UNIVERSITÄT
LEIPZIG

Detecting Hidden Meaning in Stock Images

Master’s Thesis

Supervised by Niklas Deckers

Leipzig, 01.06.2023

Pia Sülzle

MOTIVATION

THE AMBIGUITY OF STOCK IMAGES

MOTIVATION

THE AMBIGUITY OF STOCK IMAGES

A stock image's ambiguity is the result of an **intentional design** process whereby the stock photography industry presents the **maximum range of possible meanings**, and yet, falls artfully short of "deciding" any of them.

Ward, C. G. (2007). Stock Images, Filler Content and the Ambiguous Corporate Message

MOTIVATION

THE AMBIGUITY OF STOCK IMAGES

A stock image's ambiguity is the result of an **intentional design** process whereby the stock photography industry presents the **maximum range of possible meanings**, and yet, falls artfully short of "deciding" any of them.

Ward, C. G. (2007). Stock Images, Filler Content and the Ambiguous Corporate Message

The generic stock image is […] promiscuous, intended to be resold time and again for a **range of diverse uses** and products, media platforms and contexts of reception, many of which are **unanticipated** by either the photographer or stock agency.

Frosh, P. (2020). Is Commercial Photography a Public Evil?

MOTIVATION

LITERAL DESCRIPTION VS. HIDDEN MEANING

Eine Abkühlung ist bald vonnöten: In Deutschland steht die erste Hitzewelle des Jahres in den Startlöchern. © IMAGO / Shotshop

MOTIVATION

LITERAL DESCRIPTION VS. HIDDEN MEANING

Two women standing in a lake playing with a beach ball

Eine Abkühlung ist bald vonnöten: In Deutschland steht die erste Hitzewelle des Jahres in den Startlöchern. © IMAGO / Shotshop

MOTIVATION

LITERAL DESCRIPTION VS. HIDDEN MEANING

Two women standing in a lake playing with a beach ball

Eine Abkühlung ist bald vonnöten: In Deutschland steht die erste Hitzewelle des Jahres in den Startlöchern. © IMAGO / Shotshop

Heat wave, cooling

OBJECTIVE

- Automated extraction of hidden meaning in stock images
    - Is it possible to distinguish what is shown from what is meant?
    - Examine the divergence of text and image
- Analyze the usage of stock images and textual descriptions on the web

TECHNICAL BACKGROUND AND RELATED WORK

CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

- Learn to recognize a wide variety of visual concepts in images and associate them with their names
- Can be used for a wide range of applications which deal with the connection between text and image

TECHNICAL BACKGROUND AND RELATED WORK

CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

- Learn to recognize a wide variety of visual concepts in images and associate them with their names
- Can be used for a wide range of applications which deal with the connection between text and image

TECHNICAL BACKGROUND AND RELATED WORK

CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

- Learn to recognize a wide variety of visual concepts in images and associate them with their names
- Can be used for a wide range of applications which deal with the connection between text and image

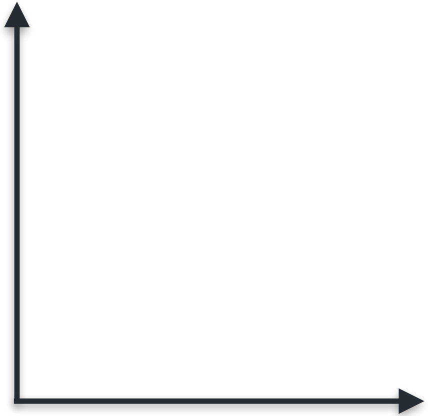
$$\longrightarrow \begin{bmatrix} 1.23 \\ 2.34 \\ 0.23 \\ 1.45 \\ \dots \\ \dots \\ \dots \end{bmatrix}$$

TECHNICAL BACKGROUND AND RELATED WORK

CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

- Learn to recognize a wide variety of visual concepts in images and associate them with their names
- Can be used for a wide range of applications which deal with the connection between text and image

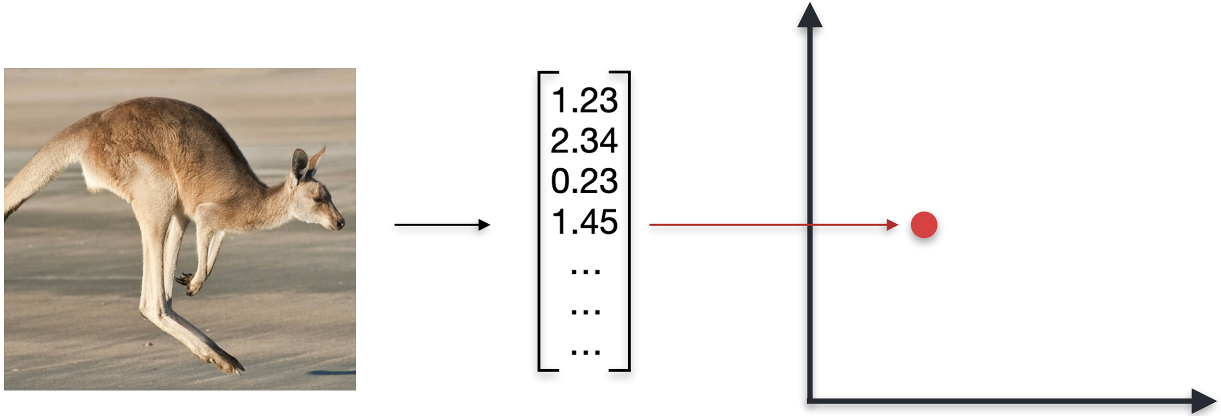
$$\longrightarrow \begin{bmatrix} 1.23 \\ 2.34 \\ 0.23 \\ 1.45 \\ \dots \\ \dots \\ \dots \end{bmatrix}$$



TECHNICAL BACKGROUND AND RELATED WORK

CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

- Learn to recognize a wide variety of visual concepts in images and associate them with their names
- Can be used for a wide range of applications which deal with the connection between text and image

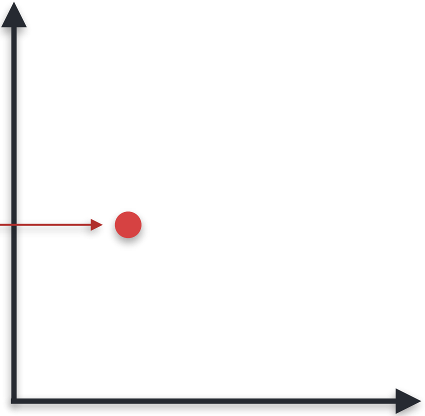


TECHNICAL BACKGROUND AND RELATED WORK

CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

- Learn to recognize a wide variety of visual concepts in images and associate them with their names
- Can be used for a wide range of applications which deal with the connection between text and image

$$\begin{bmatrix} 1.23 \\ 2.34 \\ 0.23 \\ 1.45 \\ \dots \\ \dots \\ \dots \end{bmatrix}$$

