Thematic session

Paper presentation: Group4

Wangyuan Ding

Large Concept Models:

Language Modeling in a Sentence Representation Space

LCM team, Loïc Barrault*, Paul-Ambroise Duquenne*, Maha Elbayad*, Artyom Kozhevnikov*, Belen Alastruey†, Pierre Andrews†, Mariano Coria†, Guillaume Couairon+†, Marta R. Costa-jussà†, David Dale†, Hady Elsahar†, Kevin Heffernan†, João Maria Janeiro†, Tuan Tran†, Christophe Ropers†, Eduardo Sánchez†, Robin San Roman†, Alexandre Mourachko‡, Safiyyah Saleem‡, Holger Schwenk‡

FAIR at Meta

*Core contributors, alphabetical order, †Contributors to data preparation, LCM extensions and evaluation, alphabetical order, ‡Research and project management, alphabetical order, +Initial work while at FAIR at Meta, new affiliation: INRIA, France

Large Language Models

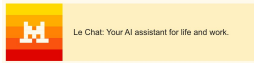
Large Language Models (LLMs) are dominating current research in natural language processing and extending to more other modalities (images, video, speech).

However, LLMs are heavily data-driven, and building an LLM from scratch requires access to enormous computational resources to process ever larger amounts of data and train models, the size of which now exceeds **four hundred billion** parameters. Extending to other modalities even need more (synthetic data). **Scaling law** is inevitably hitting the limit.

Open source:

Llama

Mistral



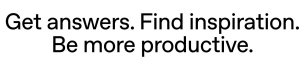
Bloom

Falcon

Close source:

Gemini

GPT



Claude.

Token-Free LLMs

LLMs, despite different model architecture and performance, are mostly based on Transformer decoder-only architecture, pre-trained on next **token** prediction task.

However, all current LLMs miss a crucial characteristic of human intelligence: **explicit reasoning** and planning at multiple levels of **abstraction**. The human brain does not operate at the word level only.

Imagine a researcher giving a fifteen-minute talk.

In such a situation, researchers do not usually prepare detailed speeches by writing out every single word they will pronounce. Instead, they **outline a flow of higher-level ideas** they want to communicate.

Should they give the same talk multiple times, the actual words being spoken may differ, the talk could even be given in different languages, but **the flow of higher-level abstract** ideas will remain the same.

Large Concept Models

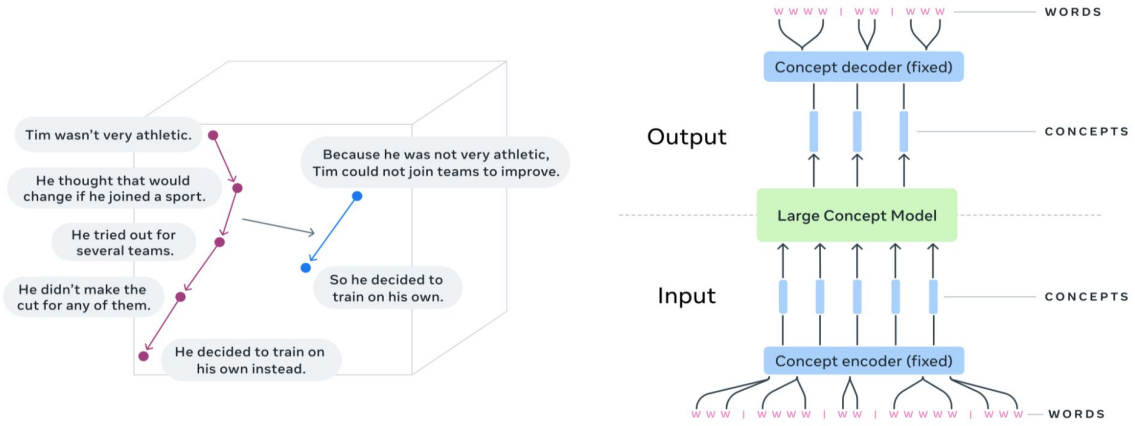


Figure 1 - Left: visualization of reasoning in an embedding space of concepts (task of summarization). Right: fundamental architecture of an LARGE CONCEPT MODEL (LCM). ⋆: concept encoder and decoder are frozen.

