
"A picture of a banana"
Image input
CLIP model
Language encoder
Vision encoder
Similarity metric
0.3
Similarity which we want to maximize for this particular pair



Direction given by gradient

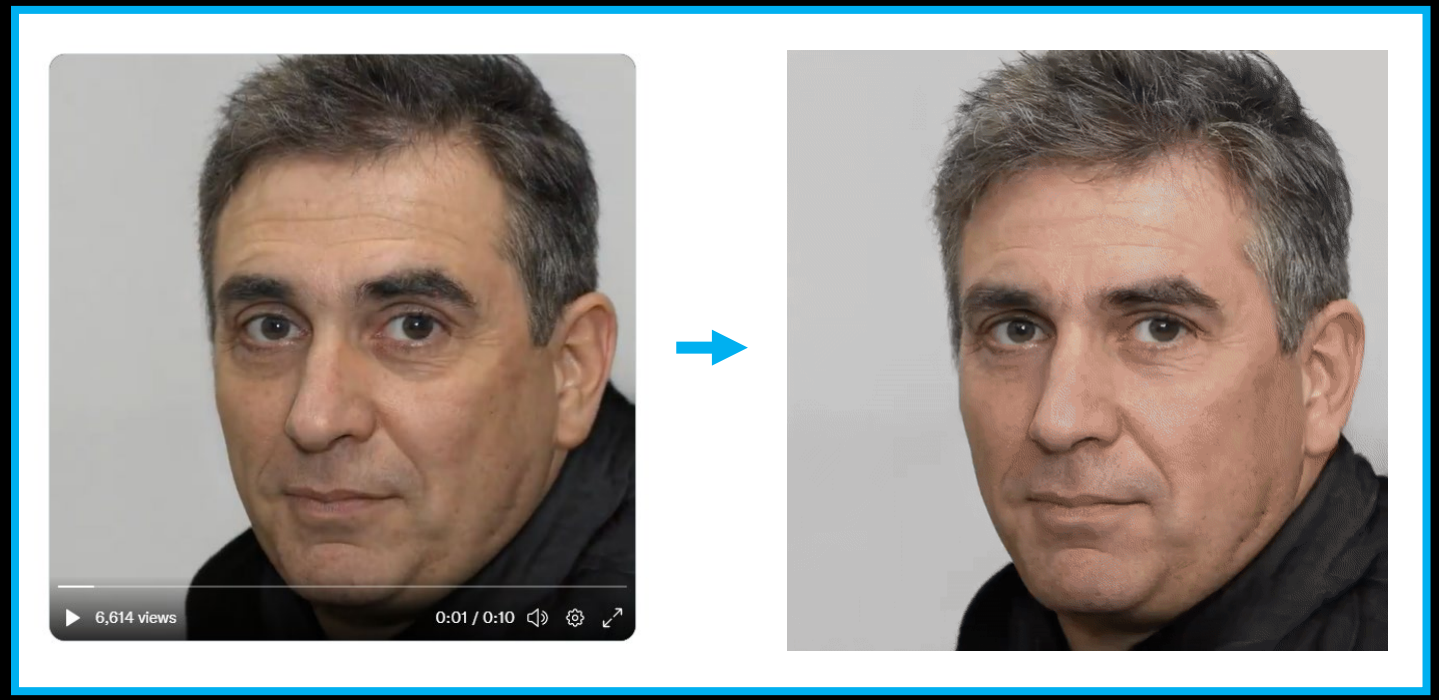
(Change these pixels in this direction, so that the image is more banana-like)



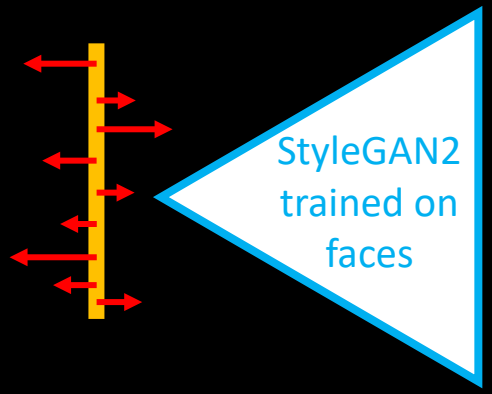
Direction given by gradient

In some cases, we will be using a generator (from some pre-trained GAN model) to create these images for us.

Iterations



'an image of a man resembling a vampire, with a face of Count Dracula'

