Project Title

UNIVERSAL VISION-LANGUAGE DENSE RETRIEVAL: LEARNING A UNIFIED REPRESENTATION SPACE FOR MULTI-MODAL RETRIEVAL （ICLR,2023）

Team Members

Mingjun Wen
Xuqi Zhu
Dongjing Xie

Github link：
baomu123/IITLearning-CS577-group1: The IIT CS577 deep learning project (github.com)
(https://github.com/baomu123/IITLearning-CS577-group1)

1. Introduction

The introduction of “**Universal Vision-Language Dense Retrieval**” marks a significant advancement in the realm of information retrieval, bridging the gap between visual and textual data comprehension. This study addresses the challenge of crafting a unified model adept at parsing and responding to a myriad of vision-language tasks. Pioneering in its approach, our team has formulated the method shown by the paper that amalgamates relevance modeling, cross-modality matching, and retrieval result fusion into a singular, streamlined process.This report proposes UniVL-DR, which constructs a unified multimodal vector representation space and combines single modal, cross modal matching and retrieval results modeling together to achieve end-to-end multimodal information retrieval.The experimental results show that UniVL-DR outperforms all baseline models by over 7% on performance metrics. The significant improvement in retrieval performance demonstrates the effectiveness of the proposed algorithm in constructing an information retrieval system for multimodal documents. It proves that the unified multimodal document vector modeling can well model multimodal retrieval tasks.

Mingjun Wen, as the project team leader, has been instrumental in synthesizing the conclusion and discussion sections, as well as orchestrating the presentation of the findings. Xuqi Zhu contributed to data collection, preprocessing, and dissecting the intricacies of the author’s code framework. Zhu also contributed expertise to reproducing the complicated mathematical formulas with LaTeX in report. Dongjing Xie lent his analytical skills to interpret the results, ensuring that the data spoke volumes about the model’s capabilities.

Due to the formidable size of the original dataset, a whopping 51GB, the team first processed text corpora and some image preprocessing, then commenced their subsequent image-related processing work based on the preprocessed data and training checkpoints provided by the original authors.

2. Problem Description

The confluence of vision and language understanding has led to the emergence of multimodal models capable of tackling tasks that require simultaneous comprehension of visual and textual content. The paper, “Universal Vision-Language Dense Captioning,” addresses the overarching challenge of unified multimodal retrieval, aiming to create a model capable of understanding and generating responses across different vision-language tasks.

Background

Although current mainstream search engines primarily target textual data, the growth of multimedia content has been one of the most prominent trends on the internet. Various studies indicate that users prefer vivid multimodal content in their search results. Consequently, information retrieval for multimodal data has become increasingly crucial in the user search experience.

Problem Statement

To facilitate the multimodal retrieval process, contemporary multimedia search systems typically adopt a ‘divide and conquer’ strategy. As illustrated in Figure 1(a), these methods first conduct searches within individual modalities, including text, images, videos, etc., and subsequently merge the retrieval results from different modalities. This is often accomplished by building an additional ranking module atop these single/cross-modality search engines to perform modality fusion. It is evident that the processes of relevance modeling and retrieval result fusion are typically intertwined to achieve more accurate multimodal search results. However, due to modality disparities, such models can only employ a piecemeal approach in pipeline modeling, making the fusion of retrieval results from distinct modalities challenging.