The input is first segmented into sentences, and each one is encoded to achieve a sequence of concepts, i.e., sentence embeddings.
This sequence of concepts is then processed by the LCM to generate at the output a new sequence of concepts.
Finally, the generated concepts are decoded into a sequence of subwords. The encoder and decoder are fixed and are not trained.

Large Concept Models

Core: concept-based understanding.

Hypothesis: Humans processing and generate language not based on single word (word-level), but rather higher level unit (phrases, sentences, even paragraph) and reason based on these.

Approach: (hierarchical) reasoning in an abstract embedding space, such as subword tokens, concepts, short description of a paragraph, and small section.

Concept: an abstract atomic idea. In practice, a concept would often correspond to a sentence in a text document, or an equivalent speech utterance.

Results: TL;DR

Better multilingual and multi-modal task performance: Multilingual translation, Cross-modality generation (text-to-speech, speech-to-text)

Cross-lingual knowledge: Multilingual Knowledge QA, Information retrieval in Low-resource languages

Why: model concept in a unified abstract semantic space, not depend on token distribution of a certain language.

Long document generate and processing: Technical Writing, Long Document Summarization, Complex Story Generation.

Creativity in Generation: Creative Writing, Conversational AI.

Why: Better inference in concept space, more like human.

Take Away

1. Concept is closer to the actual semantics compared to word or token, this reduce irrelevant signal from lower dimention, and focus on higher level semantics.
2. Modularity and extensibility: concept encoders and decoders can be independently developed and optimized without any competition or interference.
3. A concept can be translate to multiple languages or speech signal modality, no re-training or inference is needed.

	Text		Speech		Image		Video	
Model	Input	Output	Input	Output	Input	Output	Input	Output
GEMINI	47	47	62	✓	✓	✓	✓	✗
GPT	85	85	✓	✓	✓	✓	?	✗
CLAUDE	37	37	✓	✓	✓	✓	✗	✗
BLOOM	46	46	✗	✗	✓	✓	✗	✗
LLAMA 3-400B	8	8	34	✗	✓	✓	✗	✗
LCM-SONAR	200	200	76	1	✗	✗	(ASL)	✗

Table 1 - Comparison of language and modality coverage for several LLMs and our LCM operating on the SONAR embedding space. SONAR has an experimental support for American Sign Language (ASL) which is not used in this paper.

Carlo Bretti

Published as a conference paper at ICLR 2025

CAPTURED BY CAPTIONS: ON MEMORIZATION AND ITS MITIGATION IN CLIP MODELS

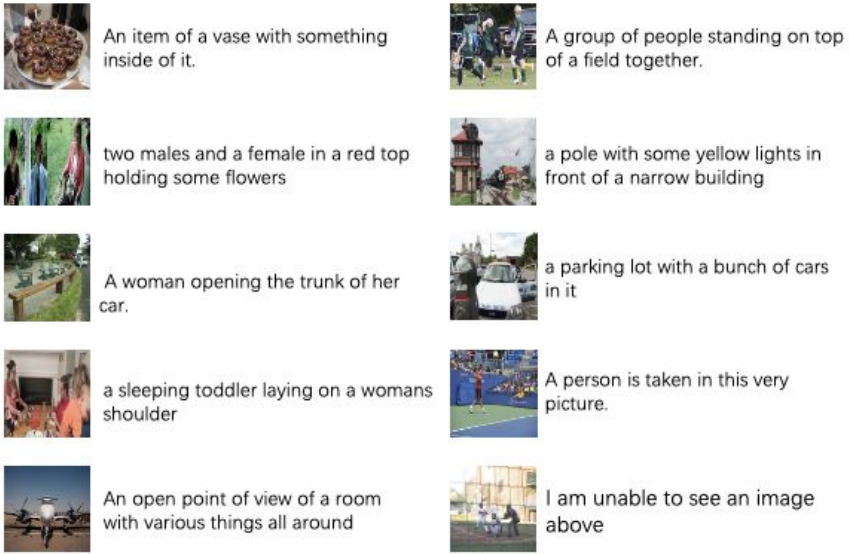
Wenhao Wang[1], **Adam Dziedzic**[1], **Grace C. Kim**[2], **Michael Backes**[1], **Franziska Boenisch**[1]*
[1]CISPA, [2]Georgia Institute of Technology

How to quantify memorization in CLIP?

