A Comparison of Late-fusion Training Strategies for Quad-modal Joint Embeddings

Domenic Luca Fürer
Department of Informatics
University of Zurich
Zurich, Switzerland
0009-0006-8784-2519

Abraham Bernstein
Department of Informatics
University of Zurich
Zurich, Switzerland
0000-0002-0128-4602

Luca Rossetto
Department of Informatics
University of Zurich
Zurich, Switzerland
0000-0002-5389-9465

***Abstract*—Multimodal methods that align the semantics of two data modalities—usually text and image—have become increasingly widespread in recent years. While such bi-modal alignments have received much attention, less work has been done toward aligning more than two modalities. This paper explores several strategies for training late-fusion joint-embedding models with up to four simultaneous modalities. We show that, even on a comparatively small dataset, there are clear and disproportional advantages when considering all modalities concurrently rather than consecutively. The results indicate that, when aligning multiple modalities, a training set with even weak alignment between all modalities substantially outperforms the use of several pair-wise alignments and that the addition of a modality can disproportionally increase alignment quality across all modalities.**

***Index Terms*—Multi-modal Embedding, Late-Fusion Joint-Embedding, Quad-modal Embedding**

I. Introduction

In recent years, machine learning has seen remarkable advancements in terms of capabilities, public interest, and accuracy. Popular models such as Dall-E [1] and Stable Diffusion [2] have gained widespread recognition and are utilized extensively across various domains. These models demonstrate the ability to generate images based on textual prompts. Related approaches exist that are not limited to the generation of images, but systems that generate audio or video based on text inputs [3], [4]. The key characteristic these models share is their incorporation of two different modalities. Beyond their generation capabilities, these multimodal models enable diverse real-life tasks, including multimodal retrieval and autonomous systems [5].

By embracing multiple modalities, these systems demonstrate improved resilience to noise [6] and achieve enhanced overall performance [5]. Consequently, they hold the potential to impact various domains by leveraging the richness of information present in diverse modalities [7]. To accomplish this objective, the semantic similarity between the different data types needs to be calculated. Typically, this is achieved through learning projections for individual modalities into a shared vector space. By mapping different modalities into this common space, direct comparison becomes possible, enabling the calculations of similarities [8].

So far, most such approaches focus on two media types, making them bi-modal. However, the rapid growth of multi-modal data, encompassing diverse formats such as text, image, audio, and video, has created the possibility for embedding more than "just" two modalities [9]. Much of the literature, however, primarily focuses on bi-modal embeddings, typically combining text and image. As a result, there exists a gap in the exploration of multimodal feature spaces that can accommodate more than just image and text modalities. Although there are a few instances of attempts to incorporate four modalities [10], [9], these approaches often utilize highly customized or novel objective functions with numerous additional terms intended to aid model convergence. Despite advancements in two modal architectures, for example, the two-tower CLIP model [11], the extension of these architectures to incorporate additional modalities has not been widely explored in the literature, which presents opportunities for investigation in multimodal embeddings.

This paper explores methods to extend the commonly used bi-modal late fusion approaches to additional modalities. Using different combinations of modalities and different training strategies, we show that the use of additional modalities can boost the alignment performance such that the whole is more than the sum of its parts. These effects can already be observed using a comparatively small dataset and few trainable parameters.

The remainder of this paper is structured as follows: After a brief overview of related work in Section II, we detail our experimental approach and the different tested conditions in Section III. The results of the conducted experiments are presented in Section IV and discussed in Section V. Finally, Section VI summarizes the findings and offers some concluding remarks.

II. Related Work

Several ways to fuse information from different modalities can be found in the literature. Fusion can be done at the input level, which is known as early-fusion, at the decision level, known as late-fusion, or intermediately, also known as hybrid-fusion [7], [12]. Relevant for this discussion is the mechanism of late fusion, which [13] defines as *"Fusion scheme that first reduces unimodal features to separately learned concept scores, then these scores are integrated to learn concepts."*

Late-fusion approaches can be found for various combinations of modalities. A popular combination is image and text, for which many different variations have been explored in recent years [14], [15], [16], [17], [11], [8]. A closely related combination is the one of video and text, where various approaches share many ideas with the image-text alignment methods [18], [19].

