

Describing Differences in Image Sets with Natural Language

Lisa Dunlap^{*}, Yuhui Zhang^{*}, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell[†], Jacob Steinhardt[†],
Joseph E. Gonzalez[†], Serena Yeung-Levy[†]

A totally tubular collab between UC Berkeley and Stanford
Equal ^{*}authorship and [†]advising contribution (author order randomly chosen)

24.4.18
Minkyu Kim
EffL@POSTECH

Index

- **Introduction**
 - Reasoning the difference between two sets
- **Set Difference Captioning**
 - What is “Difference Captioning”
 - Benchmark: VisDiffBench
 - Evaluation: how to measure ‘this model is good at difference captioning’
- **Our Method: VisDiff (Proposer + Ranker)**
 - What is Proposer, Ranker
 - Image-based, Feature-based, Caption-based
- **Results**
- **Applications**
- **Conclusion**

Introduction

Questions proposed before in ML field:

- How do visual concepts shift from a decade ago to now?



2009 /
Africa



2017 /
Africa

- What types of images are more or less memorable for humans?



a) Most memorable images (86%)



c) Least memorable images (34%)

Introduction

Questions proposed before in ML field:

- How do visual concepts shift from a decade ago to now?
- What types of images are more or less memorable for humans?

...

👉 Share a **common desideratum**:

Discovering differences between two sets of images

Introduction

How can we find **differences between two sets?**

For doing this, we will...

- Scan all images
- Find differences

Small sets okay. But, if we meet large scale set (e.g. thousands of images)?

- 🙄 → 😇

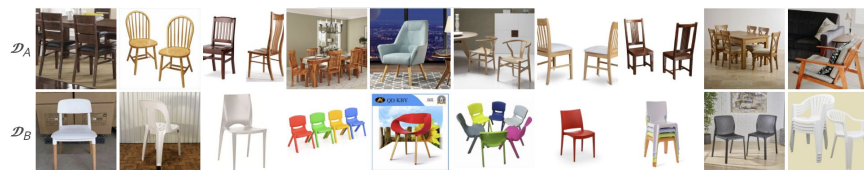
Introduction

Solution. Find differences by using learning-based methods (i.e. models)

- Exploring the task of describing differences between image sets

👉 “Difference Captioning”

- Find the most salient differences between sets



VisDiff: \mathcal{D}_A contains more..

| | Score |
|-----------------------|-------|
| Wooden furniture | 0.887 |
| Antique furniture | 0.798 |
| Dining room settings | 0.752 |
| Dining room furniture | 0.730 |
| Carved back chairs | 0.712 |

Introduction

Why find salient differences. We can find many differences!

- However, **end users** are typically interested in...
 - What can **most effectively differentiate** between the two sets
 - e.g. (below figures) “birthday”, “people posing for a picture” are valid differences
 - 👉 “people posing for a picture” **better separates** the sets

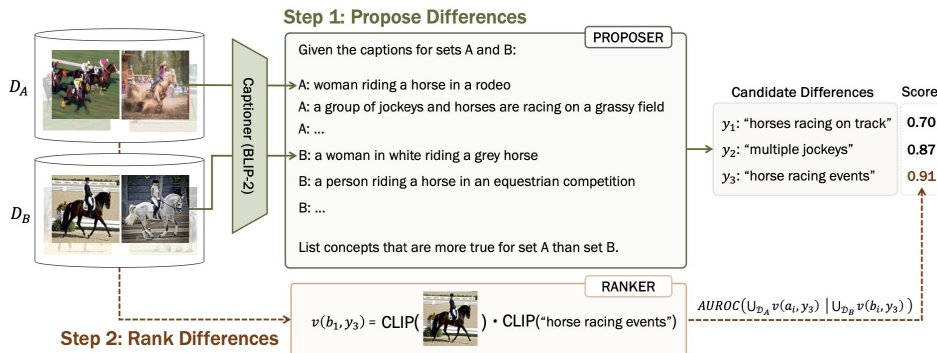


Introduction

Concern: Computational cost

- It requires reasoning over all the given images
- No existing models can effectively reason about thousands of images