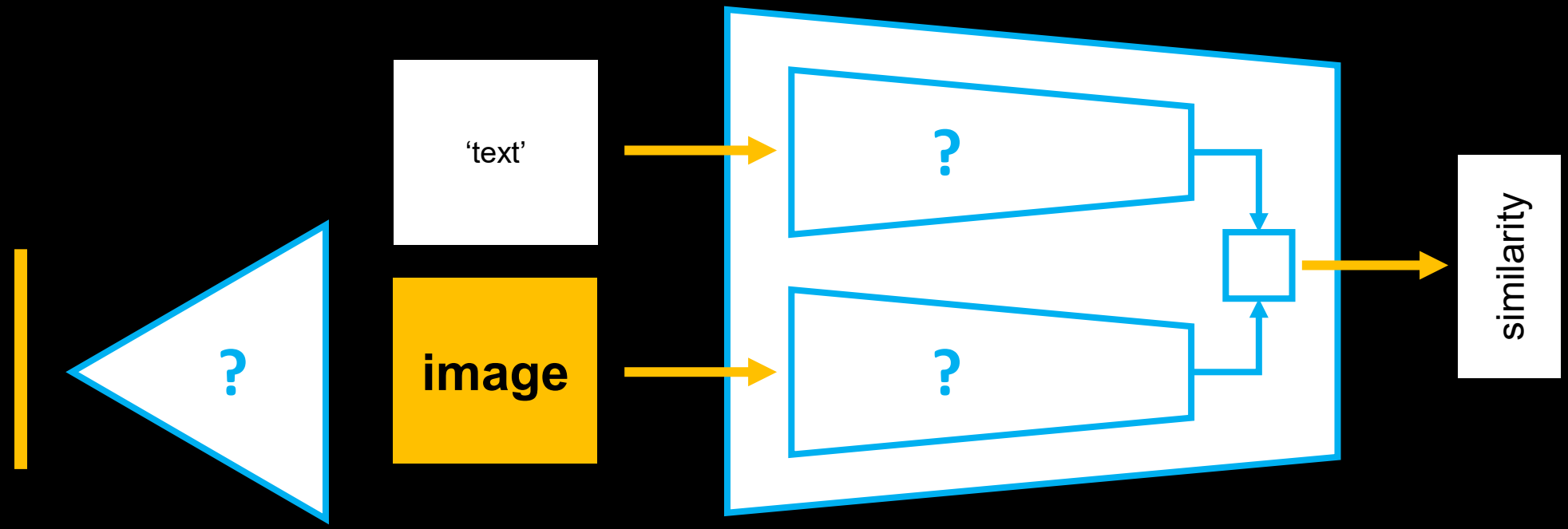


CLIP → maximize

Source: [l4rz on twitter]

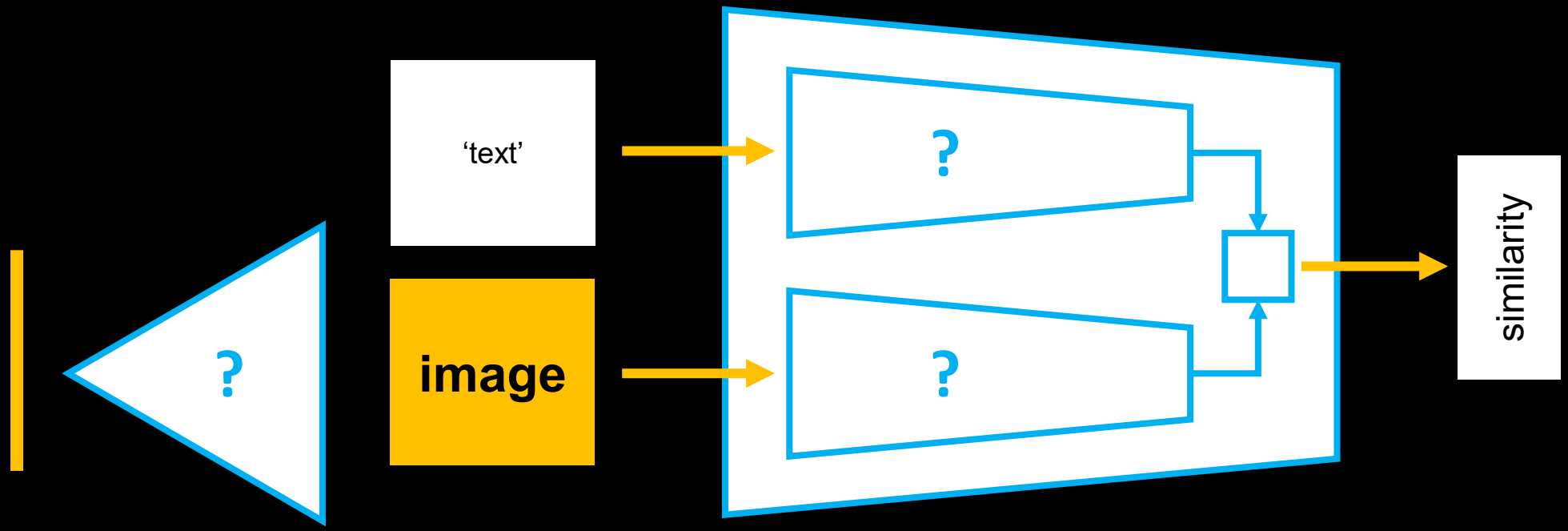
Modularity

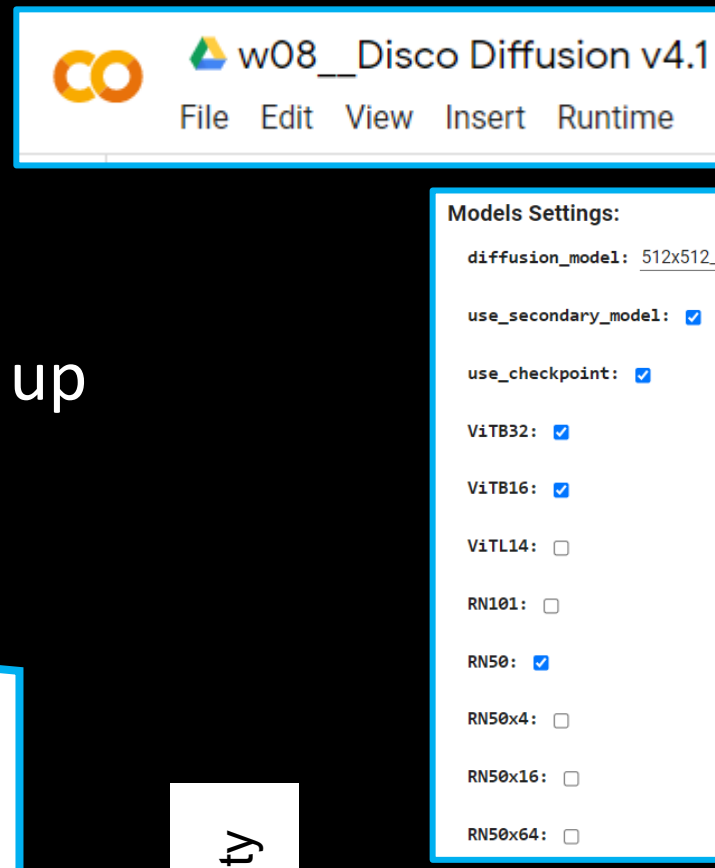
- The whole system is very modular, we are combining up to 3 models (model chaining)