Similar alignment strategies can also be found in other domains. In the domain of Knowledge Graphs, for example, [20] present a method to simultaneously align several graph embeddings from different source graphs, even if they were generated using different embedding methods. The authors find that aligning multiple embeddings concurrently improves performance on downstream tasks, such as link prediction, compared to pairwise aligned embeddings.

Other approaches simultaneously align three different modalities, such as [21] using text, image, and video, or [22] with text, image, and audio. Joint alignments of more modalities are comparatively rare. Examples include [10] combining text, image, audio, and video in an early-fusion scheme or [6] which jointly considers text, image, audio, video, and 3D.

A related area that has recently received increased attention is the one of multi-modal language models. Methods such as OneLLM [23], UnIVAL [24], or AnyGPT [25] use modality-specific tokenization mechanisms to use different modalities in a transformer-based [26] language model. Such models can then be trained to perform various tasks on multi-modal input, commonly composed of a textual prompt and a non-textual content element such as an image. While such approaches can use input from multiple modalities, their internal representations can not easily be used for the alignment of multiple modalities outside of the model in question, and any downstream tasks need to be performed by the models themselves.

III. Method

In this section, we describe our comparative methodology, provide an overview of the used dataset and model architecture, and discuss the different training strategies.

A. Dataset

While an increasing number of bi-modal datasets exist, only a few align more than two modalities. Among the largest and thematically most diverse is the PKU XMediaNet Dataset [9], [27], containing samples from 200 different WordNet [28] classes across five different types of media: text, image, audio, video, and 3D. For our purposes, we omit the 3D data, as it only consists of much fewer samples than the other modalities. Across the four relevant modalities, the dataset consists of 40 000 text-, 40 000 image-, 10 000 audio-, and 10 000 video samples.

Since the dataset authors distribute everything but the text in the form of URLs pointing to the original content sources, not all data was available anymore at the time of our experiments. By querying the provided source URLs, we were able to obtain 35 268 images (88.2%), 7 876 videos (78.8%), and 8 280 audio files (82.8%). Such link-rot-induced dataset decay is a commonly observed phenomenon in datasets distributed as URLs [29]. Since the elements in the dataset are aligned by class label and not in a one-to-one fashion and are inherently unequally distributed across the different modalities, this decay does not pose a substantial restriction in our case since the underrepresented modalities per class need to be over-sampled in any case.

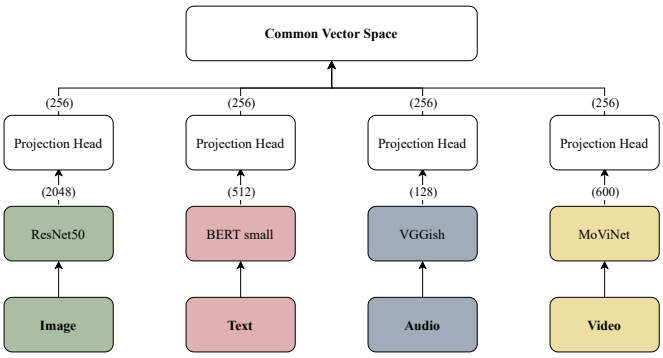


Fig. 1. Architecture of the quad-modal late-fusion model.

B. Architecture

Our joint-embedding model consists of 4 modality-specific backbones, followed by a projection head to align the modalities. We specifically chose smaller models that have been trained on uni-modal data on a classification task. While these models are no longer state-of-the-art in their respective domains, they are all well-established in literature and have been shown to be applicable to various applications. Since we are in this study not interested in the absolute performance of the resulting model but rather in the relative performance of different alignment strategies, such smaller models with comparatively small parameter counts that have been pre-trained on datasets with well-known properties are well-suited as modality-specific backbones for this task.

For the image modality, we use a ResNet50 [30] that has been pre-trained on the ImageNet [31] 1000 corpus. For videos, a MoViNet [32] is used that has been pre-trained on the Kinetics 600 [33] dataset. Audio encoding is done using a VGGish [34] model, pre-trained on the YouTube-8M [35] dataset. For text, a pre-trained BERT small [36] model is used. All these models are used as embedding backbones without their final classification layer, and all models remain frozen during the training process. The complete architecture is illustrated in Figure 1.