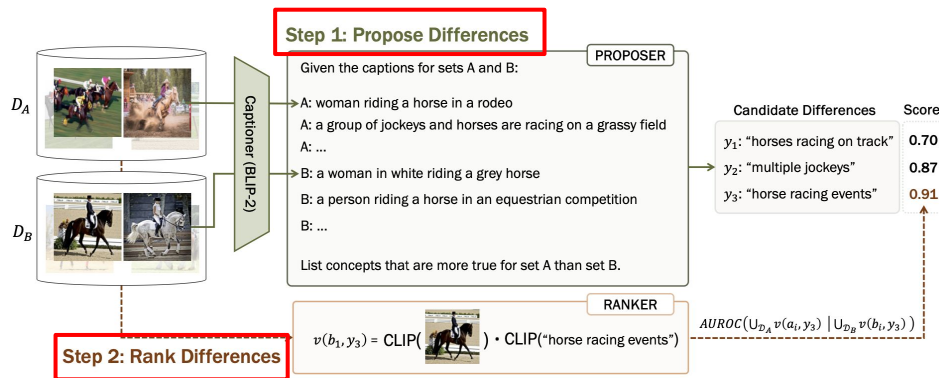
👉 Propose the framework “VisDiff”



Introduction

Framework. VisDiff

- Two-stage **proposer**-**ranker** approach
 - Proposer**. Randomly samples subsets of images from D_A and D_B to generate a set of candidate differences in natural language
 - Ranker**. Scores the salience and significance of each candidate by validating how often this difference is true for individual samples in the sets

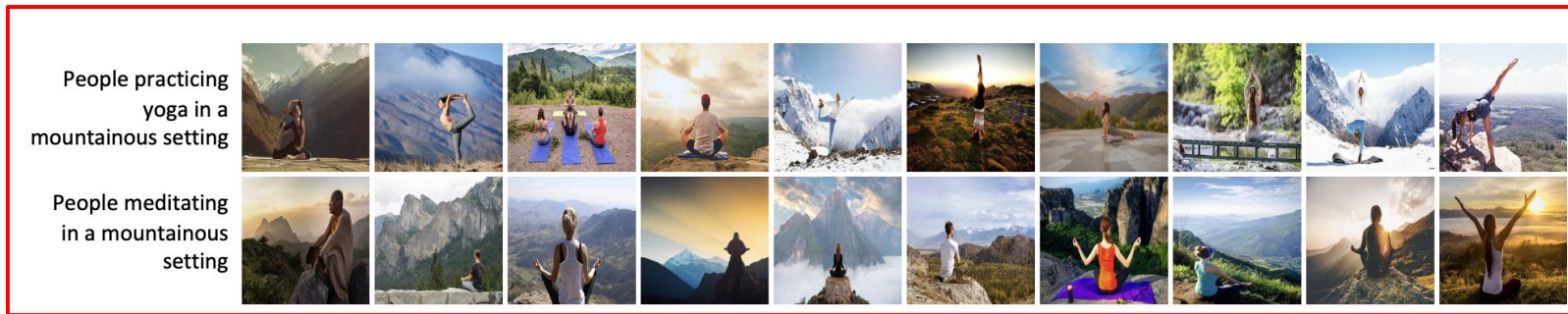


Introduction

Question. How can we evaluate the performance for this?

👉 Construct the benchmark **‘VisDiffBench’**

- 187 paired image sets with ground-truth differences
- Evaluate the predicted differences by using LLM (i.e. GPT-4)



Paired image sets

Introduction

Summary. Authors propose...

- New task: **Difference Captioning**
 - Generate differences between two datasets as language form
- New framework: **VisDiff**
 - Propose the differences + Mark the rank
 - (Will present this) Can be applied to a variety of applications
- New benchmark: **VisDiffBench**
 - Measure the performance by using LLM

Set Difference Captioning

Problem Definition.