The paper takes two models, *f* and *g*, where f is trained on the whole set S of image-text pairs, and g on S’ defined as S minus a single image-text pair.

The paper defines a alignment score based on how close the image and text representations for a single datapoint and how distant they are from other unseen images or pieces of text.

Finally, they define a score named CLIPMem for an image-text pair based on the difference in alignment between the two models.

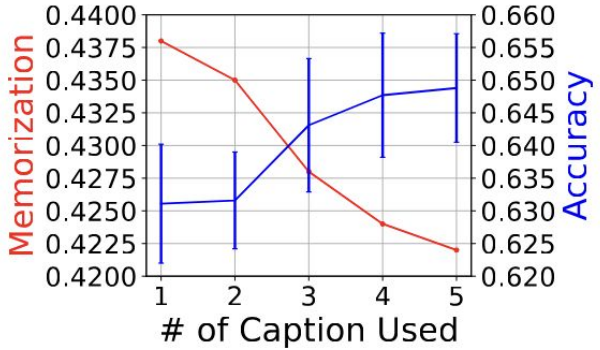


(a) Most Memorized: CLIPMem > 0.89

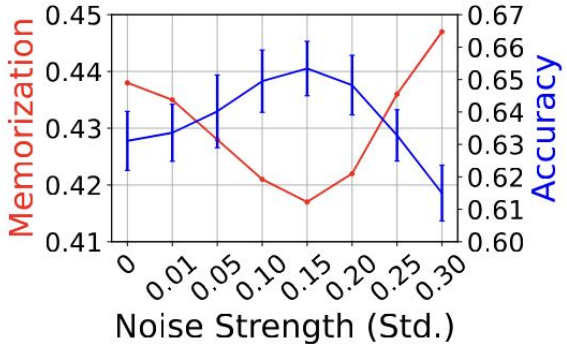
(b) Least Memorized: CLIPMem ≈ 0.0

Figure 1: **Examples of data with different levels of memorization.** Higher memorization scores indicate stronger memorization. We observe that atypical or distorted images, as well as those with incorrect or imprecise captions, experience higher memorization compared to standard samples and easy-to-label images with accurate captions. Results are obtained on OpenCLIP (Ilharco et al., 2021), with encoders based on the ViT-Base architecture trained on the COCO dataset.

In their experiments, they find that the memorization mostly happens within the text encoder, so they come up with some strategies for mitigation



(a) Different numbers of captions.



(b) Noising text embedding during training.

Figure 5: **Mitigating memorization in CLIP improves downstream generalization.** We train CLIP models with different "augmentations" in the textual domain. (a) We use multiple captions for the same image during training. (b) We directly noise the text embeddings during the training using Gaussian noise with a mean of 0 and different standard deviations (adding the Gaussian noise $\mathcal{N}(0, 0.15)$ gives us the sweet spot with the smallest memorization and highest performance). Both strategies successfully reduce memorization while improving performance.

Main takeaways

- We can define a score for memorization for multimodal models
- “CLIP highly memorizes data points with **incorrect and imprecise captions,** much like supervised models memorize mislabeled samples, but it also memorizes **atypical examples**”
- Memorization mostly happens in the text encoder - mitigating strategies can both reduce memorization and improve downstream accuracy

Floris Gisolf

MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2024.2418937

MEDICAL TEACHER
Taylor & Francis
Taylor & Francis Group

NEW WAVE

Check for updates

Pedagogy and generative artificial intelligence: Applying the PICRAT model to Google NotebookLM

Neil Mehta[a], Anoop Agrawal[b], Jennifer Benjamin[c], Seysha Mehta[d], Heather MacNeill[e] and Ken Masters[f]

[a]Professor of Medicine and Associate Dean for Curricular Affairs, Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland, OH, USA; [b]Department of Medicine, Baylor College of Medicine, Houston, TX, USA; [c]Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA; [d]Class of 2027, Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland, OH, USA; [e]Associate Professor, University of Toronto and Toronto Metropolitan University, Toronto, Canada; [f]Associate Professor of Medical Informatics, College of Medicine and Health Sciences, Sultan Qaboos University, Oman