The projection head consists of two dense layers with a drop-out and a skip-connection. Between the two dense layers, a GELU [37] is used as a non-linearity. We set the dimensionality of the output to 256, similarly to [38], [39]. Its architecture is illustrated in Figure 2.

This simple architecture is chosen to keep the parameter count low, given the comparatively small training set that is available for our experiments. Since not all backbone models

have the same output dimensionality, the parameter count in the first dense layer inevitably varies across modalities. The second dense layer has a consistent number of parameters across all four modalities.

C. Loss Function

A popular choice in loss function for the bi-modal alignment of embeddings is the InfoNCE loss, first introduced by [40] and popularized by CLIP [11]. The InfoNCE loss, as shown in Equation 1, takes two sequences of pairwise-aligned inputs $\mathbf{x}$ and $\mathbf{y}$, as well as a temperature parameter τ. N describes the sequence-length of $\mathbf{x}$ and $\mathbf{y}$. In accordance with [38], we set $\tau = 0.05$ for all our experiments.

$$InfoNCE(\mathbf{x}, \mathbf{y}) = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(\mathbf{x}_i \cdot \mathbf{y}_i/\tau)}{\sum_{j=1}^{N} exp(\mathbf{x}_i \cdot \mathbf{y}_j/\tau)} \quad (1)$$

In order to be able to use more than two modalities, we extend the loss function as shown in Equation 2. M represents the set of all considered modalities, and $\mathcal{P}(\cdot)$ is the permutation function, generating all pairwise permutations of its input.

$$\mathcal{L} = \frac{1}{|\mathcal{P}(M)|}\sum_{(m_1,m_2)\in\mathcal{P}(M)} InfoNCE(m_1, m_2) \quad (2)$$

D. Scenarios

We test and compare three different scenarios: First, we train an instance of the quad-modal model described above in a bi-modal round-robin fashion, only using two modalities per batch. This is supposed to simulate a scenario where only pairwise aligned datasets are available. All modalities are used in each epoch, but only two are used simultaneously.

Next, we omit one modality from the above model and train it in a tri-modal configuration with jointly-aligned text, image, and audio samples.

Finally, we train the model in a quad-modal configuration, using text, image, audio, and video samples jointly.

The backbones are frozen in all scenarios, and only the projection heads are trained.

Training is stopped once the loss on the test set stops decreasing.

As a performance metric, we measure recall at 1, 5, and 10 on both the training and the test set for each scenario.

IV. Results

This section presents the results for the different tested scenarios. The reported measure is Recall at 1, 5, and 10. We report the recall values on both the training and the test set to show any possible over-fitting or memorization effects. Recall treats the class labels as flat and does not consider any similarity between classes.

TABLE I
RECALL AT 1, 5, AND 10 FOR MODALITY PAIRS IN PAIR-WISE BI-MODAL EMBEDDING, TRAINED IN A ROUND-ROBIN FASHION

			Test Split			Train Split		
Recall at			1	5	10	1	5	10
Text	→	Image	0.02	0.07	0.12	0.02	0.08	0.13
Text	→	Audio	0.02	0.08	0.14	0.02	0.07	0.13
Text	→	Video	0.02	0.09	0.15	0.02	0.09	0.14
Image	→	Text	0.02	0.09	0.17	0.03	0.10	0.17
Image	→	Audio	0.02	0.10	0.18	0.03	0.10	0.17
Image	→	Video	0.03	0.11	0.19	0.04	0.13	0.21
Audio	→	Text	0.02	0.08	0.13	0.02	0.07	0.12
Audio	→	Image	0.02	0.08	0.12	0.02	0.06	0.11
Audio	→	Video	0.02	0.08	0.14	0.02	0.09	0.15
Video	→	Text	0.03	0.09	0.15	0.02	0.08	0.14
Video	→	Image	0.03	0.09	0.15	0.04	0.12	0.18
Video	→	Audio	0.02	0.09	0.15	0.03	0.11	0.17
Mean			0.02	0.09	0.15	0.03	0.09	0.15

A. Quad-modal Round-Robin

To establish a baseline for the following experiments, we train the model described in Section III-B in a pair-wise bi-modal round-robin fashion. This simulates a scenario where no alignment between more than two modalities is available, which is commonly the case in contemporary multi-modal datasets. We report the recall at 1, 5, and 10 for all pairwise modality combinations in Table I, where each line shows the results when using a query element from one modality to retrieve relevant results from one specific other modality.