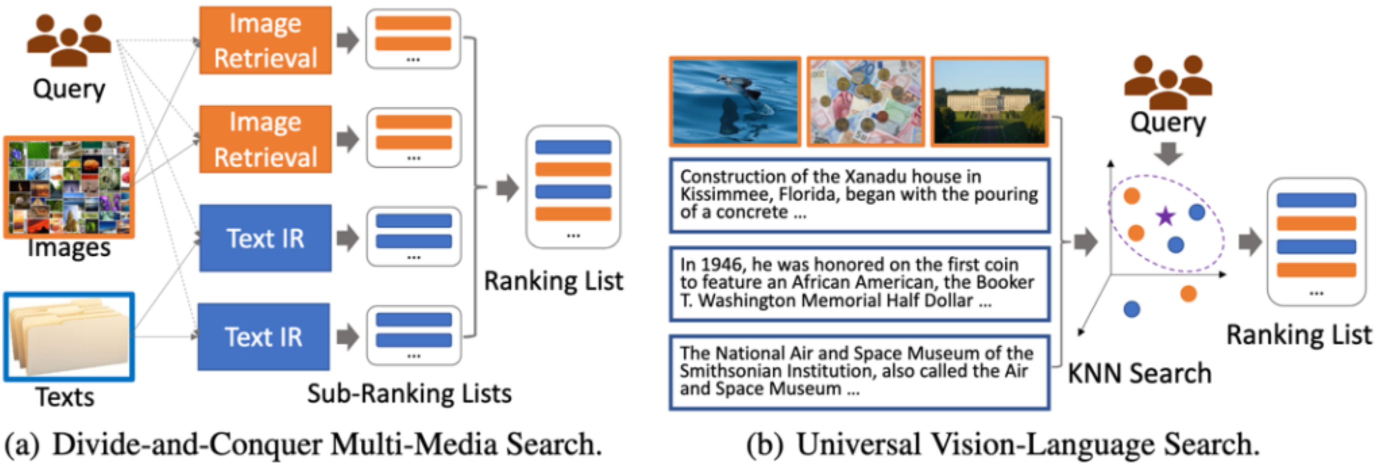
Query
Images
Texts
Image Retrieval
Image Retrieval
Text IR
Text IR
Sub-Ranking Lists
Ranking List
(a) Divide-and-Conquer Multi-Media Search.
Construction of the Xanadu house in Kissimmee, Florida, began with the pouring of a concrete ...
In 1946, he was honored on the first coin to feature an African American, the Booker T. Washington Memorial Half Dollar ...
The National Air and Space Museum of the Smithsonian Institution, also called the Air and Space Museum ...
Query
KNN Search
Ranking List
(b) Universal Vision-Language Search.

Figure 1: Different Architectures of Multi-Modal Retrieval Systems.

Methodology of UniVL-DR

In this paper, the authors introduce an end-to-end multimodal retrieval model that conducts unified retrieval on multimodal documents through user queries. As depicted in Figure 1(b), the universal multimodal retrieval maps both the query and the multimodal documents to a unified embedding space and retrieves multimodal candidates using nearest-neighbor search. Ultimately, this paper unifies the processes of relevance modeling, cross-modality matching, and retrieval result fusion in a single model as depicted in Figure 2.