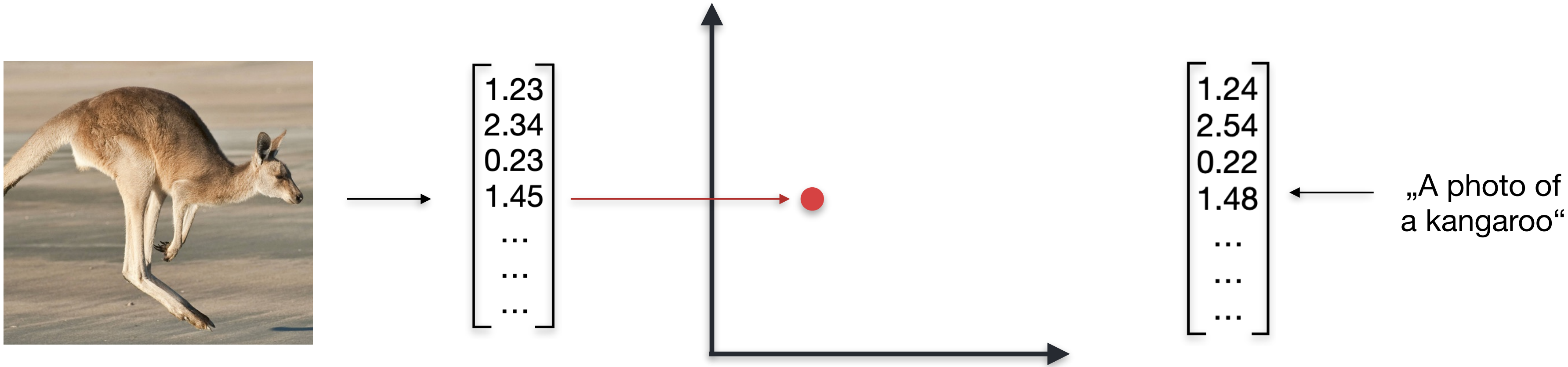


„A photo of a kangaroo“

TECHNICAL BACKGROUND AND RELATED WORK

CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

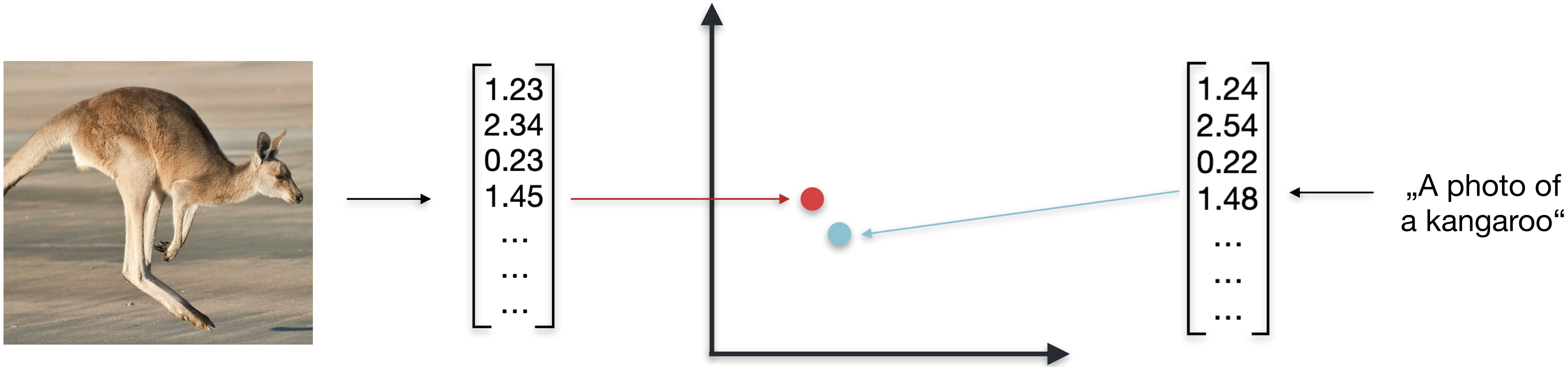
- Learn to recognize a wide variety of visual concepts in images and associate them with their names
- Can be used for a wide range of applications which deal with the connection between text and image



TECHNICAL BACKGROUND AND RELATED WORK

CLIP (CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

- Learn to recognize a wide variety of visual concepts in images and associate them with their names
- Can be used for a wide range of applications which deal with the connection between text and image



TECHNICAL BACKGROUND AND RELATED WORK

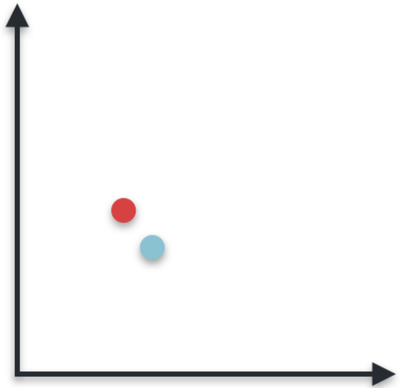
QUESTIONS ABOUT CLIP

- Is it possible to distinguish between stock and non-stock images using CLIP embeddings?
- Does CLIP describe hidden meaning or literal description?

TECHNICAL BACKGROUND AND RELATED WORK

QUESTIONS ABOUT CLIP

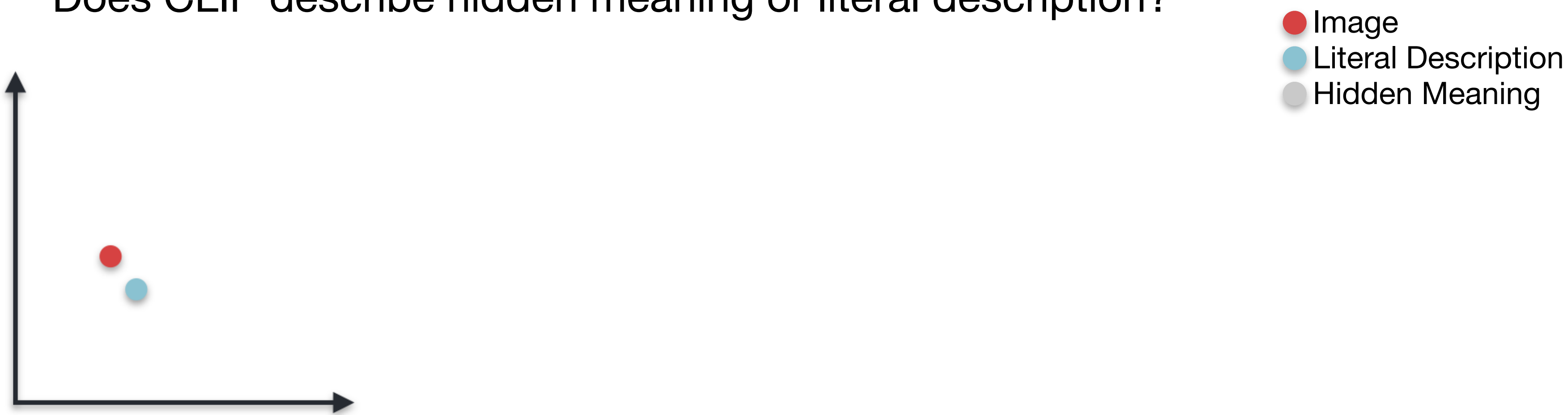
- Is it possible to distinguish between stock and non-stock images using CLIP embeddings?
- Does CLIP describe hidden meaning or literal description?



TECHNICAL BACKGROUND AND RELATED WORK

QUESTIONS ABOUT CLIP

- Is it possible to distinguish between stock and non-stock images using CLIP embeddings?
- Does CLIP describe hidden meaning or literal description?



TECHNICAL BACKGROUND AND RELATED WORK

QUESTIONS ABOUT CLIP

- Is it possible to distinguish between stock and non-stock images using CLIP embeddings?
- Does CLIP describe hidden meaning or literal description?

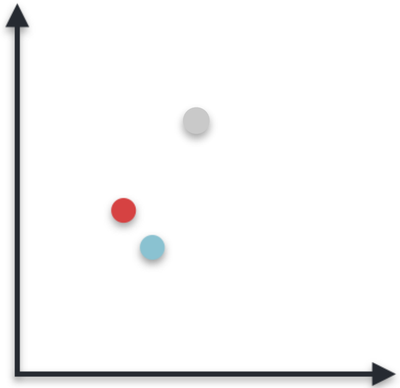


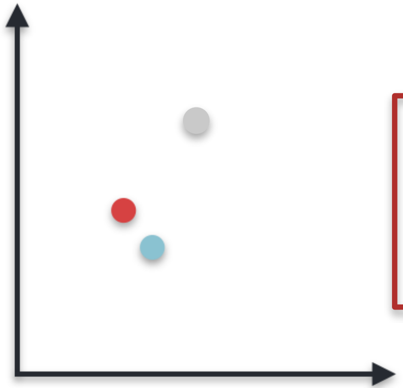
Image
Literal Description
Hidden Meaning

TECHNICAL BACKGROUND AND RELATED WORK

QUESTIONS ABOUT CLIP

- Is it possible to distinguish between stock and non-stock images using CLIP embeddings?
- Does CLIP describe hidden meaning or literal description?