In comparison, Table II shows the results when the target modality is not specified, and all elements (excluding the query element) are considered for retrieval. Since the target modality is not limited, it is labeled as *Any* in the table.

As can be seen from the presented tables, retrieval performance is generally poor for this approach. The highest recall@10 was achieved when using an image to query for video, with a value of 0.19 on the test set and 0.21 on the training set. There is, however, no substantial difference between the recall values for the different modality pairs, nor is there a substantial difference between the results for the training and the test set.

TABLE II
RECALL AT 1, 5, AND 10 FOR MODALITY PAIRS IN PAIR-WISE BI-MODAL EMBEDDING, TRAINED IN A ROUND-ROBIN FASHION

			Test Split			Train Split		
Recall at			1	5	10	1	5	10
Text	→	Any	0.00	0.00	0.01	0.00	0.00	0.01
Image	→	Any	0.00	0.00	0.01	0.00	0.00	0.01
Audio	→	Any	0.00	0.00	0.01	0.00	0.00	0.01
Video	→	Any	0.00	0.00	0.01	0.00	0.00	0.01
Mean			0.00	0.00	0.01	0.00	0.00	0.01

When not constraining the target modality, the recall at 1 and 5 drop to 0, independent of the modality of the query, for both the training and the test set.

B. Tri-modal combination

For the second experiment, we omit the video modality from the model described in Section III-B and train it in a tri-modal

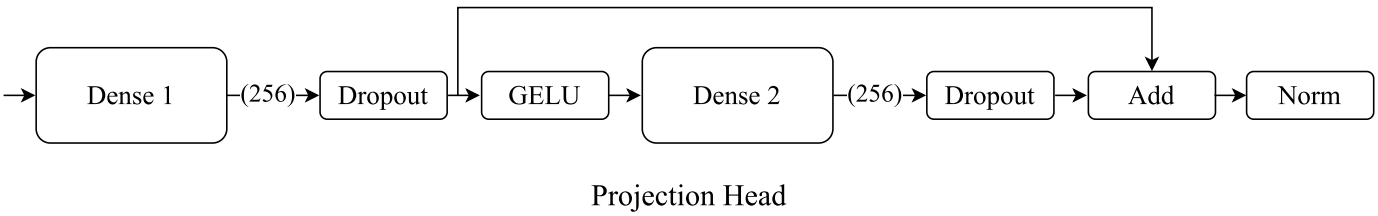


Fig. 2. Projection head used to align descriptors generated by pre-trained backbone models.

TABLE III
RECALL AT 1, 5, AND 10 FOR MODALITY PAIRS IN TRI-MODAL EMBEDDING

	Test Split			Train Split		
Recall at	1	5	10	1	5	10
Text → Image	0.08	0.20	0.28	0.09	0.21	0.27
Text → Audio	0.03	0.11	0.19	0.04	0.12	0.20
Image → Text	0.07	0.25	0.36	0.10	0.25	0.36
Image → Audio	0.05	0.16	0.25	0.07	0.18	0.28
Audio → Text	0.04	0.13	0.21	0.07	0.17	0.22
Audio → Image	0.05	0.13	0.20	0.05	0.14	0.22
Mean	0.05	0.16	0.25	0.07	0.18	0.26

fashion using only the samples of the text, image, and audio modalities using the loss function described in Section III-C. The results are shown in Tables III and IV. While some differentiation can be seen between modality pairs, with the Image-Text pair slightly outperforming the other two, the results are somewhat close across all combinations. No substantial difference between training and test set results can be observed either.

In the case of unconstrained target modalities, shown in Table IV, results are still poor. Only in one case—using image-queries on the training set—did recall@1 not remain at zero. While the overall magnitude of the measured values is still low, they at least allow for a qualitative comparison between the results obtained on the training and test set, where, similarly to Table III no substantial differences can be observed.