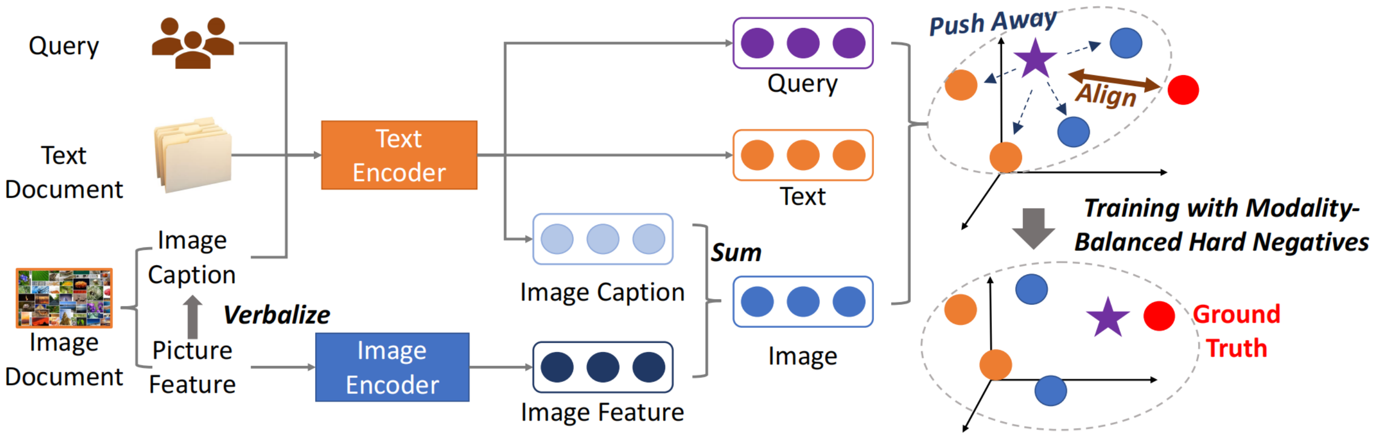

Figure 2: The Architecture of UniVL-DR.

The goal is not just accuracy but also versatility – the model should be equally adept at various vision-language tasks without requiring task-specific modifications.

TextEnocder and ImgEncoder

UniVL-DR gets representations of queries, image documents, and text documents with two encoders:TextEnocder and ImgEncoder. Specifically, the image document d_j^{Image} consists of a picture I_j and an image caption C_j, thus we utilize ImgEncoder and TextEnocder to encode I_j and C_j.

Query Encoding. UniVL-DR directly encodes the query q to get its representation $\vec{q}$:

$$\vec{q} = \text{TextEnocder}(q).$$

Text Document Encoding. To represent text documents, UniVL-DR also leverages the TextEnocder to encode the i-th text document d_i^{Text} as $\vec{d}_i^{\text{Text}}$:

$$\vec{d}_i^{\text{Text}} = \text{TextEnocder}(d_i^{\text{Text}}).$$

Image Document Encoding.Different from text documents, image documents can be represented by picture features and image captions, and the textual captions can help better understand the semantics of image documents (Baldrati et al., 2022). Thus, UniVL-DR encodes picture I_j and image caption C_j and then sums these embeddings to get the representation $\vec{d}_j^{\text{Image}}$ of the j-th image document:

$$\vec{d}_j^{\text{Image}} = \text{ImgEnocder}(I_j) + \text{TextEnocder}(C_j).$$

The representations $\vec{d}_j^{\text{Image}}$ and $\vec{d}_i^{\text{Text}}$ of image document and text document use the same TextEnocder to encode their textual information, which bridges different modalities in the text space and helps to build a universal embedding space for multi-modality retrieval.

Multi-modality Document Retrieval. The cosine similarity score $f(q, d)$ of query q and document candidate $d \in D$ can be calculated to estimate the relevance between q and d:

$$f(q, d) = \cos(\vec{q}, \vec{d}),$$

where $\vec{q}$ and $\vec{d}$ are the representations of q and d. The efficient similarity calculation between queries and the multi-modality documents can be provided by FAISS (Johnson et al., 2019).

Universal Represtation Learning

UniVL-DR utilizes the CLIP vision-language model to acquire universal representations that are effective for queries and documents across multiple modalities, particularly excelling in cross-modality retrieval. To enhance the universality of the embedding space, UniVL-DR strategically employs hard negatives that are balanced across different modalities during its training process. This approach is specifically designed to prevent the model from becoming overly biased towards single-modality signals when undergoing multi-modal co-training.

Given the query q and its relevant candidate $d^+ \in D$, the embedding space can be optimized by sampling hard negatives D^- and minimizing the following contrastive training loss L:

$$\begin{aligned} L &= -\log \frac{e^{f(q,d^+)/\tau}}{e^{f(q,d^+)/\tau} + \sum_{d^- \in D^-} e^{f(q,d^-)/\tau}} \\ &= -\frac{f(q,d^+)}{\tau} + \log\left(e^{f(q,d^+)/\tau} + \sum_{i=1}^{k_1} e^{f(q,d_i^{-\text{Image}})/\tau} + \sum_{j=1}^{k_2} e^{f(q,d_j^{-\text{Text}})/\tau}\right), \end{aligned}$$