- Given two image datasets D_A and D_B , generate a language descriptions that are more often true in D_A than D_B
 - Generate differences
 - Compare these with all images in D_A and D_B

VisDiff: \mathcal{D}_A contains more..

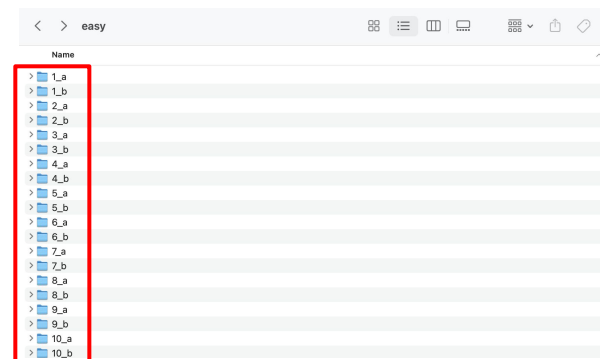
| | Score |
|--------------------------------|-------|
| Penguins standing together | 1.000 |
| Penguins in various activity | 1.000 |
| Multiple penguins in one frame | 1.000 |
| Variety of penguin species | 1.000 |
| Standing penguins | 1.000 |

Answer: \mathcal{D}_A Penguins in the snow, \mathcal{D}_B = Seals in the snow.

Set Difference Captioning

Benchmark. VisDiffBench

- 187 paired image sets each with a ground-truth difference description
 - $\{D_A, D_B\} \times 187$
 - Configuration (getting from existed datasets)
 - 150 paired sets from PairedImageSets
 - 14 paired sets from ImageNet-R
 - 23 paired sets from ImageNet*



Set Difference Captioning

Configuration. VisDiffBench

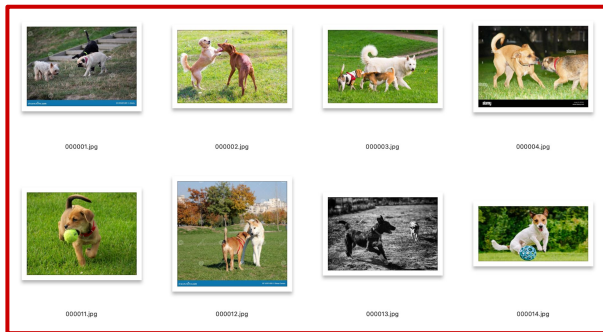
- **PairedImageSets.** (made by authors)
 - **Motivation.** Overcome the shortcoming of ImageNet-R and ImageNet*
 - Mainly capture stylistic differences
 - Only contain 37 differences in total
 - **How to make**
 - **Prompt** “generate 150 paired sentences with three difficulty levels of differences” to GPT-4
 - **easy.** “dogs playing in a park” vs. “cats playing in a park”
 - **medium.** “SUVs on the road” vs. “sedans on the road”
 - **hard.** “Sunrise over Santorini, Greece” vs. “Sunset over Santorini, Greece”
 - **Manually adjust** the annotated difficulty levels
 - **Retrieve** the top 100 images from Bing for each sentence

Set Difference Captioning

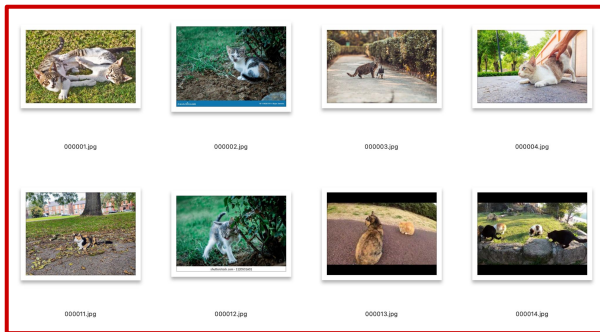
Configuration. VisDiffBench

- **PairedImageSets.** (made by authors)
 - **Results.** 50 easy, 50 medium, and 50 hard paired image sets
 - 100 images for each set
 - Paired image sets has **one ground-truth differences** (e.g. Animal species)