ABSTRACT

Healthcare educators (HPE) are challenged by rapid developments in Generative Artificial Intelligence (GenAI) tools. They need a standardized model to evaluate these new tools and to guide them in pedagogically-sound integration in the curriculum. PICRAT is an educational model designed specifically to help teachers meet this challenge. NotebookLM is a new multi-featured GenAI tool to help teachers and learners in education and research. Its newest feature allows automatic generation of an engaging podcast (called audio overview) from uploaded education or research content. Using the example of NotebookLM and, specifically, the auto-podcast feature, we illustrate how HPE can use the PICRAT model to evaluate GenAI tools for technology integration. We discuss how this model can be utilized as a standardized approach for evaluation and implementation of GenAI tools in health professions education.

ARTICLE HISTORY
Received 22 September 2024
Accepted 16 October 2024

KEYWORDS
PICRAT; Google NotebookLM; podcasts; ChatGPT; Generative AI

The PICRAT Model for Technology Integration in Teacher Preparation

Models engagement with tech and how tech alters teaching

PIC (Learner's Role):

Passive – Students consume content (e.g., listening to an AI-generated podcast).
Interactive – Students engage with content (e.g., discussing an AI-generated summary).
Creative – Students generate content (e.g., making their own AI-generated podcasts).

RAT (Teaching Change):

Replacement – AI replaces traditional tasks without teaching enhancement.
Amplification – AI enhances traditional teaching by increasing efficiency.
Transformation – AI fundamentally changes the learning process.

Passive & Transformative (PT) – An educator generates a podcast and assigns it as pre-class material.

Interactive & Transformative (IT) – Students take notes on the podcast and engage in a think-pair-share activity.

Creative & Amplifying (CA) – Students generate their own podcast and critique its accuracy.

Share Notebook with students (CT) – Teacher creates a NotebookLM and generates an audio overview and asks students to review the material and engage in a chat with the notebook to delve deep into the study materials and identify a question that they have that is not answered by the content.

Use NotebookLM to teach students something beyond the understanding of the teacher, then have the student present it to the class.