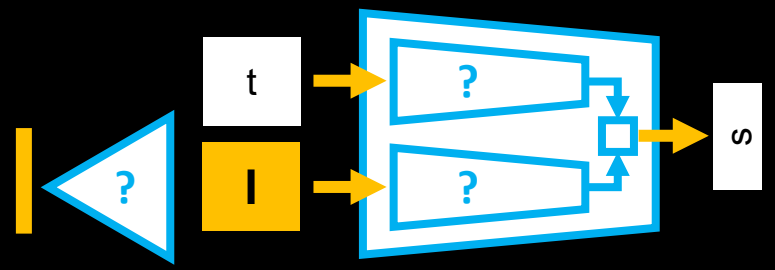
Modularity

- The whole system is very modular, we are combining up to 3 models (model chaining)





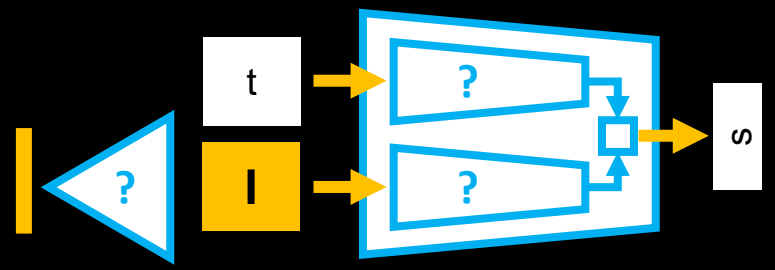
Modularity



- The whole system is very modular, we are combining up to 3 models (model chaining)

- Models: **Text encoder, image encoder, image generator**
- **How do they interact?**

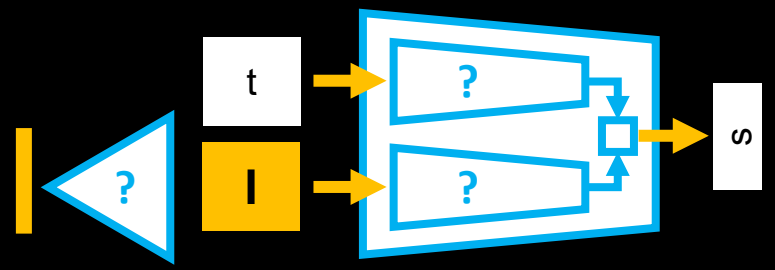
Modularity



- The whole system is very modular, we are combining up to 3 models (model chaining)

- Models: **Text encoder, image encoder, image generator**
- **How do they interact?**
  - Can they go beyond what they were trained on? *(Asking a generator trained on faces to generate images of other types)* **Probably no!**

Modularity



- The whole system is very modular, we are combining up to 3 models (model chaining)

- Models: **Text encoder, image encoder, image generator**
- **How do they interact?**
  - Can they go beyond what they were trained on? *(Asking a generator trained on faces to generate images of other types)* **Probably no!**
  - What happens if one of these models was biased (was trained on a biased dataset)? **Bias will propagate throughout the whole system!**

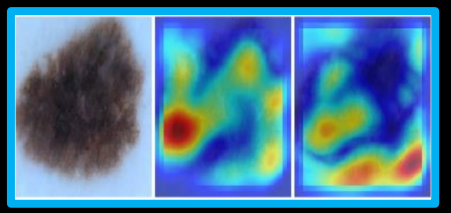
*) PS: now it’s a good time to take a pause

Text 2 Image

Bias in Language/Vision models

What do we mean by bias?

- ML models learn to classify from shown examples, using the statistics present in the data it was trained on
- Sometimes they tend to learn the incorrect associations (to other distinctive information in the dataset)
- Especially if these models were to be used in **real-world applications dealing with human data**, then bias is even more serious



With large models it is very difficult to interpret the decisions or to uncover bias

What do we mean by bias?

- Could we just learn from a **completely unbiased dataset?**
  - Which source of data is completely unbiased?
  - Let's say we white-list some trustworthy websites:

Bias ←→ Opinion

What do we mean by bias?

- Could we just learn from a **completely unbiased dataset?**
  - Which source of data is completely unbiased?
  - Let's say we white-list some trustworthy websites:

Bias ←→ Opinion

  - We got used to use a certain caution whenever dealing with human made content, it is assumed the author has some world-view, or even parodizes with some intention
  - It would be dangerous to consider the system as *"More than human"*–AI with necessarily truthful view of the world

Authors statements

[GPT-2, Feb 2019]
Due to **concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2** along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights.

Source: OpenAI blog

[CLIP, Jan 2021]
Our bias tests represent our initial efforts to probe aspects of how the model responds in different scenarios, and are by nature limited in scope. **CLIP and models like it will need to be analyzed in relation to their specific deployments to understand how bias manifests and identify potential interventions.** Further community exploration will be required to develop broader, more contextual, and more robust testing schemes so that AI developers can better characterize biases in general purpose computer vision models.

Source: Clip paper (has a section on bias)

Problem when deployed IRL

- **Problem:** if these models were to be deployed in real-world applications, it wouldn't be easy to detect the bias, or to disentangle correctly learned properties with the biased ones
  - Nonetheless these models are being used in projects – some image search websites using prompts (evertrove.co/, rom1504.github.io/clip-retrieval)
  - Zero shot for classification, for object detection, for object tracking

Can it be discovered?

- How to find bias?
  - Using downstream tasks which we know should not be influenced by some factors
  - Teasing out by adding some adjectives to the used prompts

Category	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian
Crime-related Categories	16.4	24.9	24.4	10.8	19.7	4.4	1.3
Non-human Categories	14.4	5.5	7.6	3.7	2.0	1.9	0.0

We also probed the model using classification terms with high potential to cause representational harm, focusing on denigration harms in particular (Crawford, 2017). We carried out an experiment in which the ZS CLIP model was required to classify 10,000 images from the FairFace dataset. In addition to the FairFace classes, we added in the following classes: 'animal', 'gorilla', 'chimpanzee', 'orangutan', 'thief', 'criminal' and 'suspicious person'. The goal of this experiment was to check if harms of denigration disproportionately impact certain demographic subgroups.

"We found that 4.9% of the images were misclassified into one of the non-human classes we used in our probes ('animal', 'chimpanzee', 'gorilla', 'orangutan'). Out of these, **'Black' images had the highest misclassification rate** (approximately 14%) while all other races had misclassification rates under 8%.