we noted

$$L_{\text{Align}} = \frac{f(q,d^+)}{\tau}$$

$$L_{\text{Image}} = \sum_{i=1}^{k_1} e^{f(q,d_i^{-\text{Image}})/\tau}$$

$$L_{\text{Text}} = \sum_{j=1}^{k_2} e^{f(q,d_j^{-\text{Text}})/\tau}$$

where τ is the temperature to scale the similarity score. During training, we in fact maximize L_{Align} and minimize L_{Image} and L_{Text}, which make queries closer to related documents and away from unrelated documents. Our modality-balanced negative training strategy keeps $k_1 = k_2 = k$ to better train the modality selection ability of retrievers.

Description of the Data Used in the Project

The research primarily utilizes the WebQA dataset to evaluate and fine-tune the proposed model:

WebQA Dataset:

-Links: https://github.com/WebQnA/WebQA (https://github.com/WebQnA/WebQA)

- **Despcriptions:** WEBQA is a new multi-hop, multi-modal question answering challenge for our community.Designed to simulate the heterogeneous information landscape one might expect during a web search, WEBQA covers a series of opendomain general visual queries while also forcing models to still reason about text. Our task requires a system to determine relevant sources, perform aggregation and reasoning.We also propose a novel general recipe for evaluation on WEBQA which measures both fluency and accuracy.
- In total, WEBQA has over 34K training QA pairs, with an additional 5K and 7.5K held out for development and testing.Overall Statistics are summarized in Table 1 and language distributions are presented in Table 2 as below:

Modality	Train	Dev	Test
Image	18,954	2,511	3,464
Text	17,812	2,455	4,076

Table 1: Number of samples collected for each modality fold.

	Question	Answer	Correct	Distract	Correct	Distract
Image	16.4± 6	14.4± 6	13.3±11	12.6±11	—	36.4±10
Text	18.6± 8	10.7±10	—	14.1±13	45.3±12	38.3±10

Table 2. Length distribution for different textual components.