Image
Literal Description
Hidden Meaning



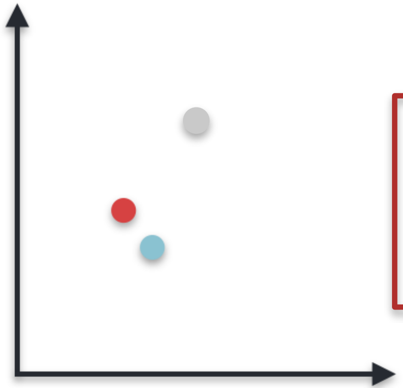
Literal description
sim(image, literal)
$\geq$
sim(image, hidden)

TECHNICAL BACKGROUND AND RELATED WORK

QUESTIONS ABOUT CLIP

- Is it possible to distinguish between stock and non-stock images using CLIP embeddings?
- Does CLIP describe hidden meaning or literal description?

Image
Literal Description
Hidden Meaning

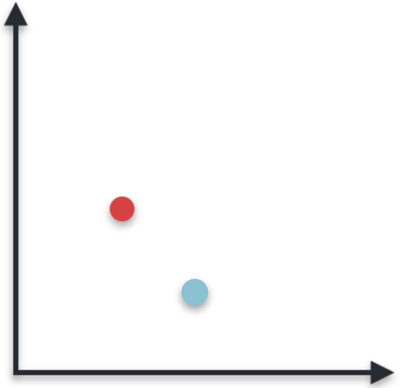


Literal description

sim(image, literal)

$\geq$

sim(image, hidden)

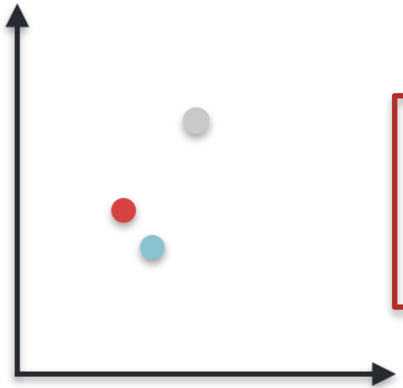


TECHNICAL BACKGROUND AND RELATED WORK

QUESTIONS ABOUT CLIP

- Is it possible to distinguish between stock and non-stock images using CLIP embeddings?
- Does CLIP describe hidden meaning or literal description?

Image
Literal Description
Hidden Meaning

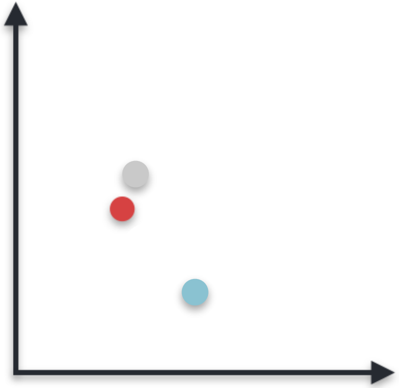


Literal description

sim(image, literal)

$\geq$

sim(image, hidden)

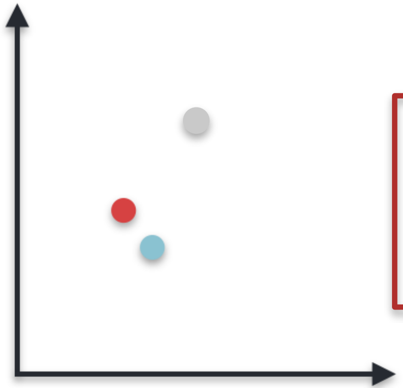


TECHNICAL BACKGROUND AND RELATED WORK

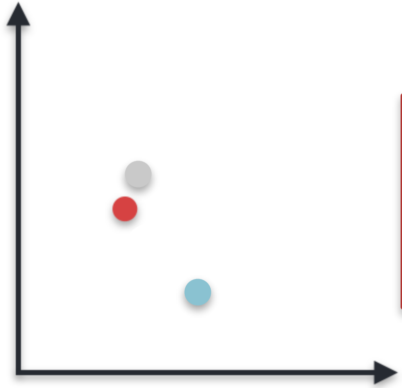
QUESTIONS ABOUT CLIP

- Is it possible to distinguish between stock and non-stock images using CLIP embeddings?
- Does CLIP describe hidden meaning or literal description?

Image
Literal Description
Hidden Meaning



Literal description
sim(image, literal)
$\geq$
sim(image, hidden)

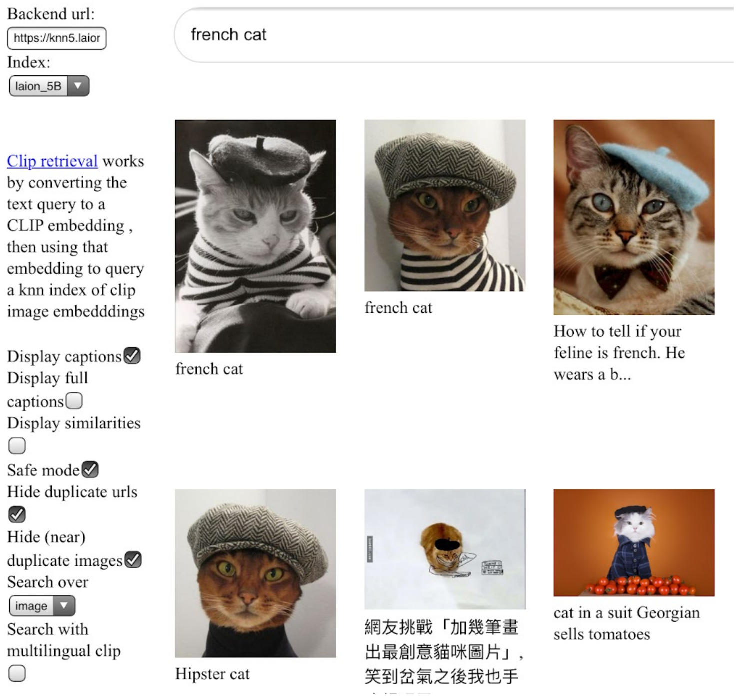


Hidden Meaning
sim(image, literal)
$\leq$
sim(image, hidden)

TECHNICAL BACKGROUND AND RELATED WORK

LAION 5B[1] & CLIP RETRIEVAL[2]

- Dataset of 5,85 billion CLIP-filtered image-text pairs
- WAT files from the Common Crawl
- Images with their corresponding alt-text attribute
- Build a large KNN index using autofaiss[3]
- CLIP Retrieval makes it possible to easily compute CLIP embeddings and build a CLIP retrieval system with them



[1] https://laion.ai/blog/laion-5b/ [2] https://rom1504.github.io/clip-retrieval [3] https://github.com/criteo/autofaiss

DATASET

- Stock image dataset
- Crawled from Pixabay[1]
- List of ~133 topics
- Per topic 500 images
- 66.277 stock images
  - With the corresponding tags