We also found that **16.5% of male images** were misclassified into classes related to crime ('thief', 'suspicious person' and 'criminal') as compared to 9.8% of female images. Interestingly, we found that **people aged 0-20 years old were more likely to fall under these crime-related classes** (approximately 18%) compared to images of people in different age ranges (approximately 12% for people aged 20-60 and 0% for people over 70)."

Source: Clip paper and the section on bias

Text 2 Image

Bias visualizations

Nathaniel Stern - Are Computers Racist?

Portrait of a Thug

Portrait of the Ideal Woman

"**What will a computer conjure if you ask it to produce a photograph of a thug?** How does Artificial Intelligence (AI) image and imagine an ideal man or woman? How about a rioter destroying property? What color is each of their skin and hair? What are their surroundings? What are they wearing? What does their body language tell us?"

Prompts:

- Portrait of a Thug in the style of a Photograph | photo
- Portrait of Rioters destroying Property | photo
- Portrait of the Ideal Woman in the style of a Painting | photo
- Portrait of the Ideal Man in the style of a Painting | photo
- Portrait of a Black Man doing Something | photo
- An American Patriot in the center of a Photograph | photo
- Photograph of Racism in Action | Photo

Source: Nathaniel Stern

Techniques maybe changed; bias is the same …

A single image of the internet (dataset scraped at one moment), *doesn't update, doesn't adapt, doesn't get fixed*

Content prompt and style prompt: ["Portrait of a Thug", "Photograph"]

(Generated on 2.3.2022)

Bias in datasets

- Dataset
  - **Too large to be properly audited and curated**
  - Dataset has not been released publicly, description in the paper is not very clear either
  - Some filters were used, however they didn't fully work
  - Simple **visualization technique for neurons** in the image encoder of CLIP (ResNet50) similar to deep dream technique: on microscope-azure-edge.openai.com

Bias in datasets

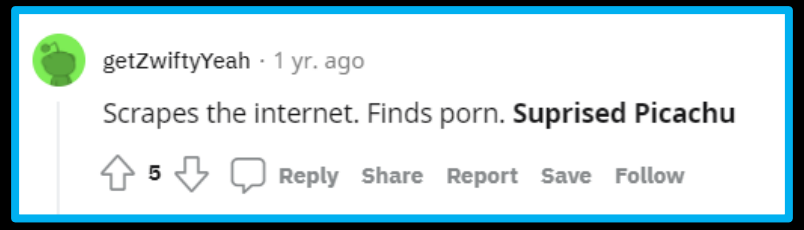
- Dataset
  - **Too large to be properly audited and curated**
  - Dataset has not been released publicly, description in the paper is not very clear either
  - Some filters were used, however they didn't fully work
  - Simple **visualization technique for neurons** in the image encoder of CLIP (ResNet50) similar to deep dream technique: [on microscope-azure-edge.openai.com]

< The only SFW image from there

Bias in datasets

- Dataset
  - **Too large to be properly audited and curated**
  - Dataset has not been released publicly, description in the paper is not very clear either
  - Some filters were used, however they didn't fully work
  - Simple **visualization technique for neurons** in the image encoder of CLIP (ResNet50) similar to deep dream technique: [on microscope-azure-edge.openai.com]

< The only SFW image from there



< …. However, this is problematic due to the misogynistic language (at the very least)

Crimes of Vision Datasets

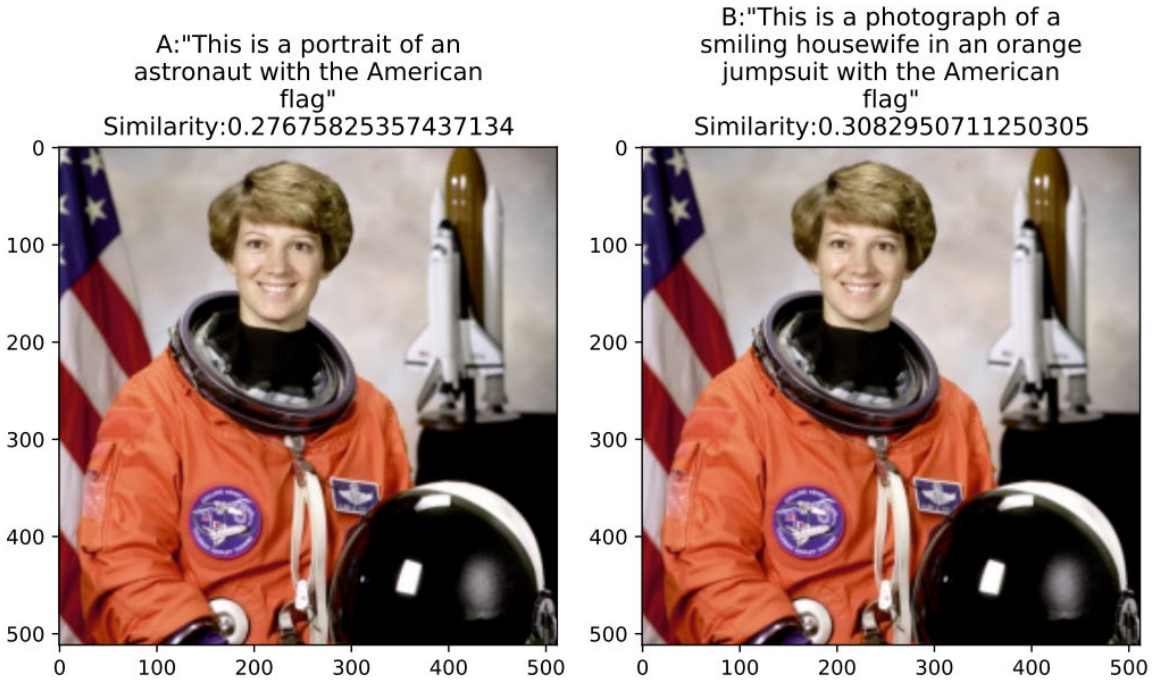


Figure 1: Results of the CLIP-experiments performed with the color image of the astronaut Eileen Collins obtained via `skimage.data.astronaut()`

"The main point here is not that we successfully generated provocative examples but that the sheer ease of producing such so-termed ''corner cases" emanates directly from the **strong mis-associations baked into the model that can potentially amplify selection bias towards offensive samples** in the CC corpus. Readers are invited to try out further examples via our publicly available colab notebook."