TABLE IV
RECALL AT 1, 5, AND 10 FOR MODALITY PAIRS IN TRI-MODAL EMBEDDING

	Test Split			Train Split		
Recall at	1	5	10	1	5	10
Text → Any	0.00	0.03	0.06	0.00	0.03	0.07
Image → Any	0.00	0.05	0.10	0.01	0.06	0.13
Audio → Any	0.00	0.02	0.05	0.00	0.03	0.06
Mean	0.00	0.03	0.07	0.00	0.04	0.09

C. Quad-modal Combination

For the third and final experiment, we train the model with all four modalities simultaneously, using the loss function from Section III-C. Tables V and VI again show the modality-constrained and unconstrained recall values, respectively.

For the pair-wise modality-constrained results shown in Table V, we can see some differentiation between the different modalities. The best results are achieved when using an image to query for videos. This is despite the fact that the image and video backbones have been pre-trained on different datasets, and the training set does not have any content duplication between the image and video samples. Nevertheless, both modalities being visual appears to help with alignment. The modality pairs containing the audio modality consistently underperform the other pairings, indicating a systematic limitation in either the used backbone network or the training set. There is, again, little difference between the result obtained from the training and the test set, validating the employed training mechanism.

TABLE V
RECALL AT 1, 5, AND 10 FOR MODALITY PAIRS IN QUAD-MODAL EMBEDDING

	Test Split			Train Split		
Recall at	1	5	10	1	5	10
Text → Image	0.11	0.27	0.37	0.12	0.28	0.37
Text → Audio	0.04	0.14	0.23	0.04	0.15	0.25
Text → Video	0.07	0.21	0.33	0.09	0.26	0.37
Image → Text	0.17	0.43	0.56	0.20	0.45	0.57
Image → Audio	0.07	0.26	0.38	0.10	0.30	0.43
Image → Video	0.20	0.49	0.64	0.26	0.55	0.68
Audio → Text	0.04	0.14	0.23	0.05	0.17	0.25
Audio → Image	0.06	0.18	0.25	0.08	0.20	0.27
Audio → Video	0.05	0.16	0.25	0.07	0.19	0.28
Video → Text	0.09	0.25	0.37	0.11	0.31	0.43
Video → Image	0.19	0.38	0.47	0.26	0.46	0.55
Video → Audio	0.05	0.17	0.26	0.06	0.21	0.33
Mean	0.10	0.26	0.36	0.12	0.29	0.40

When looking at the modality-unconstrained results of Table VI, again some differentiation between the used query modalities becomes apparent. Querying using images performs by far the best in this experiment, while using an audio sample as a query performs the worst.

V. DISCUSSION

In this section, we discuss the insights that can be gained from the presented results, as well as the limitations of the performed experiments.

TABLE VI
RECALL AT 1, 5, AND 10 FOR MODALITY PAIRS IN QUAD-MODAL EMBEDDING

	Test Split			Train Split		
Recall at	1	5	10	1	5	10
Text → Any	0.00	0.02	0.07	0.00	0.03	0.08
Image → Any	0.01	0.09	0.21	0.02	0.13	0.24
Audio → Any	0.00	0.01	0.05	0.00	0.03	0.06
Video → Any	0.00	0.04	0.09	0.01	0.07	0.14
Mean	0.00	0.04	0.11	0.01	0.06	0.13

A. Insights

What can be clearly observed from the results presented in Section IV is that none of the modality-constrained constellations performs as well as a state-of-the-art bi-modal co-embedding model. This is not surprising given the comparatively small training set with rather coarse, class-label-based alignment, and the frozen backbone networks that were pre-trained on different data. For this discussion, we are not interested in the absolute performance when compared to other bi-modal methods trained and evaluated on different datasets. Instead, we are interested in the relative performance between the three presented scenarios and the insights that can be gained when comparing them.

We first tested a bi-modal pair-wise round-robin approach as a baseline, since such an approach would be an intuitive and straightforward way to combine multiple bi-modal datasets that cannot be easily aligned directly. While the rather poor results obtained from this approach tell us little in isolation, they become much more insightful when we compare the matching rows in Tables I and III. While there was less data available in total for the tri-modal embedding when compared to the round-robin one, due to the omission of the video modality, the results in Table III still consistently and substantially outperform those shown in Table I. For both image-to-text and text-to-image retrieval, for example, the recall@10 values are more than twice for the tri-modal joint embedding compared to what was achieved in the round-robin approach. Therefore, we can conclude that introducing an additional jointly aligned modality is more effective than adding further not fully aligned data.