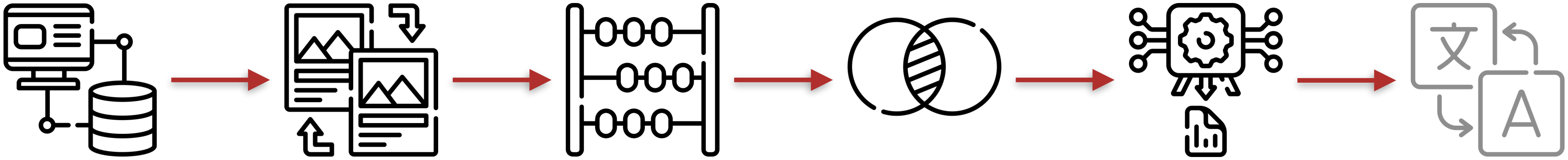

```
{
"total": 4692,
"totalHits": 500,
"hits": [
    {
        "id": 195893,
        "pageURL": "https://pixabay.com/en/blossom-bloom-flower-195893/",
        "type": "photo",
        "tags": "blossom, bloom, flower",
        "previewURL": "https://cdn.pixabay.com/photo/2013/10/15/09/12/flower-195893_15
        "previewWidth": 150,
        "previewHeight": 84,
        "webformatURL": "https://pixabay.com/get/35bbf209e13e39d2_640.jpg",
        "webformatWidth": 640,
        "webformatHeight": 360,
        "largeImageURL": "https://pixabay.com/get/ed6a99fd0a76647_1280.jpg",
        "fullHDURL": "https://pixabay.com/get/ed6a9369fd0a76647_1920.jpg",
        "imageURL": "https://pixabay.com/get/ed6a9364a9fd0a76647.jpg",
        "imageWidth": 4000,
        "imageHeight": 2250,
        "imageSize": 4731420,
        "views": 7671,
        "downloads": 6439,
        "likes": 5,
        "comments": 2,
        "user_id": 48777,
        "user": "Josch13",
        "userImageURL": "https://cdn.pixabay.com/user/2013/11/05/02-10-23-764_250x250.
    },
    {
        "id": 73424,
        ...
    },
    ...
]
}
```


Sample response for a Pixabay API request

[1] https://pixabay.com/

EXPERIMENTAL SETUP

A ROADMAP

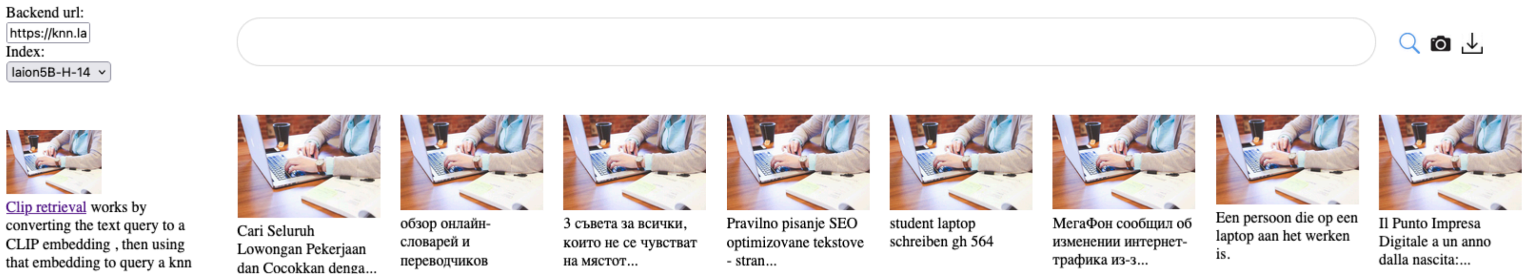


EXAMPLE STOCK IMAGE

EXPERIMENTAL SETUP

FINDING DUPLICATES OF STOCK IMAGES

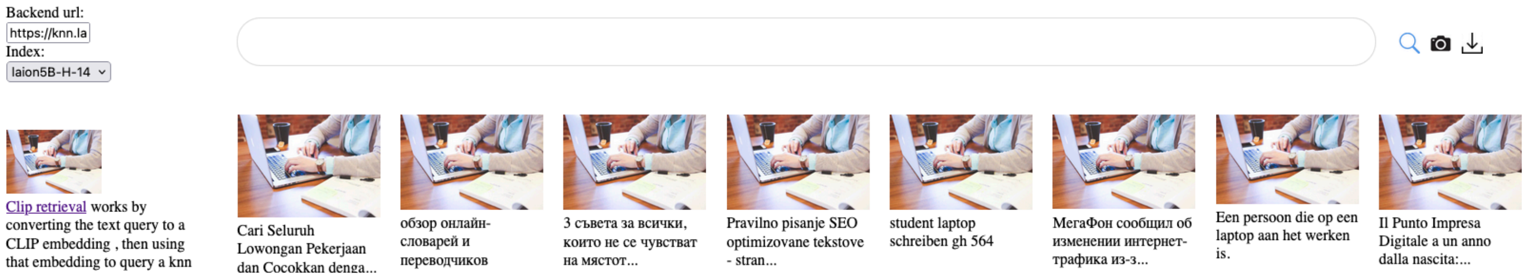
- Iteratively find as many captions as possible for each image
  - Filter English captions for later use



EXPERIMENTAL SETUP

FINDING DUPLICATES OF STOCK IMAGES

- Iteratively find as many captions as possible for each image
  - Filter English captions for later use



Extracting many different hidden meanings (as well as some literal descriptions)

EXPERIMENTAL SETUP

DATA CREATION

- **Goal:** To have several hidden meanings and literal descriptions for an image.

[1] https://huggingface.co/ydshieh/vit-gpt2-coco-en

EXPERIMENTAL SETUP

DATA CREATION

- **Goal:** To have several hidden meanings and literal descriptions for an image.

- As many captions as possible are crawled per image
- As a literal representation of the image a caption created by an Image2Text[1] model is used
  - Assumption: generated description is representative for what can be seen on the image

[1] https://huggingface.co/ydshieh/vit-gpt2-coco-en

EXPERIMENTAL SETUP

DATA CREATION

EXPERIMENTAL SETUP

DATA CREATION

- Literal description:
    - „A woman is typing on a laptop computer“

EXPERIMENTAL SETUP

DATA CREATION

- Literal description:
  - „A woman is typing on a laptop computer“
- Excerpt from the captions of the duplicates:
  - Top 5 Blogging Platforms
  - 10 Ways to Earn Money Online from Home Without Investment
  - Should Small Businesses Go For Enterprise Resource Planning?
  - Two hands typing on laptop on desktop with coffee

EXPERIMENTAL SETUP

WORD FREQUENCIES

- **Goal:** Find the most common hidden meaning of an image.

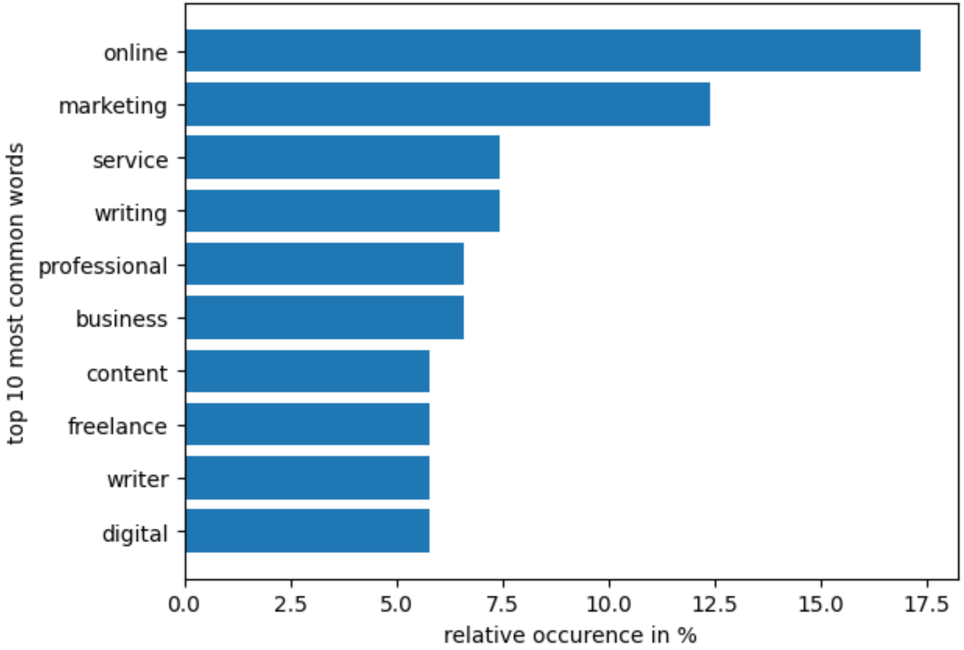
EXPERIMENTAL SETUP

WORD FREQUENCIES

- **Goal:** Find the most common hidden meaning of an image.
- Create preprocessed word set for every caption and for the generated description
- Remove all words that appear in the literal description from the word sets of the captions
  - Assumption: only words that do not appear in the literal description belong to the hidden meaning
- Count the occurrences of all remaining words

EXPERIMENTAL SETUP

WORD FREQUENCIES - EXAMPLE



EXPERIMENTAL SETUP

SIMILARITY ANALYSIS - CLIP

- **Goal:** Find out whether CLIP captures the **hidden** or the **literal** meaning.