The PICRAT Model for Technology Integration in Teacher Preparation

Royce Kimmons, Charles R. Graham, & Richard E. West
Brigham Young University

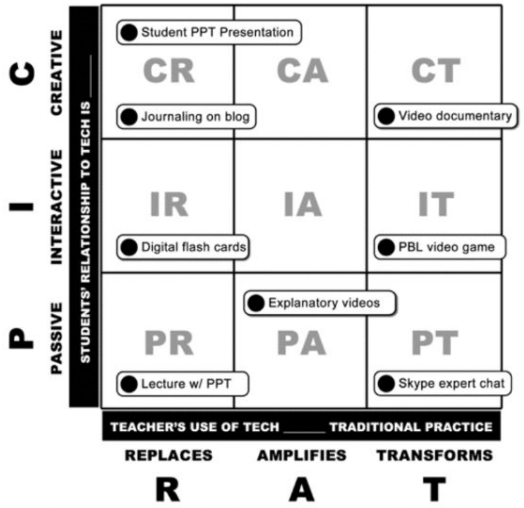
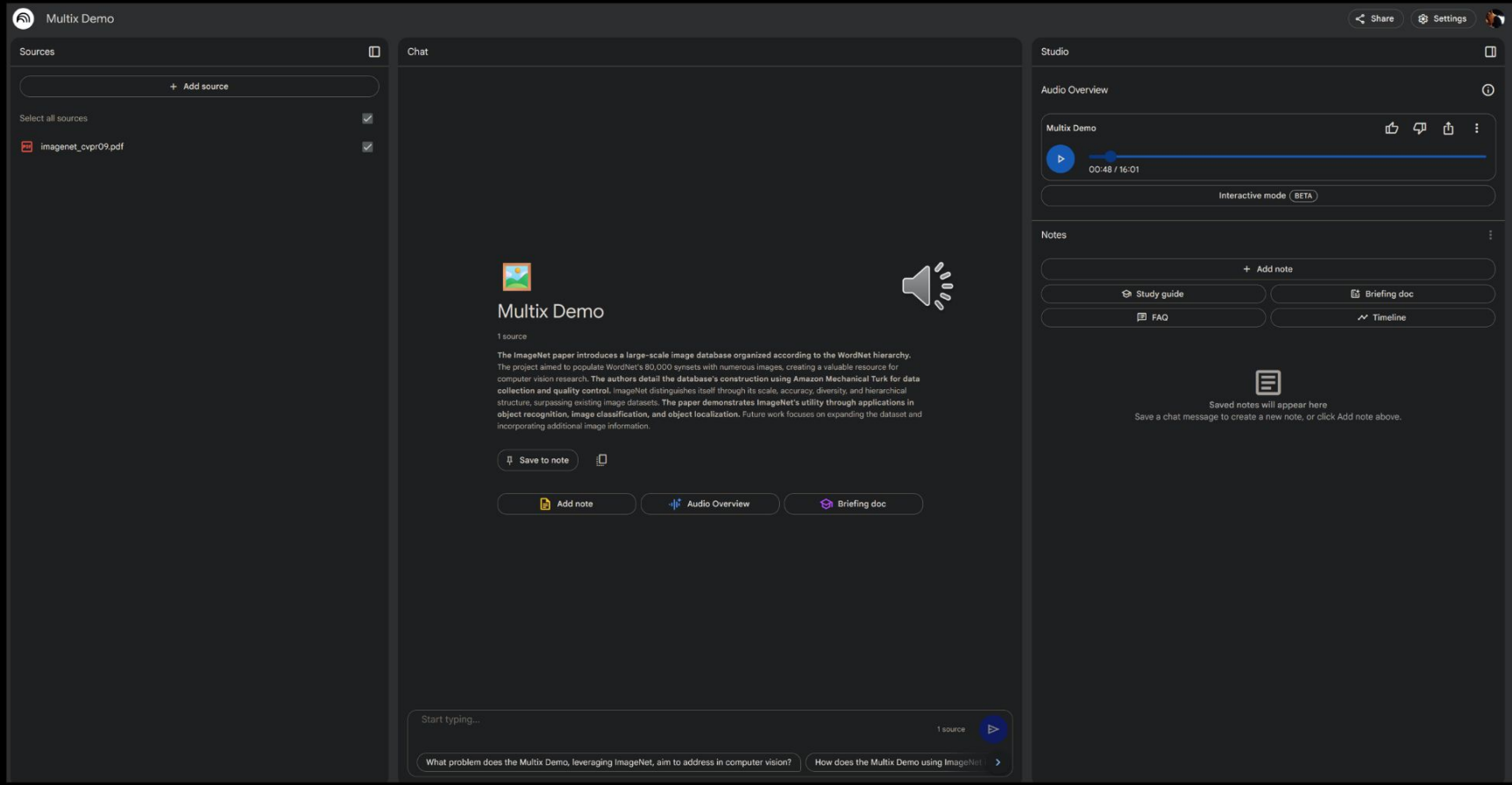


Figure 6. *An example of unit activities mapped to PICRAT.*


Multix Demo
Share
Settings
Sources
Add source
Select all sources
imagenet_cvpr09.pdf
Chat
Multix Demo
1 source
The ImageNet paper introduces a large-scale image database organized according to the WordNet hierarchy. The project aimed to populate WordNet's 80,000 synsets with numerous images, creating a valuable resource for computer vision research. The authors detail the database's construction using Amazon Mechanical Turk for data collection and quality control. ImageNet distinguishes itself through its scale, accuracy, diversity, and hierarchical structure, surpassing existing image datasets. The paper demonstrates ImageNet's utility through applications in object recognition, image classification, and object localization. Future work focuses on expanding the dataset and incorporating additional image information.
Save to note
Add note
Audio Overview
Briefing doc
Start typing...
1 source
What problem does the Multix Demo, leveraging ImageNet, aim to address in computer vision?
How does the Multix Demo using ImageNet
Studio
Audio Overview
Multix Demo
00:48 / 16:01
Interactive mode
BETA
Notes
Add note
Study guide
Briefing doc
FAQ
Timeline
Saved notes will appear here
Save a chat message to create a new note, or click Add note above.