“Dogs playing in a park”



“Cats playing in a park”



Set Difference Captioning

Configuration. VisDiffBench

- **ImageNet-R.**
 - Renditions of 200 ImageNet classes across 14 categories
 - Ground-truth. **Name of the category** (e.g. art, cartoon, painting)
- **ImageNet*.**
 - Synthetic images transformed from original ImageNet images using textual inversion
 - Ground-truth. **Name of the category** (e.g. in the beach)



Set Difference Captioning

Evaluation. Performance on VisDiffBench

- Whether the difference is semantically similar to the ground truth
- **How to measure.** Using GPT-4
 - Categorize similarity into three levels
 - 0: no match, 0.5: partially match, 1.0: perfect match
 - Check whether it is a believable method
 - Computed the correlation of GPT-4's scores with the average score across four independent annotators (i.e. humans)
 - 👉 High Pearson correlation of 0.80
 - (Consistent with prior findings that LLM can align well with human evaluations)
 - **Metric.** ACC@K
 - The highest score of any of the top-k proposals

Our method: VisDiff

Two-stage framework. Proposer + Ranker

Roles

- **Proposer.**
 - Takes random subsets $S_A \subseteq D_A$ and $S_B \subseteq D_B$ + Proposes differences
- **Ranker.**
 - Takes these proposed differences
 - Evaluates them across all of D_A and D_B to assess which ones are most true

Our method: VisDiff

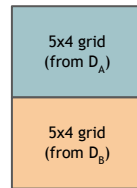
Proposer.

- $(S_A, S_B) \rightarrow$ proposed differences $'y' = (y_1, y_2, \dots)$
 - Authors set $|S_A| = |S_B| = 20$
 - Proposed differences $'y'$. (Ideally) more true on S_A than S_B
 - Running the proposer **multiple times** over different sampled sets
 - Take the **union of the proposed differences**
 - Leverage visual language models (VLM) in **three different ways**
 - Image-based
 - Feature-based
 - Caption-based

Our method: VisDiff

Proposer.

- Three different ways
 - Image-based (Use LLaVA-1.5 as VLM)
 - Arrange images into a single 4-row, 10-column grid
 - Top half of images (20 images) + bottom half of images (20 images)
 - Prompt the VLM to propose differences between the top and bottom half of images
 - Feature-based (Use BLIP-2 (consist of encoder, language model) as VLM)
 - Subtract the mean embeddings of S_A and S_B
 - $\text{mean}(\text{VLM_encoder}(S_A)) - \text{mean}(\text{VLM_encoder}(S_B))$
 - Subtracted embedding is fed into VLM's language model to generate the difference
 - Caption-based (Use BLIP-2 as VLM)
 - Generate captions from S_A and S_B using VLM
 - Generate differences from the captions by using GPT-4



Our method: VisDiff

Ranker.

- Mark the rank of proposed differences ‘y’
 - Validate and rank the proposed differences y
 - Sorts proposed differences by computing a difference score s_y
 - $s_y = \mathbb{E}_{x \in \mathcal{D}_A} v(x, y) - \mathbb{E}_{x \in \mathcal{D}_B} v(x, y)$
 - $v(x, y)$. Measure of how well the image ‘x’ satisfies the proposed differences ‘y’
 - Running t-test on the two score distributions
 - for filtering out proposed differences that are not statistically significant
- Leverage VLMs to compute the ranking score $v(x, y)$ in three ways
 - Image-based
 - Feature-based
 - Caption-based

Our method: VisDiff

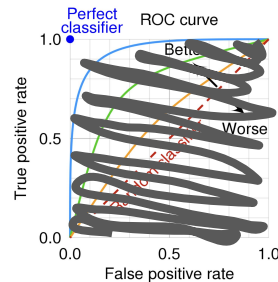
Ranker.