When further comparing the results of the joint tri-modal training with those from the joint quad-modal training from Table V, we can again see increased recall values for all pairings already present in Table III. This holds despite there having been no increase in training data for the text, image, or audio modalities. These results again suggest that introducing an additional aligned modality that can be considered jointly during training can inform the alignment between all modalities.

These results consistently suggest that in order to achieve a well-aligned embedding across multiple modalities, having a training set with aligned samples from all modalities disproportionately improves performance. The results do not preclude the possibility that comparable multimodal performance could not also be achieved by using a training set with several only pairwise-aligned training samples. They do, however, seem to indicate that in order to achieve it, comparatively more data, and hence more compute, will be required.

B. Limitations

While we consider the presented results to be insightful, there are some limitations related to them and the performed experiments that we want to acknowledge here. While the dataset we used for these experiments is the largest and semantically most diverse one with (more than) four modalities that can (to the best of our knowledge) be currently found in literature, it is rather small and offers only coarse-grained alignment between the samples when compared to state-of-the-art bi-modal datasets. It is, therefore, also unsurprising that the recall values we obtained here are lower than other bi-modal methods can achieve on larger bi-modal datasets. The limited availability of such multi-modally aligned data also means that our experiments had to be limited to one dataset with one specific set of properties. We are, therefore, unable to say to what degree our findings are influenced by these properties and how many of the insights would replicate on other data.

We show the results of experiments performed using a late-fusion approach with pre-trained and frozen backbone models. While we also performed similar experiments with early-fusion architectures that were trained end-to-end, model convergence was not reached on the available data and the obtained results were inconclusive. We, therefore, omit them here. While it can be considered plausible that the insights presented above are also valid for other fusion schemes, we could not empirically examine this.

VI. CONCLUSION

While the combination of two media modalities, primarily image and text, has received much attention in recent years, multimodal approaches with more than two modalities have so far received comparatively little attention. In this paper, we presented a comparison of training strategies for late-fusion joint-embeddings of up to four modalities: text, image, audio, and video. We were able to show that, when it comes to the joint alignment of multiple modalities, jointly aligned training samples are more important than more pair-wise aligned samples and that introducing an additional aligned modality can boost the retrieval performance across all modalities, as measured by recall.

In future work, it might be interesting to invest in creating larger and more finely aligned multimodal datasets and expand these experiments to different model architectures and fusion schemes. Using a larger dataset, the insights gained here could lead the way toward new foundation models that go beyond bi-modality.

REFERENCES

[1] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine

Learning Research, vol. 139. PMLR, 2021, pp. 8821–8831. [Online]. Available: http://proceedings.mlr.press/v139/ramesh21a.html

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 10 674–10 685. [Online]. Available: https://doi.org/CVPR52688.2022.01042

[3] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. P. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 21 450–21 474. [Online]. Available: https://proceedings.mlr.press/v202/liu23f.html

[4] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-a-video: Text-to-video generation without text-video data," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/pdf?id=nJfylDvgzlq

[5] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, 2017. [Online]. Available: https://doi.org/MSP.2017.2738401

[6] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, "Semi-supervised cross-media feature learning with unified patch graph regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 583–596, 2016. [Online]. Available: https://doi.org/TCSVT.2015.2400779

[7] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *Vis. Comput.*, vol. 38, no. 8, pp. 2939–2970, 2022. [Online]. Available: https://doi.org/10.1007/s00371-021-02166-7

[8] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 5005–5013. [Online]. Available: https://doi.org/10.1109/CVPR.2016.541

[9] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, 2018. [Online]. Available: https://doi.org/10.1109/TCSVT.2017.2705068

[10] X. He, Y. Peng, and L. Xie, "A new benchmark and approach for fine-grained cross-media retrieval," in *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*. ACM, 2019, pp. 1740–1748. [Online]. Available: https://doi.org/10.1145/3343031.3350974