Ivona Najdenkoska

DRCT: Diffusion Reconstruction Contrastive Training towards Universal Detection of Diffusion Generated Images

Baoying Chen [*1] Jishen Zeng [*1] Jianquan Yang [2] Rui Yang [1]

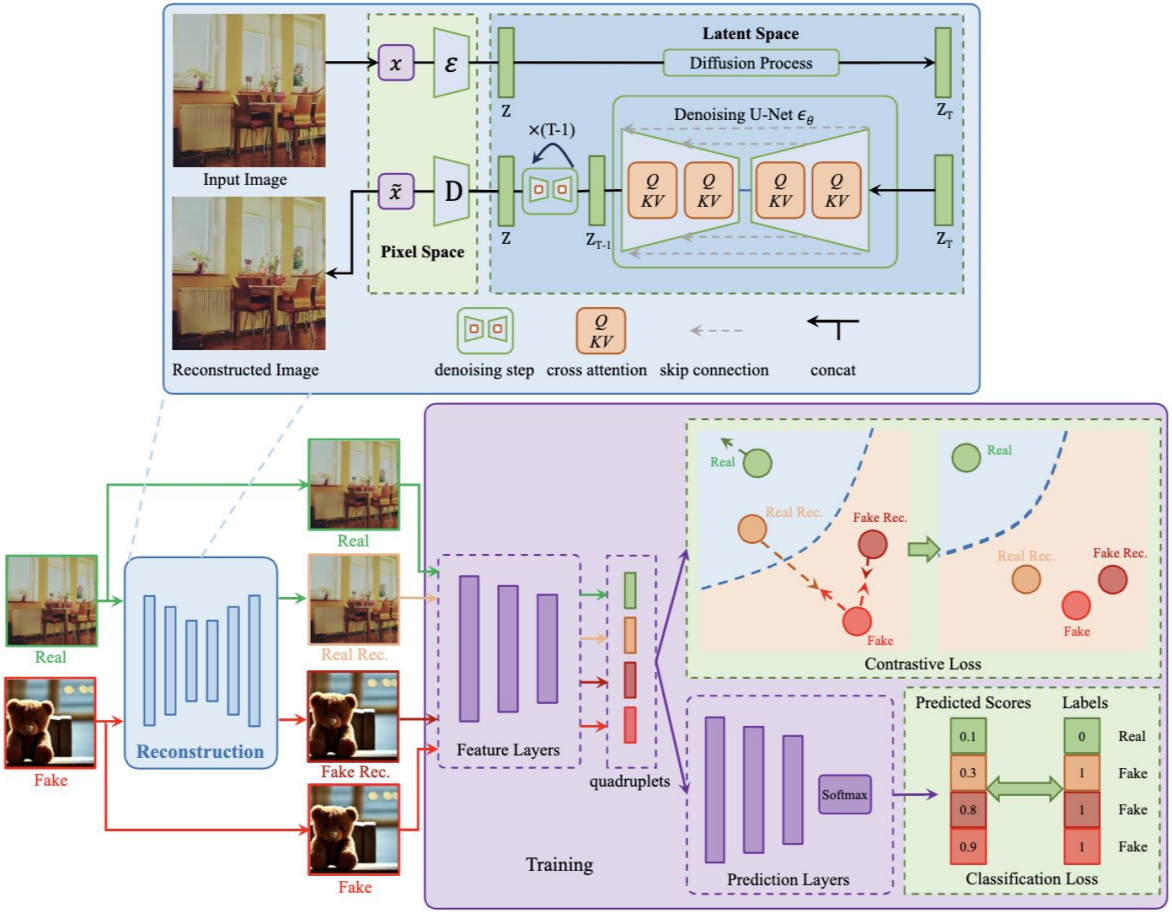
ICML 2024

Intro

- The diffusion reconstruction on real images can preserve the image content while leaving the ***fingerprint*** of the diffusion model on the output images.
- These reconstructed images can serve as ***informative yet hard samples*** for detectors to learn the subtle differences between real and generated images.