EXPERIMENTAL SETUP

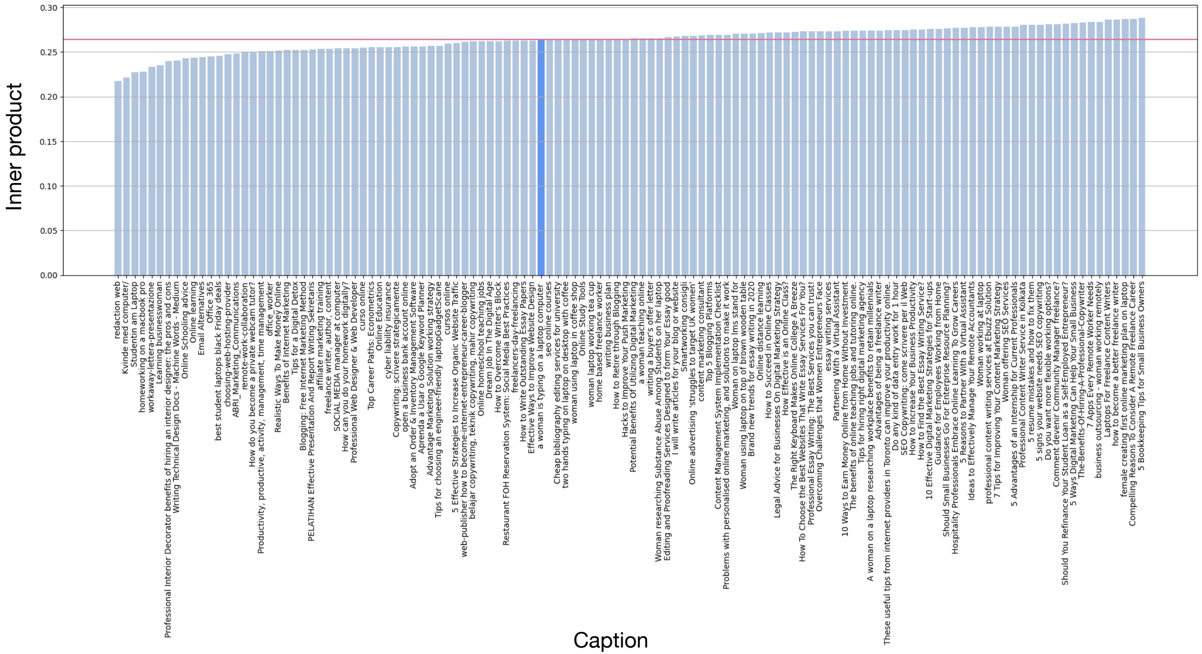
SIMILARITY ANALYSIS - CLIP

- **Goal:** Find out whether CLIP captures the **hidden** or the **literal** meaning.

- Used Model: CLIP
- Create embedding for the image, the literal description and all captions
- Calculate inner product between the embeddings of the image and every texts (description + captions)

