Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models

Jiuxiang Gu  
Jianfei Cai  
Shafiq Joty  
Li Niu  
Gang Wang
Goal

Image-to-Text Retrieval

A young man doing a skateboard trick while others watch
A man doing a skate trick during a competition event with an audience
Guys on a course made for skateboarding
A group of people doing skateboarding tricks on a car
A boy riding on his skateboard at a skate park while other guys watch

Text-to-Image Retrieval

Bright room with a couch and various different dressers
Classical Pipeline

Image Encoder

\[ i \]

\[ \text{Image Feature} \]

\[ \nu_h \]

Similarity

\[ t_h \]

Text Feature

Text Encoder

\[ C \]

Bright room with a couch and various different dressers
Motivation:
Look $\rightarrow$ Imagine $\rightarrow$ Match

Image-to-Text Retrieval

Global Similarity

Local Similarity

Imagine

Text-to-Image Retrieval

Global Similarity

Local Similarity

Imagine
Look ➔ Imagine

A living room with wooden floors and furniture

A big, open room with large windows and wooden floors
Match

A living room with wooden floors and furniture

A big, open room with large windows and wooden floors
Look ➔ Imagine

A big, open room with large windows and wooden floors
A big, open room with large windows and wooden floors
Proposed Approach

A big, open room with large windows and wooden floors

A living room with wooden floors and furniture
Cross-Modal Retrieval with Generative Learning

A big, open room with large windows and wooden floors

A living room with wooden floors and furniture
Cross-Modal Retrieval with Generative Learning

\[ \hat{i} = G_i(z, t_c) \]
\[ z \sim \mathcal{N}(0, 1) \]
\[ t_c \sim \mathcal{N}(\mu(p(t_i)), \sigma(p(t_i))) \]
Results

**Image-to-Text Retrieval**

**Query**
- sheep graze in a lushly green mountain meadow
- a group of sheep grazing in a grassy valley

**GT Caption**

**Text-to-Image Retrieval**

**Query**
A tall church sitting next to a lush green field

**GT Image**
Results (Classical Pipeline)

Image-to-Text Retrieval

**Look**

*Query*

- sheep graze in a lushly green mountain meadow
- a group of sheep grazing in a grassy valley

**Match**

A herd of sheep standing on top of a hill

Text-to-Image Retrieval

**Look**

*Query*

A tall church sitting next to a lush green field

**Match**

GT Image

GT Caption
Results (Ours)

Image-to-Text Retrieval

Query

GT Caption
- sheep graze in a lushly green mountain meadow
- a group of sheep grazing in a grassy valley

Look

Imagine

Match

A herd of sheep grazing on a lush green field

Sheep graze in a lushly green mountain meadow

Text-to-Image Retrieval

Query

A tall church sitting next to a lush green field

GT Image

Look

Imagine

Match

Looking at a church in a lush green field.
At the Poster:

- Additional details
- Quantitative results
- Discussion