- Three different ways
 - Image-based (Use LLaVA-1.5 as VQA model 👉 Binary output)
 - Query the VQA model to ask whether the image contains differences
 - $v(x, y) = VQA(x, y)$
 - Caption-based (Use BLIP-2 as VLM, Vicuna-1.5 as QA model 👉 Binary output)
 - Generate captions 'c' from x using VLM → Ask Vicuna-1.5 whether the c contains y
 - $v(x, y) = QA(c, y)$

Our method: VisDiff

Ranker.

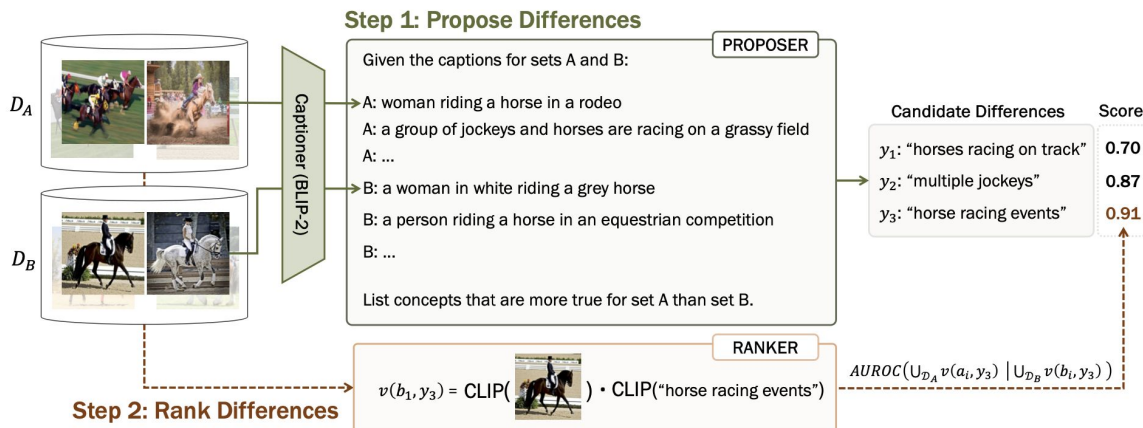
- Three different ways
 - Feature-based (Use CLIP ViT-G/14 as embedding model)
 - Compute the cosine similarity
 - between the image embedding e_x and text embedding e_y
 - $v(x, y) = \frac{e_x \cdot e_y}{\|e_x\| \|e_y\|}$
 - Since it is continuous,
we compute s_y as the AUROC of using v to classify between D_A and D_B
 - AUROC = Area Under ROC Curve
 - $s_y = AUROC(U_{D_A} v(a_i, y) \mid U_{D_B} v(b_i, y))$



Our method: VisDiff

Best combination. Caption-based Proposer + Feature-based Ranker

- Set Caption-based Proposer, Feature-based Ranker as ‘main method’
 - Others are considered as **baselines**



Results

Which proposer/ranker works best?

Results on VisDiffBench

- Caption-based proposer outperforms others
- Feature-based Ranker outperforms others

| Proposer | Ranker | PIS-Easy | | PIS-Medium | | PIS-Hard | | ImageNet-R/* | |
|-----------------|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|
| | | Acc@1 | Acc@5 | Acc@1 | Acc@5 | Acc@1 | Acc@5 | Acc@1 | Acc@5 |
| LLaVA-1.5 Image | CLIP Feature | 0.71 | 0.81 | 0.39 | 0.49 | 0.28 | 0.43 | 0.27 | 0.39 |
| BLIP-2 Feature | CLIP Feature | 0.48 | 0.69 | 0.13 | 0.33 | 0.12 | 0.23 | 0.68 | 0.85 |
| GPT-4 Caption | Vicuna-1.5 Caption | 0.60 | 0.92 | 0.49 | 0.77 | 0.31 | 0.61 | 0.42 | 0.70 |
| GPT-4 Caption | LLaVA-1.5 Image | 0.78 | 0.99 | 0.58 | 0.80 | 0.38 | 0.62 | 0.78 | 0.88 |
| GPT-4 Caption | CLIP Feature | 0.88 | 0.99 | 0.75 | 0.86 | 0.61 | 0.80 | 0.78 | 0.96 |