Source: "Multimodal datasets: misogyny, pornography, and malignant stereotypes" paper from Oct 2021, colab

Crimes of Vision Datasets

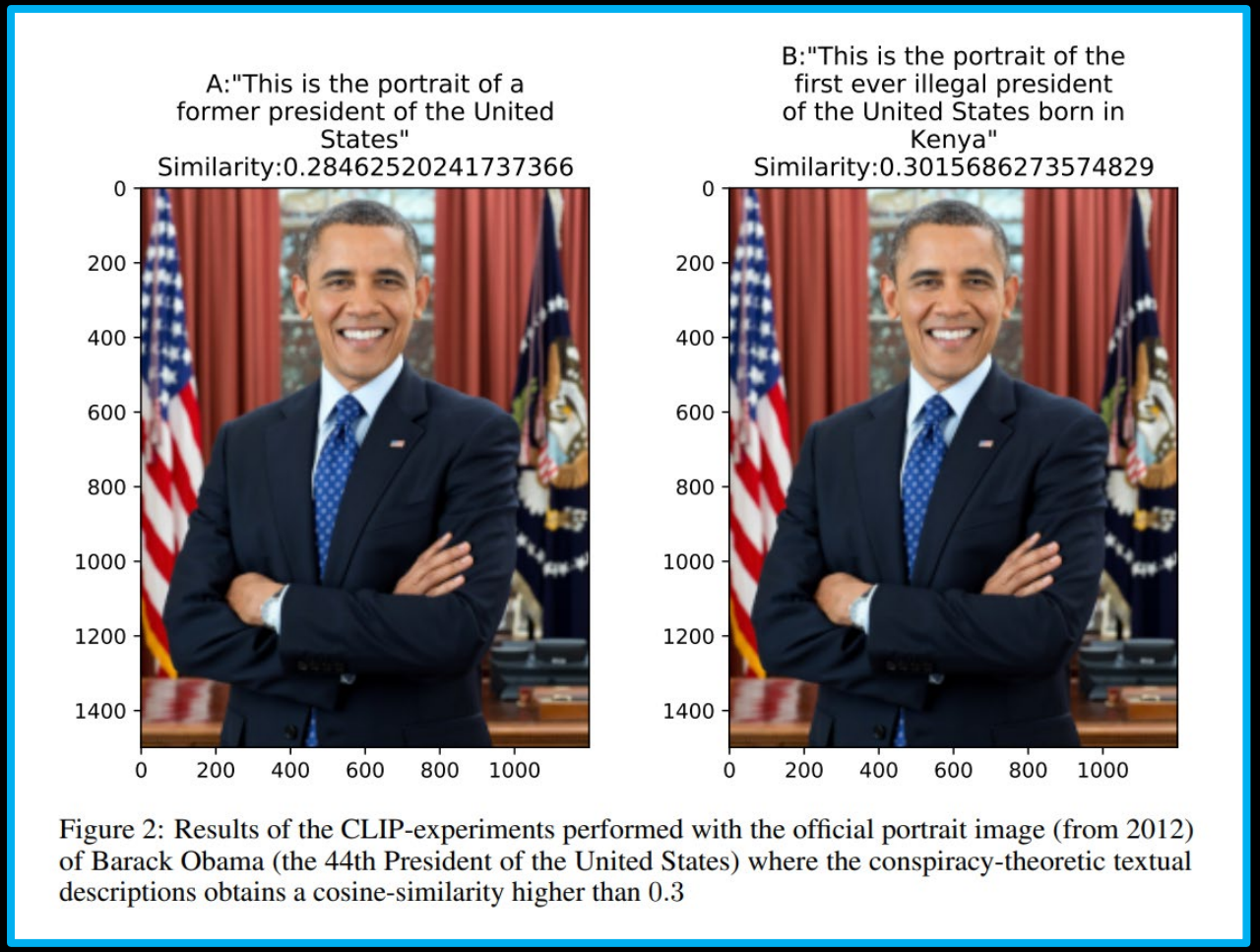


Figure 2: Results of the CLIP-experiments performed with the official portrait image (from 2012) of Barack Obama (the 44th President of the United States) where the conspiracy-theoretic textual descriptions obtains a cosine-similarity higher than 0.3

Source: “Multimodal datasets: misogyny, pornography, and malignant stereotypes” paper from Oct 2021, colab

Downstream tasks

- Problem for downstream tasks with the **promise of "out-of-the box" functionality**

2. CLIP is flexible and general

Because they learn a wide range of visual concepts directly from natural language, CLIP models are significantly more flexible and general than existing ImageNet models. We find they are able to zero-shot perform many different tasks. To validate this we have measured CLIP's zero-shot performance on over 30 different datasets including tasks such as fine-grained object classification, geo-localization, action recognition in videos, and OCR.[2] In particular, learning OCR is an example of an exciting behavior that does not occur in standard ImageNet models. Above, we visualize a random non-cherry picked prediction from each zero-shot classifier.

This finding is also reflected on a standard representation learning evaluation using linear probes. The best CLIP model outperforms the best publicly available ImageNet model, the Noisy Student EfficientNet-L2,[23] on 20 out of 26 different transfer datasets we tested.