- This paper proposes a novel training framework named **Diffusion Reconstruction Contrastive Training (DRCT).**
- DRCT significantly improves the detection accuracy and generalization ability of diffusion-generated image detectors.

Diffusion Reconstruction Contrastive Training (DRCT)



DRCT consists of a reconstruction stage and a training stage:

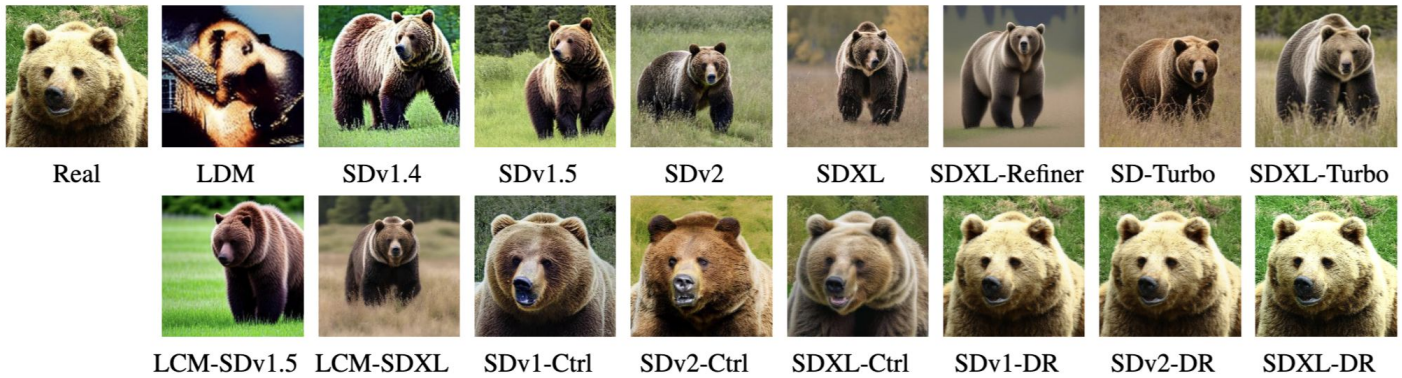
1. **Reconstruction stage:** a large number of images are produced by reconstructing both real and generated image using selected diffusion models, which are then prepared for the training of the classifier.
2. **Training stage**: 4 types of samples: real images, real reconstructed images, fake images, and fake reconstructed images, are used for computing the contrastive loss and the classification loss.

DRCT-2M Dataset

- Collection of 2 million images for training and evaluation. It consists of two parts:
  - Images automatically generated by diffusion-based models (prompts are derived from the MSCOCO)
  - Images collected from real-world scenarios (Midjourney and CIVITAI)

The **DRCT-2M** dataset involves ***16 types of stable diffusion models***, including LDM, SDv1.4, SDv1.5, SDv2, SDXL, SDXL-refiner, SD-Turbo, SDXL-Turbo, LCM-SDv1.5, LCM-SDXL, SDv1-Ctrl, SDv2-Ctrl, SDXL-Ctrl, SDv1-DR, SDv2-DR and SDXL-DR.

The prompt used for image generation is "*A big burly grizzly bear is shown with grass in the background.*"



Some experimental details

Data: The compared methods are trained on the DRCT-2M dataset (utilizing real images from MSCOCO) and the GenImage.

Evaluation metric: Accuracy (ACC) as the metric to evaluate detection performance, using a threshold of 0.5.