Results

Which **proposer** works best?: Analysis

Caption-based proposer. Image → Caption → Difference

- “Image → Caption” may result in some loss of information
- However...
 - Strong reasoning capabilities of LLM (i.e. GPT-4) effectively compensate for this by identifying diverse and nuanced differences between image sets

Results

Which **ranker** works best?: Analysis

Feature-based ranker. Difference → Features → Score

- Score = Similarity 🙌 Continuous value
 - Others (Image and Caption-based). score is **binary value** (contain, not contain)
 - Continuous value allows for **more fine-grained image annotation and improved calibration**

Results

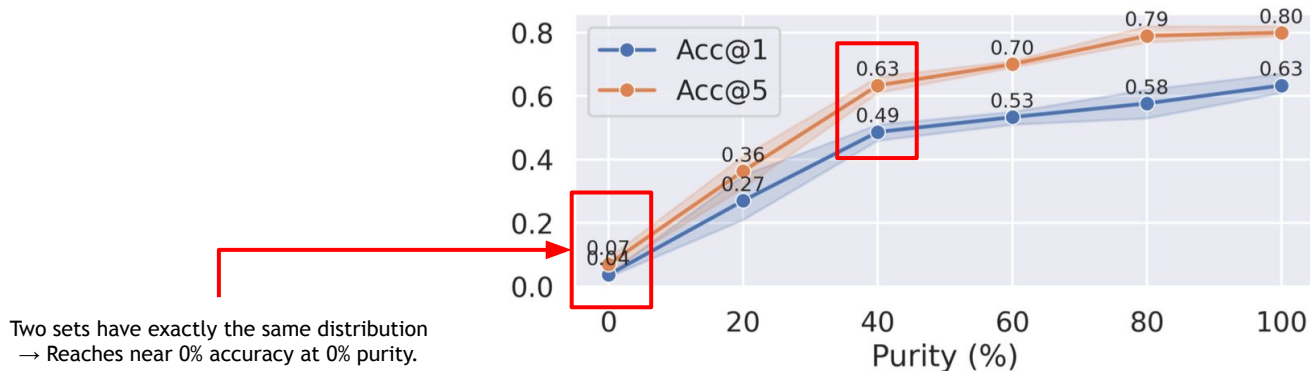
Performance Under Noisy Data Splits

- 100% of purity image set is rare in real-world
 - Example of a 100% of purity image set. Only consists of same type images
- Experiment. Swapping a certain % of images between DA and DB
 - Purity 0%. 50% image swapping and an equal distribution of two sets
 - Purity 100%. no image swapping

Results

Performance Under Noisy Data Splits

- **Results.** Tested on 50 paired sets within PairedImageSets-Hard
 - Decline in purity correlates with a drop in accuracy
 - Since identifying the difference becomes harder
 - Even at 40% purity, Acc@1 remains at 49%, only reduced from 63% at 100% purity (= Our method has robustness!)



Applications

ImageNetV2 vs. ImageNet

- Discover **more interpretable differences** between two datasets
- Time shift.
 - ImageNet. Collected prior to 2012
 - ImageNetV2. Collected between 2012 and 2014

👉 Discovered by using VisDiff (e.g. Instagram, Whatsapp)

| Class | More True for ImageNetV2 |
|--------------------|------------------------------------|
| Dining Table | People posing for a picture |
| Wig | Close up views of dolls |
| Hand-held Computer | Apps like Twitter and Whatsapp |
| Palace | East Asian architecture |
| Pier | Body of water at night |
| Schnauzer | Black dogs in different settings |
| Pug | Photos possibly taken on Instagram |
| Horizontal Bar | Men's gymnastics events |

ImageNetV2



ImageNet

Applications

CLIP ViT-H vs. ResNet-50

Want to know. CLIP's behavior compared with ResNet

- **Settings.**

- Data. Top 5 classes in ImageNet where CLIP outperforms ResNet
- D_A . Images correctly identified by CLIP but not by ResNet
- D_B . All other images