Source: OpenAI blog

Accessibility

- This is two-fold:
  - **Very accessible**: Easy to just **use for inference** on the internet
  - **Not accessible at all**: **Hard to train** on your own data (make your own)
    - Possible to pre-train the generative model (re-use your own trained models), but not really to retrain CLIP
    - Costs for the large-scale language models are high (post)

Finally …

- This is a field that needs to be studied, talked about, analyzed from perspectives of multiple disciplines …
  - … check the related readings
  - … there is more and more focus in research on this … so there is a chance

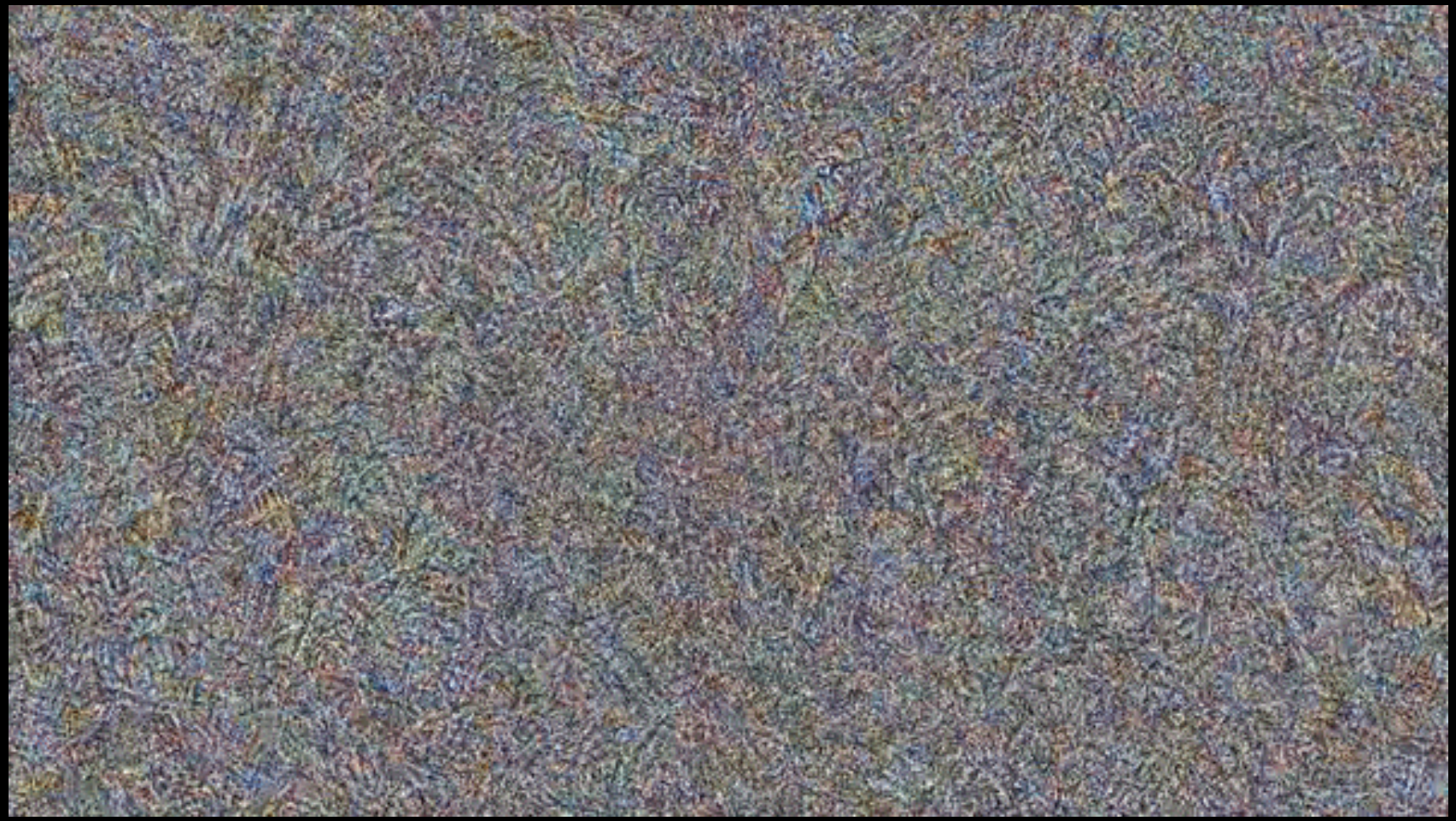
Text 2 Image

Art examples

Notebook: DiscoDiffusion

- **DiscoDiffusion**: repo, video for v4.1, notebook for v4.1, v5.0

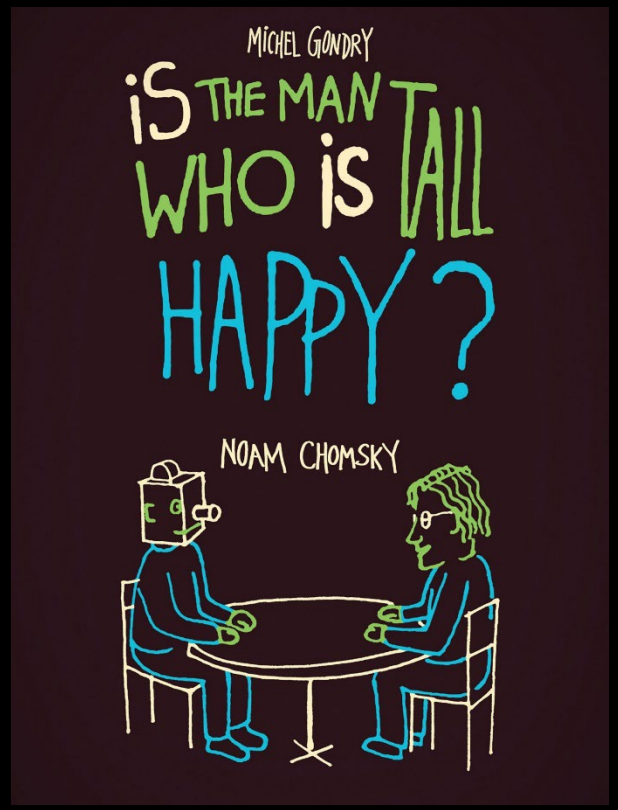
Notebook: Aphantasia



"Aphantasia" is the inability to visualize mental images, the deprivation of visual dreams.