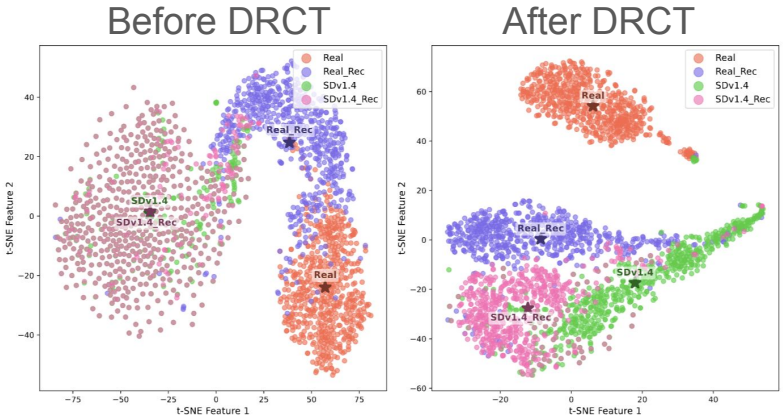


Table 1. Accuracy (ACC, %) comparisons of our DRCT and other generated image detectors on DRCT-2M. Except for DIRE and DRCT, all methods are only trained on SDv1.4 and then evaluated on different testing subsets on DRCT-2M. For the training data of DIRE and DRCT, when the Diffusion Reconstructed (DR) model is SDv1, the original fake images were generated by SDv1.4. When the DR model is SDv2, the original fake images were generated by SDv2.

Method	DR	SD Variants						Turbo Variants		LCM Variants		ControlNet Variants			DR Variants			Avg.
		LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	SDv1-DR	SDv2-DR	SDXL-DR	
CNNSpot	-	99.87	99.91	99.90	97.55	66.25	86.55	86.15	72.42	98.26	61.72	97.96	85.89	82.84	60.93	51.41	50.28	81.12
F3Net	-	99.85	99.78	99.79	88.66	55.85	87.37	68.29	63.66	97.39	54.98	97.98	72.39	81.99	65.42	50.39	50.27	77.13
CLIP/RN50	-	99.00	99.99	99.96	94.61	62.08	91.43	83.57	64.40	98.97	57.43	99.74	80.69	82.03	65.83	50.67	50.47	80.05
GramNet	-	99.40	99.01	98.84	95.30	62.63	80.68	71.19	69.32	93.05	57.02	89.97	75.55	82.68	51.23	50.01	50.08	76.62
De-fake	-	92.1	99.53	99.51	89.65	64.02	69.24	92.00	93.93	99.13	70.89	58.98	62.34	66.66	50.12	50.16	50.00	75.52
Conv-B	-	**99.97**	**100.0**	**99.97**	95.84	64.44	82.00	80.82	60.75	99.27	62.33	99.80	83.40	73.28	61.65	51.79	50.41	79.11
UnivFD	-	98.30	96.22	96.33	93.83	91.01	93.91	86.38	85.92	90.44	88.99	90.41	81.06	89.06	51.96	51.03	50.46	83.46
DIRE	SDv1	98.19	99.94	99.96	68.16	53.84	71.93	58.87	54.35	**99.78**	59.73	99.65	64.20	59.13	51.99	50.04	49.97	71.23
DIRE	SDv2	54.62	75.89	76.04	**99.87**	59.90	93.08	**99.77**	57.55	87.29	72.53	67.85	99.69	64.40	49.96	52.48	49.92	72.55
DRCT/Conv-B (ours)	SDv1	99.91	99.90	99.90	96.32	83.87	85.63	91.88	70.04	99.66	78.76	**99.90**	95.01	81.21	**99.90**	95.40	75.39	90.79
DRCT/Conv-B (ours)	SDv2	99.66	98.56	98.48	99.85	96.10	**98.68**	99.59	83.30	98.45	93.78	96.68	**99.85**	**97.66**	93.91	**99.87**	**90.39**	**96.55**
DRCT/UnivFD (ours)	SDv1	96.74	96.26	96.33	94.89	**96.24**	93.46	93.43	92.94	91.17	95.01	95.60	92.68	91.95	94.10	69.55	57.43	90.49
DRCT/UnivFD (ours)	SDv2	94.45	94.35	94.24	95.05	95.61	95.38	94.81	**94.48**	91.66	**95.54**	93.86	93.48	93.54	84.34	83.20	67.61	91.35

Takeaways

- The paper proposes a universal framework - **Diffusion Reconstruction Contrastive Training (DRCT)**, for enhancing the generalizability of existing methods for detecting diffusion-based generated images.

- While DRCT also boosts the detection accuracy for GAN-generated images, the improvement is not as ***significant***.
- This difference is mainly due to the ***unique generative artifacts*** produced by GANs versus those produced by diffusion-based methods - opportunity for future work :)