- **Analysis.**

- CLIP shows better performance at the image with presence of people, label

| Class | Acc_C | Acc_R | More Correct for CLIP |
|---------------|---------|---------|--|
| Tobacco Shop | 0.96 | 0.50 | Sign hanging from the side of a building |
| Digital Watch | 0.88 | 0.52 | Watches displayed in a group |
| Missile | 0.78 | 0.42 | People posing with large missiles |
| Pot Pie | 0.98 | 0.66 | Comparison of food size to coins |
| Toyshop | 0.92 | 0.60 | People shopping in store |

← Results

Applications

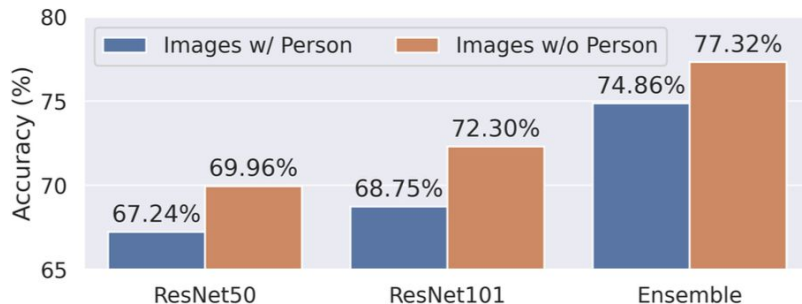
Finding Failure modes of ResNet

- Identify failure modes of a model by contrasting images...
 - that are **correctly predicted against those that are erroneously classified**
- **Settings.**
 - D_A . ImageNet images misclassified by both ResNet-50 and ResNet- 101
 - D_B . Correctly classified images

Applications

Finding Failure modes of ResNet

- Results.
 - Proposed differences. “Humanized object items”, “People interacting with objects”
- Analysis.
 - ResNet models perform worse when the images include human subjects
 - Same observation when measuring Accuracy



Applications

Stable DiffusionV2 (SDv2) vs. Stable DiffusionV1 (SDv1)

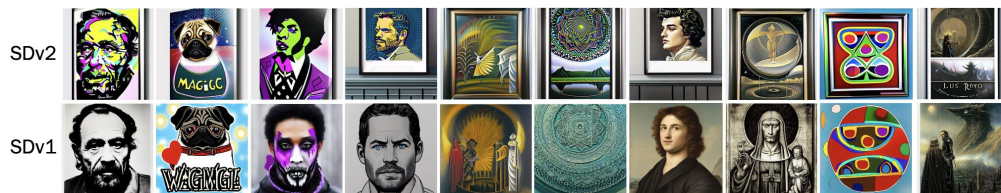
- **Settings.**

- D_A . Generated images from SDv2
- D_B . Generated images from SDv1

- **Results.**

- Proposed differences “vibrant and contrasting colors”, “images with frames or borders”
- Confirm “vibrant and contrasting colors” by computing the avg. saturation
 - (Prompt in PartiPrompts) SDv2: **112.61**, SDv1: 110.45
 - (Prompt in DiffusionDB) SDv2: **97.96**, SDv1: 93.49

👉 Same observation when using VisDiff



👉 Qualitative results when using the prompt in DiffusionDB

Applications

Describing Memorability in Images

Use VisDiff in addressing diverse real-world questions

- **Settings.**

- Dataset. **LaMem** (Large-scale Image Memorability) dataset
 - Assigned a memorability score by humans
- D_A . Most memorable 25th percentile
- D_B . Least memorable 25th percentile

- **Results.**

- Most memorable images “presence of humans”, “humorous settings”
- Least memorable images “landscapes”, “urban environments”

👉 These findings are consistent with those of Isola et al.

Conclusion

Authors introduce...

- The task “**Difference Captioning**”
- **VisDiff** that is designed to identify and describe differences in sets
- Show VisDiff’s utility in finding interesting insights across a variety of real-world applications

My opinion. Tools for finding new research direction?

- Describe the differences in natural language form
 - Human can understand this differences