ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Lightweight Cross-Modal Representation Learning

Bilal FAYE[1], Hanane AZZAG[1], Mustapha LEBBAH[2], Djamel BOUCHAFFRA[3]

1- LIPN, UMR CNRS 7030 Sorbonne Paris Nord University, Villetaneuse, France
2- DAVID Lab, University of Versailles, University Paris-Saclay, Versailles, France
3- Center for Development of Advanced Technologies, Algiers, Algeria

Abstract. Low-cost cross-modal representation learning is crucial for deriving semantic representations across diverse modalities such as text, audio, images, and video. Traditional approaches typically depend on large specialized models trained from scratch, requiring extensive datasets and resulting in high resource and time costs. To overcome these challenges, we introduce a novel approach named Lightweight Cross-Modal Representation Learning (LightCRL). This method uses a single neural network titled Deep Fusion Encoder (DFE), which projects data from multiple modalities into a shared latent representation space. This reduces the overall parameter count while still delivering robust performance comparable to more complex systems. The code is available via https://github.com/b-faye/lightweightCRL/

1 Introduction

Modalities are channels for exchanging information between humans and the environment, including text, audio, images, and video. Cross-modal learning bridges these channels using alignment and fusion strategies. According to [8], cross-modal fusion combines data into a unified representation—early, late, intermediate, or hybrid fusion. Early fusion integrates raw features at the input level, late fusion at the output level, and intermediate fusion at multiple stages. Hybrid fusion employs a mix of these methods. Cross-modal alignment aligns modalities within a shared semantic space using techniques like contrastive learning and encoder-decoder architectures [1]. Contrastive learning enhances alignment between similar samples and reduces it for dissimilar ones, as used in ConVIRT and CLIP [9, 12]. Masked modeling, like in VisualBERT, learns by predicting unseen parts of data [3]. Encoder-decoder structures, seen in visualGPT, facilitate modality interactions [6]. These methods, however, require large models for each modality and extensive aligned datasets, posing high computational costs and data availability challenges.

To address the computational and data challenges in cross-modal learning, we introduce Lightweight Cross-Modal Representation Learning (LightCRL), a method designed to acquire high-level semantic representations while minimizing computational resources efficiently. LightCRL significantly reduces the number of parameters required and operates independently of large aligned datasets by leveraging large pre-trained models that remain frozen during the process. At the core of LightCRL is the Deep Fusion Encoder (DFE), which stands as the

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

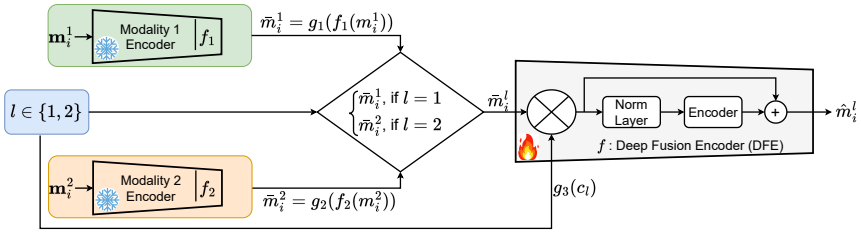

Fig. 1: LightCRL framework: Only DFE denoted f is trained for cost-effectiveness, while f_1 and f_2 remain static. $\mathbf{m}_i^1$, $\mathbf{m}_i^2$ represent modalities 1 and 2 with their respective embedding $\bar{\mathbf{m}}_i^1$ and $\bar{\mathbf{m}}_i^2$, using respective frozen encoders. $\hat{\mathbf{m}}_i^1$ and $\hat{\mathbf{m}}_i^2$ are embeddings on the common latent space. c_l is the context identifier for modality l, and g_1, g_2, and g_3 ensure uniform dimensions.

unique trainable element shared across different modalities. The DFE integrates a learnable context vector specific to each modality, combining these vectors with the embeddings of the modalities to concentrate training efforts solely on the parameters of the DFE. Our contributions are threefold: (i) we maximize efficiency by using frozen pre-trained models to curtail extensive training and conserve resources; (ii) we streamline the learning process and enhance information integration across modalities with the unified DFE, even in scenarios with limited datasets; (iii) and we enable effective differentiation and representation of diverse data through context-aware fusion within the DFE, using consistent parameters across different modalities.