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html

[12] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: multimodal transfer module for CNN fusion," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 13 286–13 296. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Joze_MMTM_Multimodal_Transfer_Module_for_CNN_Fusion_CVPR_2020_paper.html

[13] C. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November 6-11, 2005*. ACM, 2005, pp. 399–402. [Online]. Available: https://doi.org/10.1145/1101149.1101236

[14] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 7181–7189. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Gu_Look_Imagine_and_CVPR_2018_paper.html

[15] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2598339

[16] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 12 888–12 900. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html

[17] D. Peri, S. Sah, and R. W. Ptucha, "Show, translate and tell," in *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*. IEEE, 2019, pp. 295–299. [Online]. Available: https://doi.org/10.1109/ICIP.2019.8802922

[18] J. Dong, X. Li, and C. G. M. Snoek, "Predicting visual features from text for image and video caption retrieval," *IEEE Trans. Multim.*, vol. 20, no. 12, pp. 3377–3388, 2018. [Online]. Available: https://doi.org/10.1109/TMM.2018.2832602

[19] H. Xu, G. Ghosh, P. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 2021, pp. 6787–6800. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.544

[20] M. Baumgartner, D. Dell'Aglio, H. Paulheim, and A. Bernstein, "Towards the web of embeddings: Integrating multiple knowledge graph embedding spaces with fedcoder," *J. Web Semant.*, vol. 75, p. 100741, 2023. [Online]. Available: https://doi.org/10.1016/j.websem.2022.100741

[21] X. Li, C. Xu, G. Yang, Z. Chen, and J. Dong, "W2VV++: fully deep learning for ad-hoc video search," in *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*. ACM, 2019, pp. 1786–1794. [Online]. Available: https://doi.org/10.1145/3343031.3350906

[22] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *CoRR*, vol. abs/2109.01797, 2021. [Online]. Available: https://arxiv.org/abs/2109.01797

[23] J. Han, K. Gong, Y. Zhang, J. Wang, K. Zhang, D. Lin, Y. Qiao, P. Gao, and X. Yue, "Onellm: One framework to align all modalities with language," *CoRR*, vol. abs/2312.03700, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.03700

[24] M. Shukor, C. Dancette, A. Rame, and M. Cord, "UnIVAL: Unified model for image, video, audio and language tasks," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://openreview.net/forum?id=4uflhObpcp

[25] J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li, H. Yan, J. Fu, T. Gui, T. Sun, Y. Jiang, and X. Qiu, "Anygpt: Unified multimodal llm with discrete sequence modeling," *CoRR*, 2024. [Online]. Available: https://arxiv.org/abs/2402.12226

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[27] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5585–5599, 2018. [Online]. Available: https://doi.org/10.1109/TIP.2018.2852503

[28] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995. [Online]. Available: https://doi.org/10.1145/219717.219748

[29] V. Lakic, L. Rossetto, and A. Bernstein, "Link-rot in web-sourced multimedia datasets," in *MultiMedia Modeling - 29th International Conference, MMM 2023, Bergen, Norway, January 9-12, 2023, Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 13833. Springer, 2023, pp. 476–488. [Online]. Available: https://doi.org/10.1007/978-3-031-27077-2_37

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[31] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 2009, pp. 248–255. [Online]. Available: https://doi.org/10.1109/CVPR.2009.5206848

[32] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "Movinets: Mobile video networks for efficient video recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 16 020–16 030. [Online]. Available: https://doi.org/CVPR46437.2021.01576

[33] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *CoRR*, vol. abs/1907.06987, 2019. [Online]. Available: http://arxiv.org/abs/1907.06987

[34] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. IEEE, 2017, pp. 131–135. [Online]. Available: https://doi.org/10.1109/ICASSP.2017.7952132

[35] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *CoRR*, vol. abs/1609.08675, 2016. [Online]. Available: http://arxiv.org/abs/1609.08675

[36] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[37] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," *CoRR*, vol. abs/1606.08415, 2016. [Online]. Available: http://arxiv.org/abs/1606.08415

[38] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 1708–1718. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.00175

[39] C. Hori, T. Hori, T. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 4203–4212. [Online]. Available: https://doi.org/10.1109/ICCV.2017.450

[40] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: http://arxiv.org/abs/1807.03748