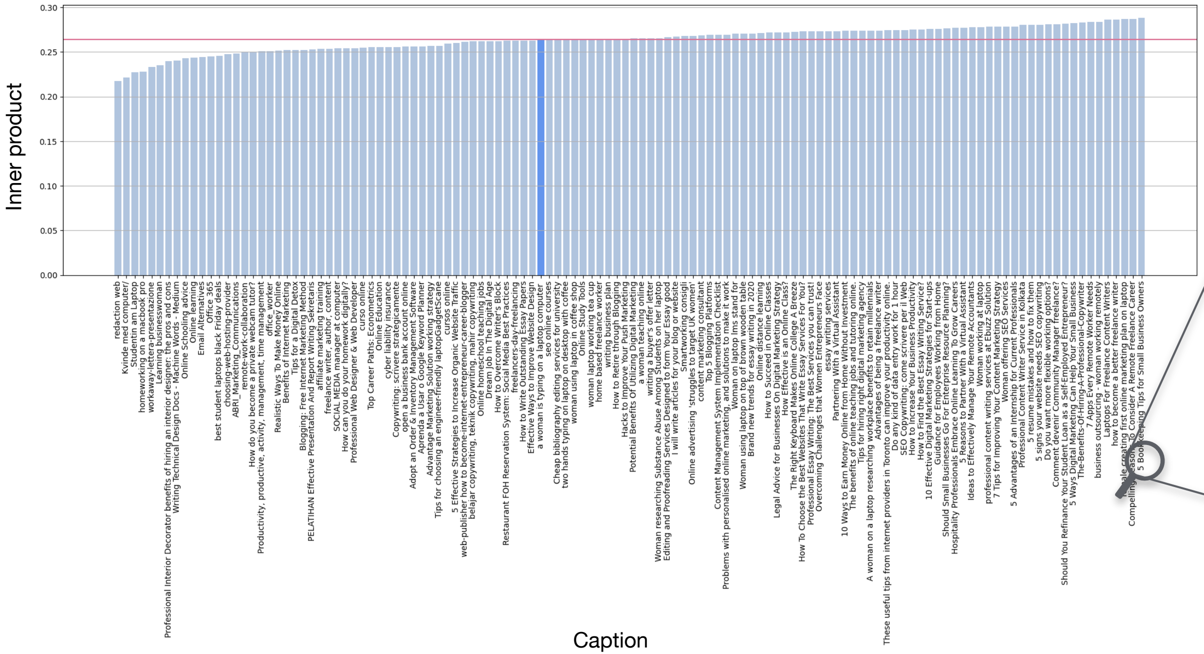
EXPERIMENTAL SETUP

SIMILARITY ANALYSIS - CLIP - EXAMPLE



EXPERIMENTAL SETUP

SIMILARITY ANALYSIS - CLIP - EXAMPLE

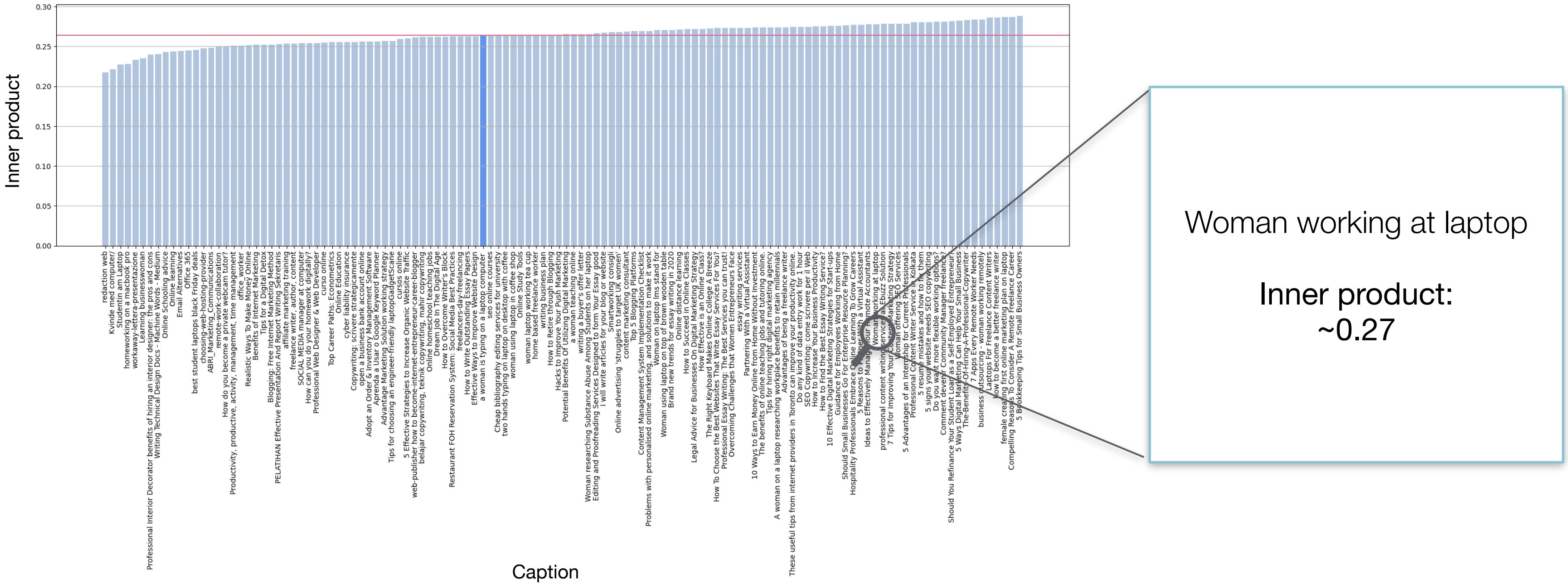


5 Bookkeeping Tips for Small Business Owners

Inner product: ~0.28

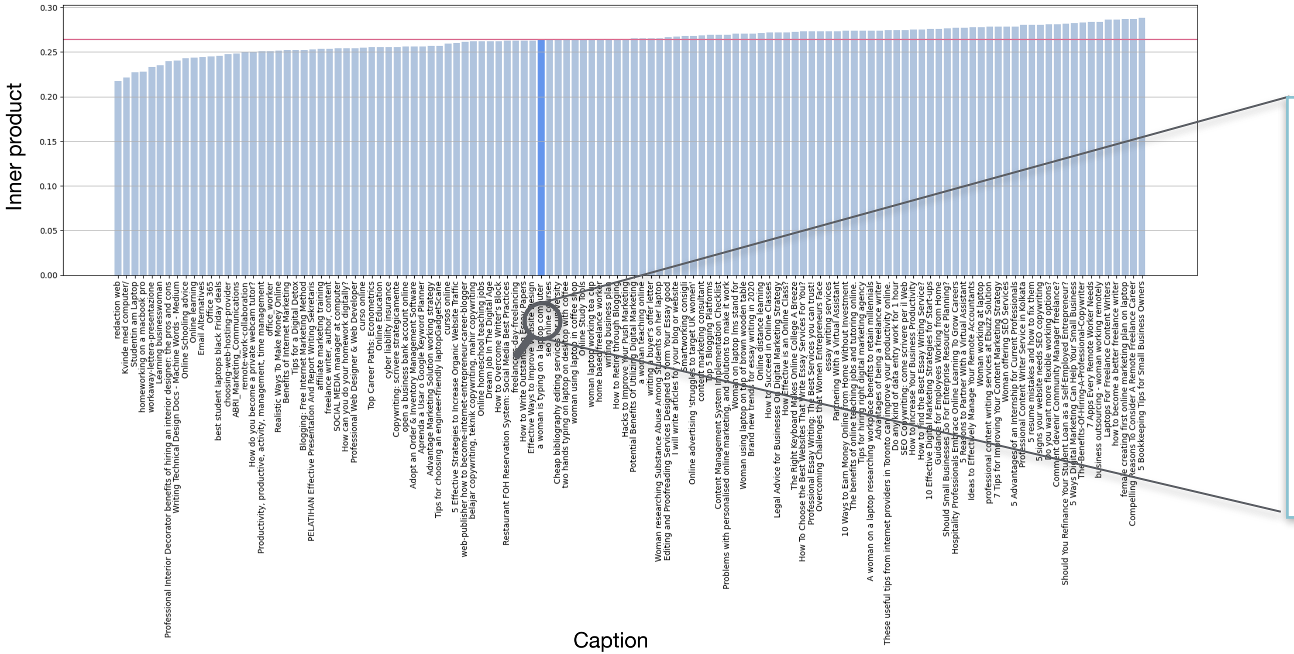
EXPERIMENTAL SETUP

SIMILARITY ANALYSIS - CLIP - EXAMPLE



EXPERIMENTAL SETUP

SIMILARITY ANALYSIS - CLIP - EXAMPLE

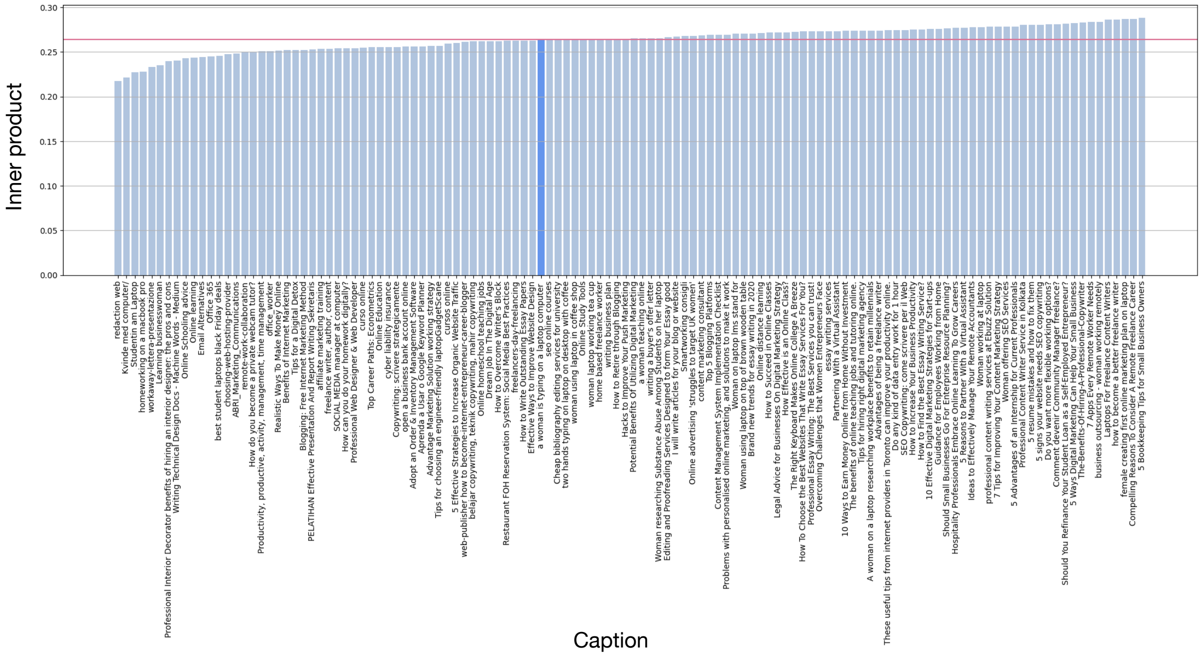


A woman is typing on a laptop computer (Image2Text model generated description)