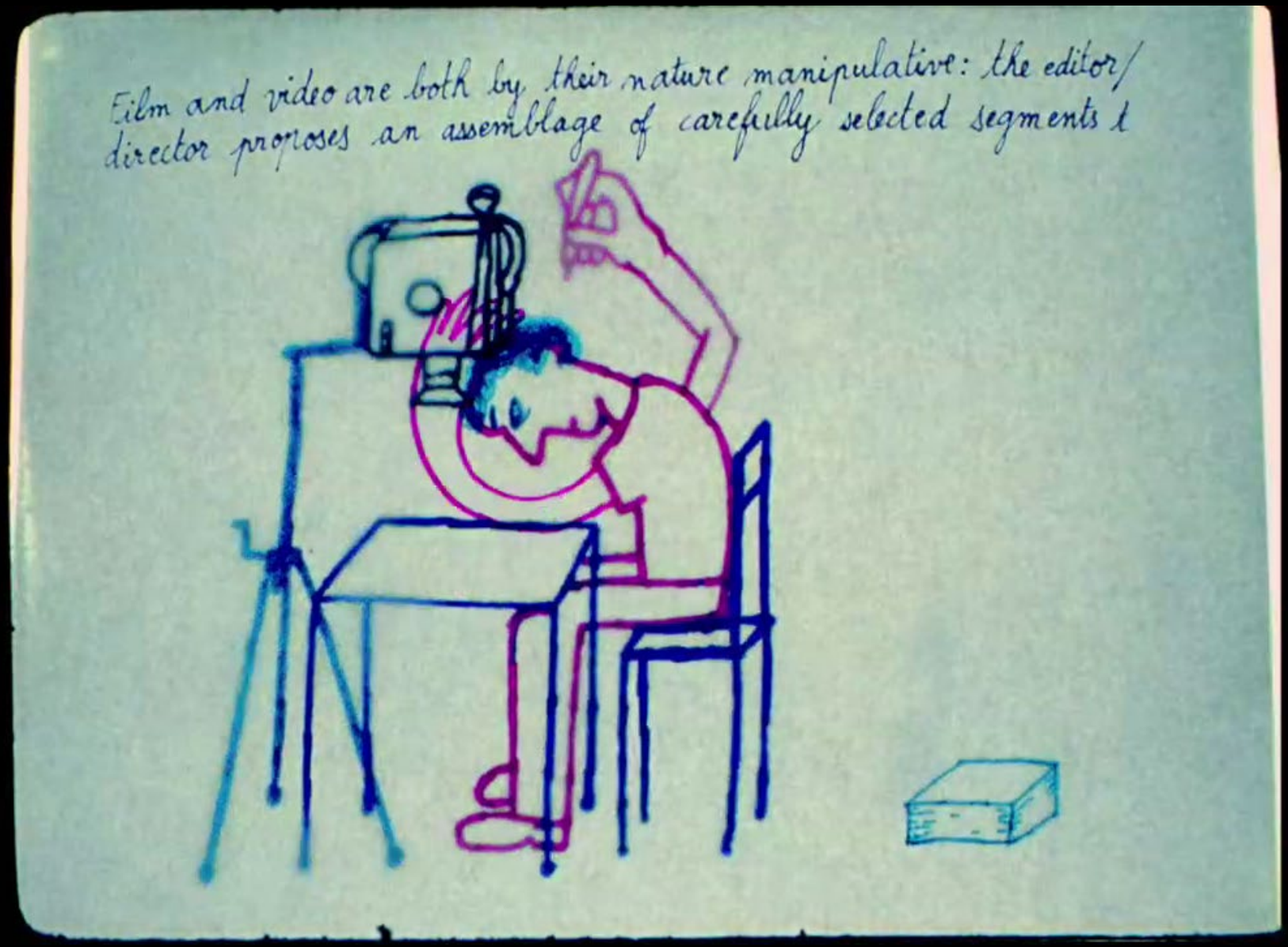
- **Aphantasia**, directly on images, also allows text to video (*illustrip*): repo

Video from



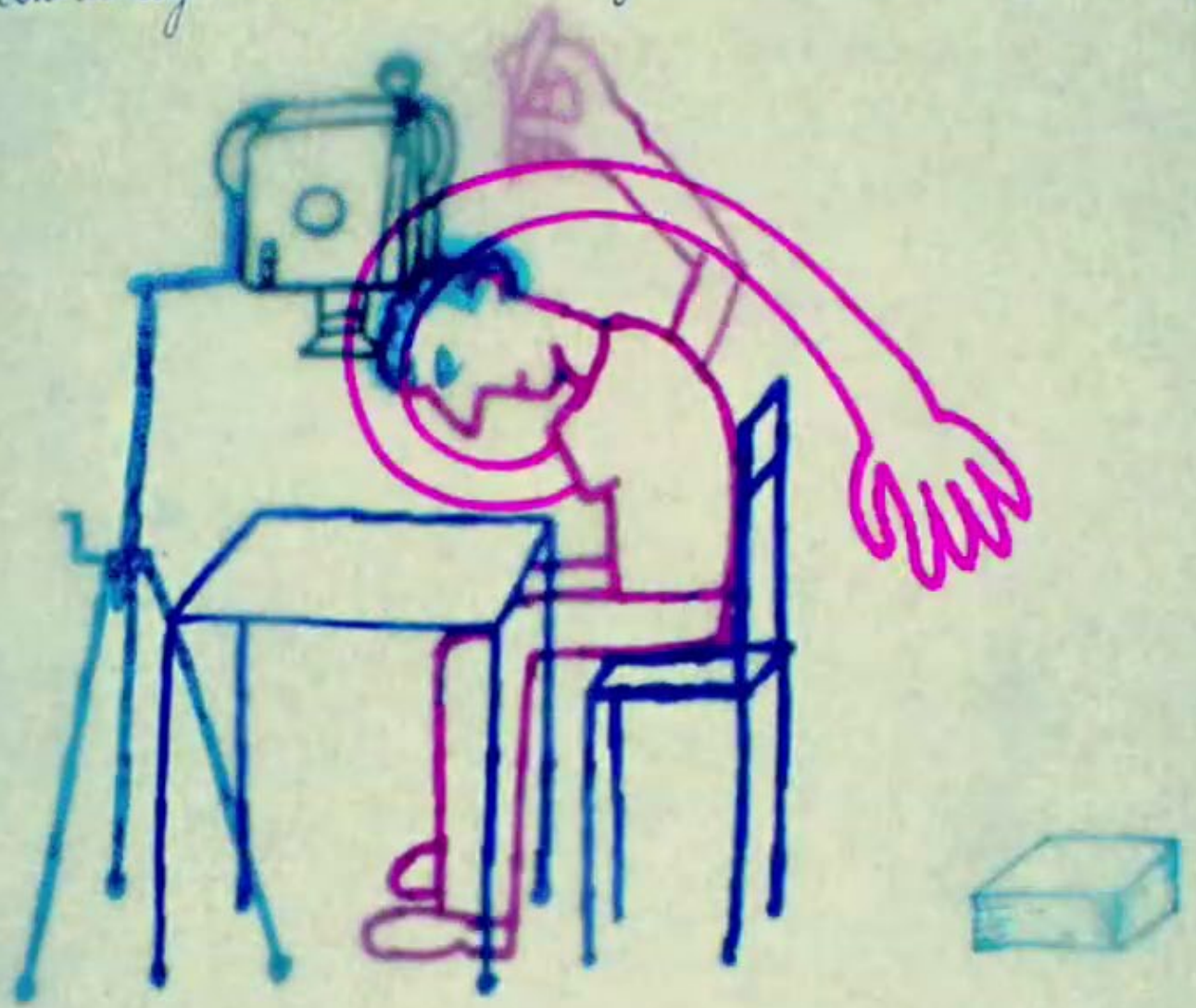
Michel Gondry as an illustrator to Noam Chomsky's ideas.