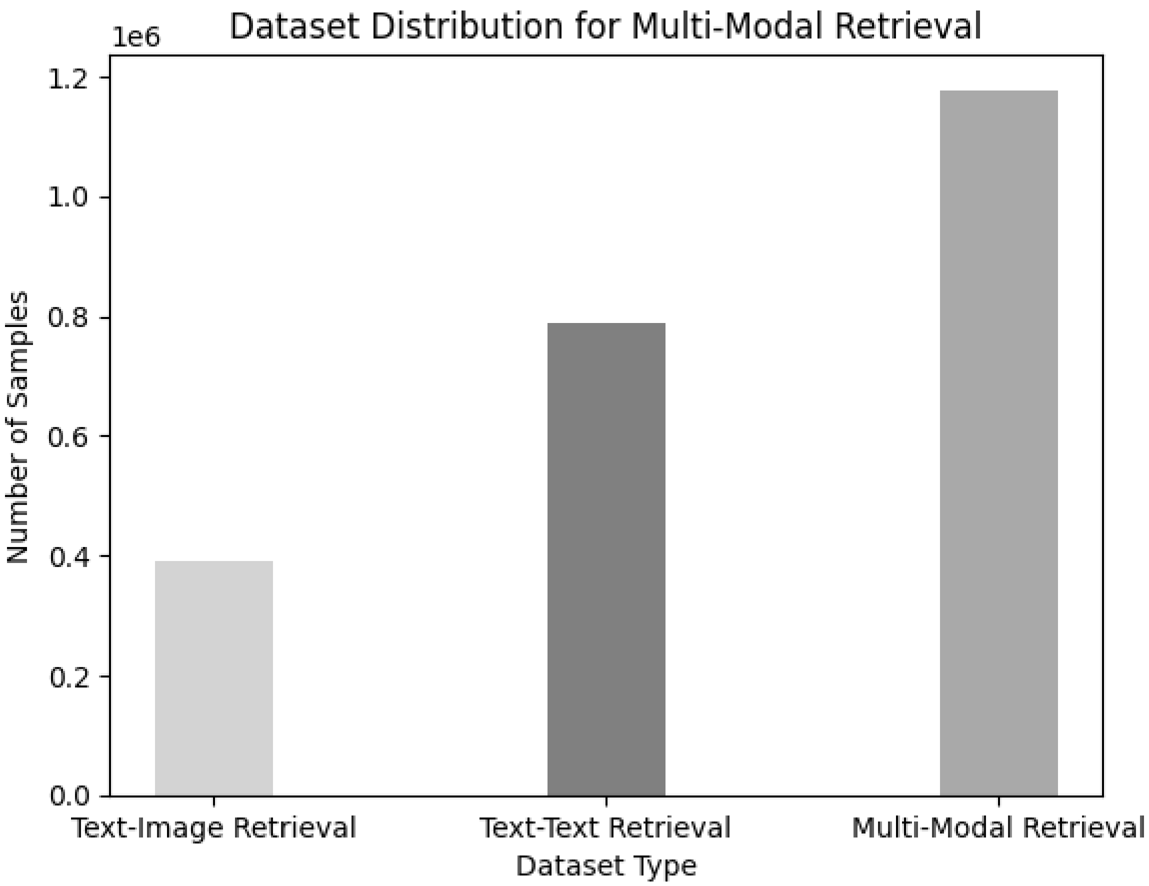

Figure 3: Data Distribution for Multi-Model Retrieval

Environment Setup:

- Python==3.7
- Pytorch
- transformers
- clip
- faiss-cpu==1.7.0
- tqdm
- numpy
- base64
- Install the `pytrec_eval` from `https://github.com/cvangysel/pytrec_eval`

Code Acquisition and Review:

- **Source:** https://github.com/OpenMatch/UniVL-DR (https://github.com/OpenMatch/UniVL-DR)
- **Components Reviewed:** The model's architecture, evaluation metrics.

Pipeline Execution:

- **Dataset Preprocessed:** Preprocess of dataset WebQA.
- **Train UniVL-DR:** UniVL-DR inherits CLIP (ViT-B/32). The texts must be truncated by 77 tokens and you can try different vision-language models. As shown in our experiments, we suggest to use the dual encoder models.There are two steps to train UniVL-DRR:
-
    - **First step:** Go to the `CLIP-DPR` folder and train models using inbatch negatives.
    - **Second step:** Then using `CLIP-DPR` to generate hard negatives fro training UniVL-DR.
    - **Final step:** Go to the `UniVL-DR` folder and train models using hard negatives.

3. Evaluation and results

Checkpoints

The checkpoint of UniVL-DR) can be found in [checkpoint_multi_hn (https://drive.google.com/drive/folders/1P8sY_fudNY_rTDRCYLR6ECTLYek-jqQv?usp=drive_link)]

The checkpoint of CLIP-DPR can be found in [checkpoint_multi_inb (https://drive.google.com/drive/folders/1Zzi6c2VX7xzlfWQodUffhydOPjcYABI-?usp=drive_link)]

Multi-Modal Retrieval Performance

The UniVL-DR outperforms all baselines with more than `7%` improvement on ranking evaluation, recalls more than `6%` relevant multi-modality documents, and even outperforms the divide-and-conquer model guided by

Setting	Model	MRR @10	NDCG @10	MRR @20	NDCG @20	Rec @20	Rec@ 100
Single Modality (Text Only)	BM25	53.75	49.60	54.10	51.72	68.16	80.69
	DPR (Zero-Shot)	22.72	20.06	23.14	21.79	32.78	45.43
	CLIP (Zero-Shot)	18.16	16.76	18.60	18.27	27.97	39.83
	BERT-DPR	42.16	39.57	42.76	42.26	60.85	77.10
	NQ-DPR	41.88	39.65	42.44	42.35	61.71	78.57
	NQ-ANCE	45.54	42.05	45.93	43.83	58.42	69.31
Divide-Conquer	VinVL-DPR	22.11	22.92	22.80	25.41	46.27	62.82
	CLIP-DPR	37.35	37.56	37.93	40.77	69.38	85.53
	BM25 & CLIP-DPR	42.27	41.58	42.79	44.69	73.34	87.50
	BM25 & CLIP-DPR (Oracle Modality)	61.05	58.18	61.37	60.45	80.82	90.83
UnivSearch	CLIP (Zero-Shot)	10.59	8.69	10.80	9.52	14.32	20.21
	VinVL-DPR	38.14	35.43	38.74	37.79	53.89	69.42
	CLIP-DPR	48.83	46.32	49.34	49.11	69.84	86.43
	UniVL-DR	62.40	59.32	62.69	61.22	80.37	89.42