Inner product: ~0.26

EXPERIMENTAL SETUP

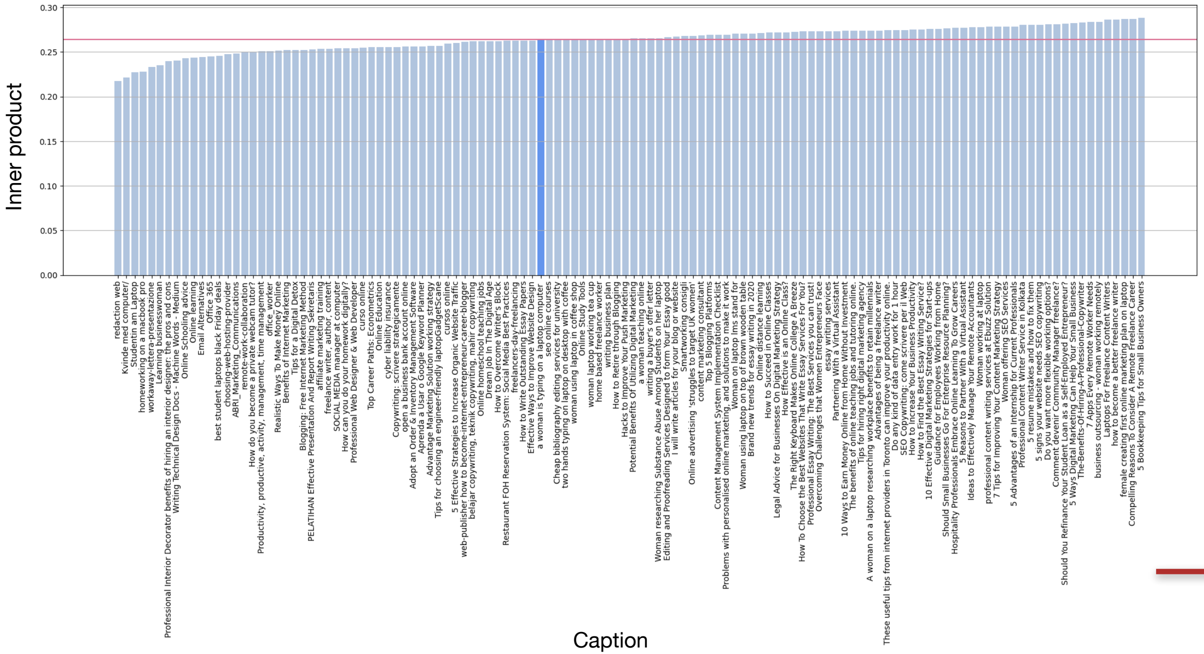
SIMILARITY ANALYSIS - CLIP - EXAMPLE



- Using CLIP does not help to distinguish between hidden or literal meaning
- CLIP can capture the literal meaning as well as the hidden meaning

EXPERIMENTAL SETUP

SIMILARITY ANALYSIS - CLIP - EXAMPLE



- Using CLIP does not help to distinguish between hidden or literal meaning
- CLIP can capture the literal meaning as well as the hidden meaning

A language model could show the difference between hidden and literal better

EXPERIMENTAL SETUP

SIMILARITY ANALYSIS - SBERT

- **Goal:** Find a measure to distinguish hidden meaning from literal description.

[1] https://www.sbert.net/

EXPERIMENTAL SETUP

SIMILARITY ANALYSIS - SBERT

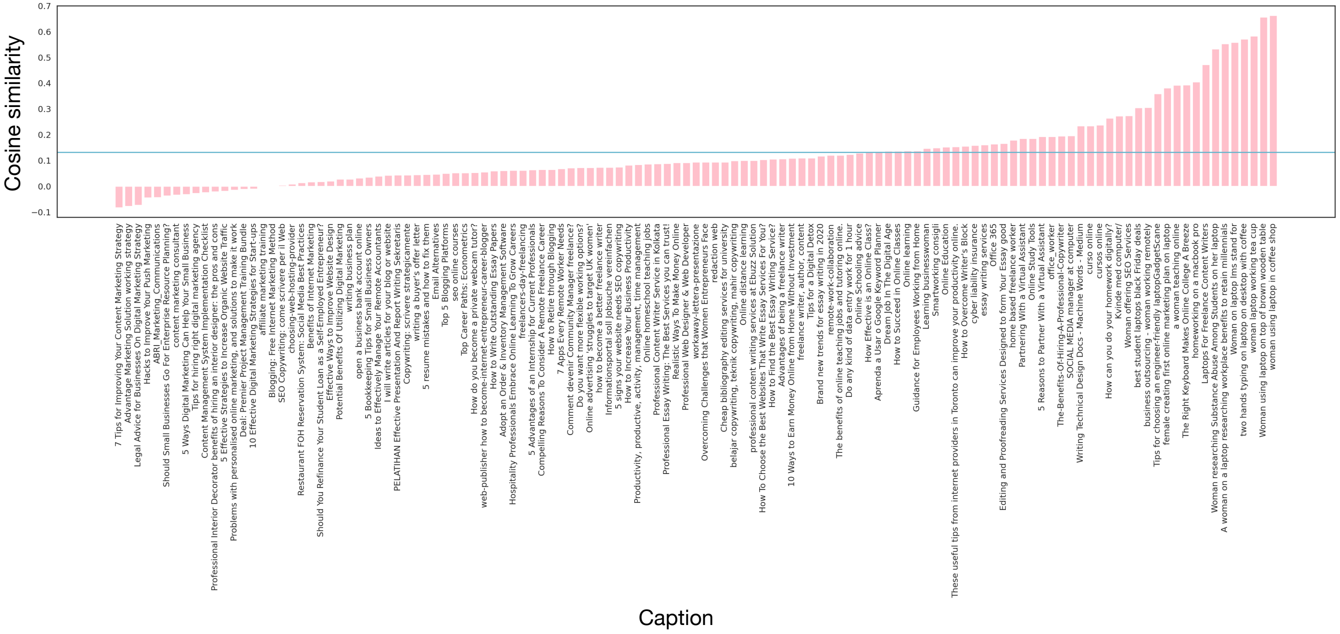
- **Goal:** Find a measure to distinguish hidden meaning from literal description.

- Used Model: SBERT[1]
- Create embedding for the literal description and all captions
- Use literal description as representation of the image
  - Calculate inner product between the literal description and every caption