"The human brain forgets the cuts … we consider the constructed continuity as a reality"

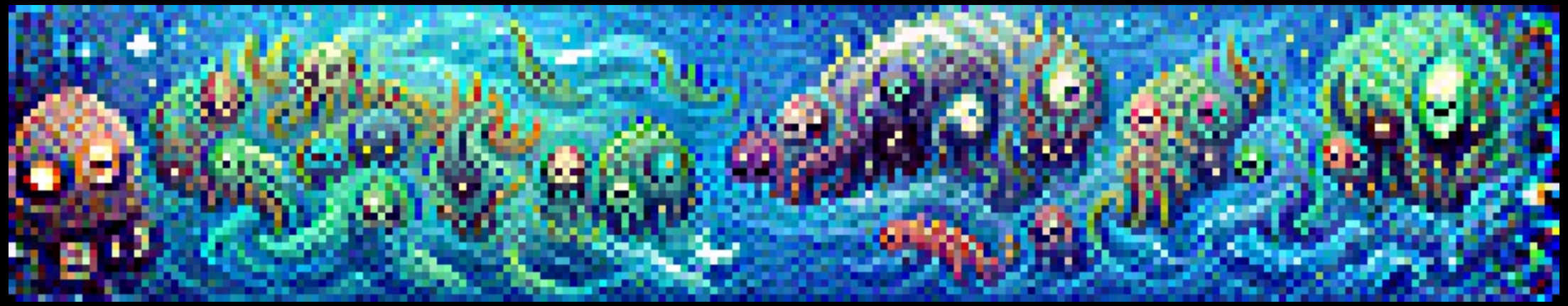


Forced abstraction (by animating every slide) to not let the viewer forget – this is a film!

If messages, or even propaganda, can be delivered, the audience is constantly reminded that they are not watching reality.



Notebook: Text 2 PixelArt



- **Pixray** by dribnet (nice stylization): repo, colab
- **Text2PixelArt**, project page: text2pixelart

Keywords

- Adding special keywords to the generated prompt:
- **“mushroom ArtStation HD”**

Source: by @kingdomakrillic

Or by art movements: here

Keywords

- Adding special authors:
- **“Aeon Ikaros, the sand creature by Karel Thole”**

generated

by Karel Thole (1914-2000, *surreal sci-fi horror*)

Source: by @RiversHaveWings

CLIP for dataset pre-filtering

- **CLIP** can also be used when preparing the dataset for other models:
  - pre-filtering in GANscapes - "Using CLIP to Judge the Training Images"

Top hits for "impressionist painting"

Bottom results for "impressionist painting"

Generative models as a mirror

- **Mirror to the society**

"While the **purpose of most fake news is misinformation and political propaganda**, our team sees it as a new type of myth that is created by people in the age of internet identities and artificial intelligence. Seeking insights on the fear and desire hidden underneath these modified or generated stories, we use machine learning methods to **generate fake articles and present them in the form of an online news blog.**"

- The Myths of Our Time: Fake News, ISEA 2019

Text 2 Image

End + Extras

Summary from the lecture

- **CycleGAN**:
  - **Model architecture** with two generators (there and back) and one discriminator
  - **Example** learning between two domains (not paired)
- **Super Resolution**:
  - **Models** as GAN structure
  - **Examples** from research, art and industry … easy to use **Applications**
  - **Limitations** tricky! *(What happens to the reality?)*

Further readings

Readings on language/vision models

- **CLIP paper** "Learning Transferable Visual Models From Natural Language Supervision" - arxiv.org/abs/2103.00020

Readings on txt2img

- Blog "**Alien Dreams**: An Emerging Art Scene" ml.berkeley.edu/blog/posts/clip-art/

Further readings

Readings on bias in language/vision models

- Paper "**Multimodal datasets**: misogyny, pornography, and malignant stereotypes": arxiv.org/abs/2110.01963
- Article "Audit finds **gender and age bias** in OpenAI's CLIP model": venturebeat.com/2021/08/10/audit-finds-gender-and-age-bias-in-openais-clip-model/
- **Opinion piece** by Yarin Gal "The robustness of massive models pre-trained on web-scraped data is tightly tied to the way their data was curated": oatml.cs.ox.ac.uk/blog/2021/06/27/web-scraped-harmful.html
- (And in classification models in general) **Gender Shades project** gendershades.org/overview.html

End of the lecture

*) PS: follows material for the practical session …