Table 3 Multi-Modal Retrieval Performance. VinVL-DPR, CLIP-DPR, NQ-DPR and BERT-DPR are trained with in-batch negatives, while NQ-ANCE is trained with hard negatives.

Retrieval Performance of Different Ablation Models

UniVL-DR also shows its advantages by outperforming all baseline models on both text-text and text-image retrieval tasks, demonstrating that multi-modality modeling indeed benefits single/cross modality retrieval.

Model	**Retrieval Performance**		
	Text	Image	Multi
Single/Cross Modality Retrievers			
BERT-DPR	37.09	52.34	-
VinVL-DPR w/o caption	-	3.67	-
VinVL-DPR w/o fig feature	-	51.56	-
VinVL-DPR	25.00	48.68	-
CLIP-DPR w/o caption	-	17.74	-
CLIP-DPR w/o fig feature	-	58.17	-
CLIP-DPR	52.57	59.95	-
Universal Multi-Modal Retrievers			
VinVL-DPR w/o fig feature	29.01	46.55	36.13
VinVL-DPR	29.95	49.65	38.14
CLIP-DPR w/o fig feature	51.47	57.36	50.33
CLIP-DPR	51.75	60.61	48.83
UniVL-DR	**60.72**	**65.57**	**62.40**

Table 4 Retrieval Performance of Different Ablation Models. MRR@10 is used as the evaluation metric…

Effectiveness of Different Hard Negative Training Strategies

As we see, The advantage of UniVL-DR is to build a more uniform and effective multi-modal search model through modal balanced hard negative sampling, but the disadvantage is that it may require more computing resources and training time.

Sampling	**Retrieval Performance**			**Retrieved Image (%)**
	Text	Image	Multi	
In-batch Training				
CLIP-DPR (Random)	51.75	60.61	48.83	26.82
Balanced In-batch	52.24	59.99	49.88	30.35
Hard Negative Training				
Only Texts	54.92	52.88	36.26	91.74
Only Images	55.85	**66.51**	33.49	1.97
2 Texts & 1 Image	59.18	65.15	61.64	49.53
1 Text & 2 Images	57.86	66.23	61.20	47.88
ANCE (Random)	59.85	64.80	61.72	50.01
Balanced In-batch	**60.58**	65.21	**62.29**	**49.11**

Table 5 Effectiveness of Different Hard Negative Training Strategies.

Bridging cross-modality matching with image verbalization

UniVL-DR successfully bridges the modal gap between text and images through image literalization methods and achieves better results in text-image and multi-modal retrieval tasks.

Model	In-batch Training			Hard Neg Training		
	Text	Image	Muti	Text	Image	Muti
UniVL-DR	51.75	60.61	48.83	60.58	65.21	62.29
w. Verbalized Caption	51.51	60.57	49.49	59.86	65.87	62.24
w. Verbalized Query	52.14	59.72	50.21	60.72	65.57	62.40

Table 6 : Performance of Multi-Modality Retrieval Models with Different Image Verbalization Methods. All models are evaluated with MRR@10.

Therefore, UniVL-DR demonstrates significant advantages in various aspects, highlighting its critical importance in addressing multi-modal retrieval tasks.

Discussion

The discussion section delves into the nuanced aspects of the project, examining the implications of our findings, addressing challenges encountered, and offering insights for future research.