[1] https://www.sbert.net/

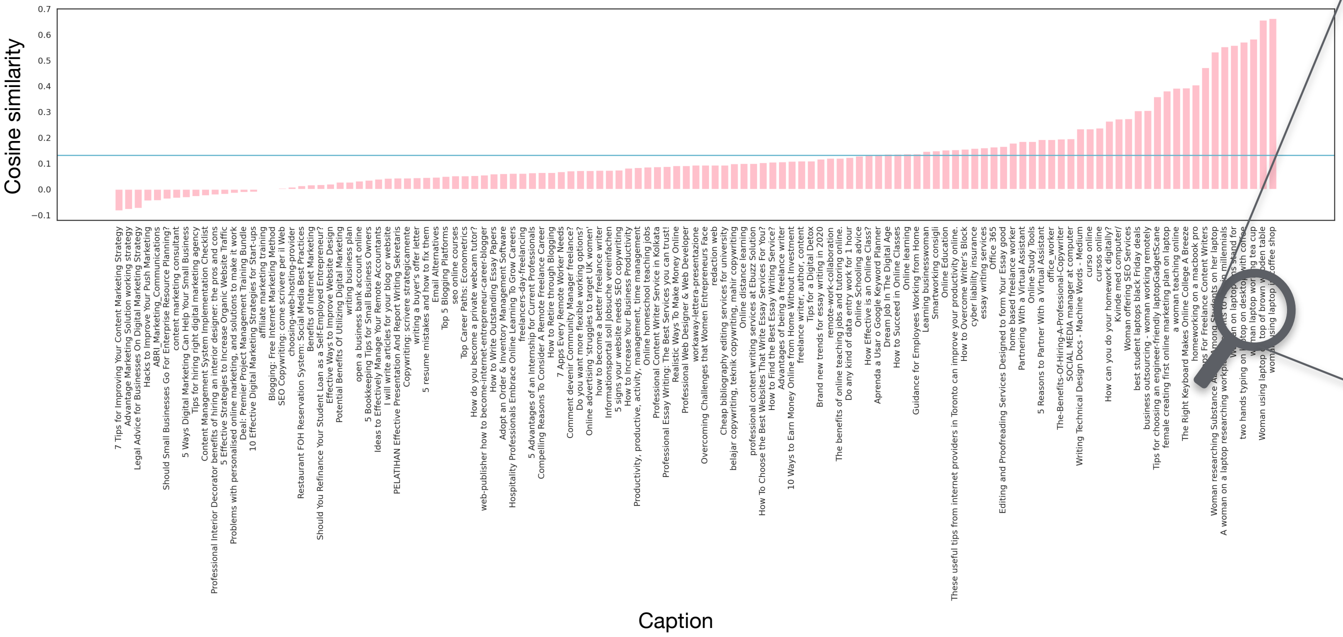
EXPERIMENTAL SETUP

SIMILARITY ANALYSIS - SBERT - EXAMPLE



EXPERIMENTAL SETUP

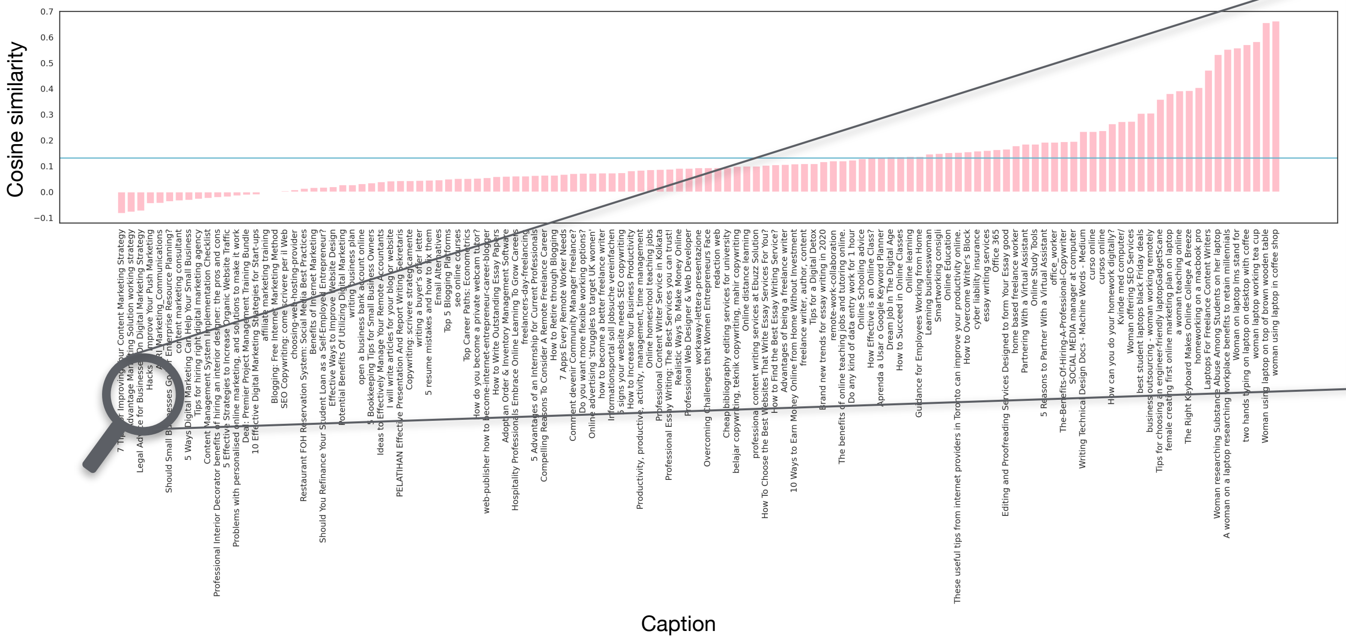
SIMILARITY ANALYSIS - SBERT - EXAMPLE



1. Woman using laptop in coffee shop
2. Woman using laptop on top of brown wooden table
3. Woman laptop working tea cup
4. Two hands typing on laptop on desktop with coffee

EXPERIMENTAL SETUP

SIMILARITY ANALYSIS - SBERT - EXAMPLE



116. 7 Tips For Improving Your Content Marketing Strategy
117. Advantage Marketing Solution working strategy
118. Legal Advice for Businesses On Digital Marketing Strategy
119. Hacks to improve your push Marketing

EXPERIMENTAL SETUP

NEURAL NETWORK TO CREATE HIDDEN MEANING EMBEDDING (EVALUATION IN PROGRESS)

- **Goal:** To create a CLIP embedding which contains one of the hidden meanings of the image.

EXPERIMENTAL SETUP

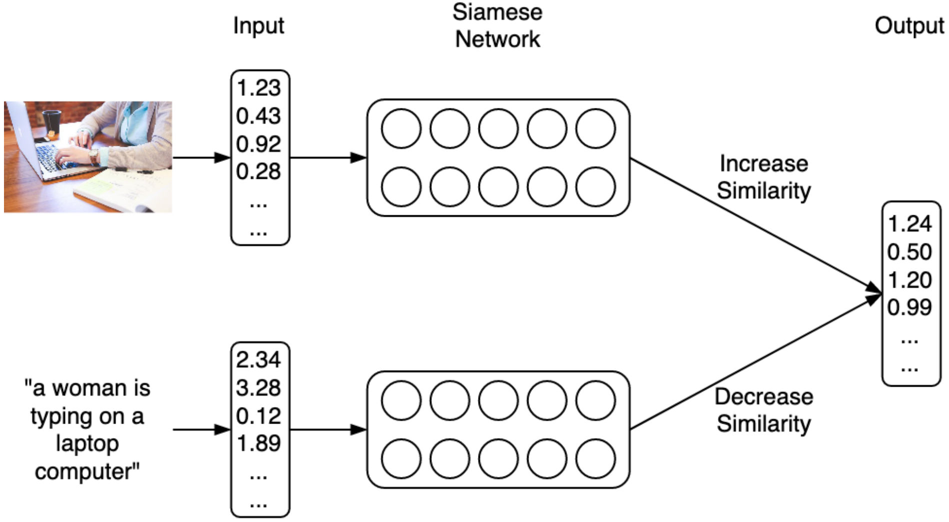
NEURAL NETWORK TO CREATE HIDDEN MEANING EMBEDDING (EVALUATION IN PROGRESS)

- **Goal:** To create a CLIP embedding which contains one of the hidden meanings of the image.
- Input: CLIP embedding of the image, CLIP embedding of the Image2Text-generated description
- Loss:
  - Maximize cosine similarity between image & output embedding
  - Minimize cosine similarity between literal description & output embedding

EXPERIMENTAL SETUP

NEURAL NETWORK TO CREATE HIDDEN MEANING EMBEDDING (EVALUATION IN PROGRESS)

- **Goal:** To create a CLIP embedding which contains one of the hidden meanings of the image.
- Input: CLIP embedding of the image, CLIP embedding of the Image2Text-generated description
- Loss:
  - Maximize cosine similarity between image & output embedding
  - Minimize cosine similarity between literal description & output embedding



CONCLUSION

- First **stock image dataset** with 66.277 stock images and their tags
- A simple analysis of the **word frequencies** of the captions can already give information about the hidden meaning of an image
- CLIP alone is **not suitable** to find a hidden meaning in images
- SBERT can capture a **difference between** the **literal description** of an image and the **hidden meaning** of an image
- Neural Network which outputs a **hidden meaning embedding**

OUTLOOK

WITHIN THIS THESIS

- Convert the hidden meaning embedding to text to see if it captures the hidden meaning
- Improvement of the network (through additional use of the LM)
- Final experiment „in the wild“
  - On a News dataset

FOLLOWING THIS THESIS

- Find the passages in a whole text that are related to a picture OR
- Find a picture that represents certain text passages
  - E.g. through the connection to generative Text2Image models
- Use the data set for further classification tasks

BIBLIOGRAPHY

- Frosh, P. (2020). Is Commercial Photography a Public Evil? Beyond the Critique of Stock Photography. 10.5040/9781350054998.ch-010.
- Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. In: International conference on machine learning (pp. 8748-8763).
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.
- Schuhmann, C. et al. (2022). Laion-5B: An open large-scale dataset for training next generation image-text models.
- Ward, C. G. (2007). Stock Images, Filler Content and the Ambiguous Corporate Message. M/C Journal, 10(5). https://doi.org/10.5204/mcj.2706

Image

- Hitzewelle in Deutschland läuft an – doch für Italien bedeutet das neue Hoch vor allem neue Fluten. (2023, 30. Mai). https://www.tz.de/welt/italien-unwetter-hochwasser-ueberschwemmung-hitzewelle-deutschland-sommerwetter-spanien-zr-92304148.html