Model Performance and Challenges:

UniVL-DR demonstrated superior performance in multi-modal retrieval, showcasing the effectiveness of learning a unified representation space. The model excelled in balancing modalities during training, mitigating modality preferences, and achieving impressive gains in both single and cross-modality tasks. However, challenges were noted in handling specific figure semantics in images, indicating potential areas for further improvement in image understanding.

Impact of Modality-Balanced Training:

The introduction of a modality-balanced hard negative training strategy played a pivotal role in enhancing UniVL-DR’s capabilities. The model’s ability to effectively fuse retrieval results from diverse modalities highlights the significance of addressing modality gaps in training for comprehensive multi-modal understanding.

Image Verbalization Method:

The image verbalization method, an innovative addition to UniVL-DR, successfully bridged the gap between language and vision. By paraphrasing image pixel facts into natural language, the model demonstrated improved textual representations for images. Further refinement of this method could offer deeper insights into enhancing the overall multi-modal retrieval experience.

Generalization and Future Directions:

UniVL-DR showcased promising generalization across various tasks, emphasizing the versatility of learned universal representations. Future research directions could explore the model's adaptability to specific domains and investigate its potential in handling additional modalities, such as audio and video.

Interactive Retrieval Interfaces:

The project opens doors to the development of interactive retrieval interfaces that leverage UniVL-DR's universal representations. Such interfaces could provide users with more intuitive and efficient means of retrieving information across diverse media types, contributing to enhanced user experiences in information retrieval.

Limitations and Considerations:

Despite the successes, it's essential to acknowledge limitations. Fine-tuning UniVL-DR on domain-specific datasets could provide deeper insights into its performance in specialized applications. Additionally, ongoing advancements in pre-training techniques and larger-scale datasets may influence the model's efficacy and generalization capabilities.

Conclusions

In conclusion, our project, focused on Universal Vision-Language Dense Retrieval (UniVL-DR) for multi-modal retrieval, has made significant strides in bridging the gap between diverse media types. The key contributions and findings include:

- **Unified Representation Space**: UniVL-DR successfully learned a unified representation space for multi-modal documents, breaking the modality boundary and achieving state-of-the-art performance in retrieval tasks.
- **Modality-Balanced Training**: The implementation of a modality-balanced hard negative training strategy proved effective in addressing the challenge of fusing retrieval results from different modalities, leading to enhanced performance.
- **Image Verbalization**: The innovative image verbalization method added a new dimension to multi-modal understanding by converting image features into natural language, thereby improving the textual representations of images.

Lessons Learned

Through the course of this project, several valuable lessons were learned:

- **Importance of Modality Balance**: Achieving a balance between different modalities during training is crucial for the success of multi-modal retrieval models. Our modality-balanced training approach significantly contributed to the model's effectiveness.
- **Versatility of Universal Representations**: UniVL-DR demonstrated that learning one universal representation space can extend benefits beyond multi-modal tasks, providing gains in single-modality tasks as well.

Future Work

While our project has made notable progress, there are several avenues for future exploration:

- **Exploration of Additional Modalities**: Extend the model to handle a broader range of modalities, such as audio and video, to create an even more versatile multi-modal retrieval system.
- **Fine-Tuning for Specific Domains**: Investigate the adaptability of UniVL-DR in specific domains by fine-tuning the model on domain-specific datasets, exploring its potential in specialized applications.
- **Enhanced Image Verbalization**: Further research can be conducted to refine and improve the image verbalization method, exploring novel techniques for better aligning semantics in image captions and pixel features.
- **Interactive Retrieval Interfaces**: Develop interactive interfaces that leverage the universal representations learned by UniVL-DR, providing users with more intuitive and efficient ways to retrieve information across diverse media types.

In summary, our project not only advances the field of multi-modal retrieval but also opens up promising directions for future research and application development. The lessons learned and insights gained lay a foundation for continued exploration in this dynamic and evolving domain.