2 Proposed Method

Lightweight Cross-Modal Representation Learning (LightCRL) offers a novel approach to efficient representation learning across modalities. It uses large pre-trained encoders for each modality, kept fixed during training. LightCRL introduces a unique neural network, "Deep Fusion Encoder", to handle cross-modal representations. The DFE (f) in Figure 1 plays a pivotal role in our framework. Nonlinear projection functions g_1, g_2, and g_3 ensure input dimension consistency for f [12]. f combines trainable context vector c with frozen pre-trained encoders f_1 and f_2, tailored to each modality for coherent fusion. This approach allows DFE to represent diverse modalities effectively with shared parameters. We propose a shallow neural network for DFE, maintaining semantic representation efficiently and cost-effectively. Consider datasets $\mathbf{M}_1$ and $\mathbf{M}_2$ for modalities 1 and 2 respectively, with data pairs $(\mathbf{m}_i^1, \mathbf{m}_i^2)$ where $\mathbf{m}_i^1$ is from modality 1 and $\mathbf{m}_i^2$ from modality 2. The paired dataset is $\{(\mathbf{m}_1^1, \mathbf{m}_1^2), (\mathbf{m}_2^1, \mathbf{m}_2^2), ..., (\mathbf{m}_N^1, \mathbf{m}_N^2)\}$ with N samples. LightCRL, as depicted in Figure 1, trains f for cross-modal representation via alignment in a shared latent space. The context modality identifier c_l with $l \in \{1, 2\}$ enhances the Deep Fusion Encoder (DFE) representation. This approach enables uniform parameter utilization across different modalities,

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

thereby eliminating the necessity to learn individual encoder weights for each modality. This method is called "Lightweight" due to its reduced computational overhead and increased efficiency in managing multiple modalities within a single model framework. For each input pair $(\mathbf{m}_i^1, \mathbf{m}_j^2)$, LightCRL's transformation is defined as $\hat{\mathbf{m}}_i^1 = f(g_1(f_1(\mathbf{m}_i^1)) \otimes g_3(c_1))$, and $\hat{\mathbf{m}}_i^2 = f(g_2(f_2(\mathbf{m}_i^2)) \otimes g_3(c_2))$. The symbol $\otimes$ represents the fusion operation between the context modality vector and embeddings from frozen encoders. Fusion strategies include element-wise methods (addition, multiplication, concatenation) outlined in [10], or cross-attention fusion [5]. During training, we sample a minibatch of K input pairs $(\mathbf{m}_i^1, \mathbf{m}_i^2)$ from the training dataset of N pairs. LightCRL's training objective function includes a contrastive loss between modality 1 and 2 for pairs $(\hat{\mathbf{m}}_i^1, \hat{\mathbf{m}}_j^2)$:

$$\ell_{ij} = -\log\left(\frac{\exp(\langle \hat{\mathbf{m}}_i^1, \hat{\mathbf{m}}_j^2 \rangle / \tau)}{\sum_{k=1}^{K} \exp(\langle \hat{\mathbf{m}}_i^1, \hat{\mathbf{m}}_k^2 \rangle / \tau)}\right) \tag{1}$$

The term $\langle \hat{\mathbf{m}}_i^1, \hat{\mathbf{m}}_i^2 \rangle$ represents cosine similarity, with $\tau \in \mathbb{R}^+$ as a learnable temperature parameter. This loss function preserves mutual information between true pairs through representation functions. To ensure symmetry, we introduce a similar contrastive loss from modality 2 to modality 1: ℓ_{ji}. The matching pairs are situated along the diagonal of the similarity matrix $(\hat{\mathbf{m}}_i^1, \hat{\mathbf{m}}_i^2)$, which serves as the target for the loss function:

$$t_{ij} = \frac{\exp((\langle \hat{\mathbf{m}}_i^1, \hat{\mathbf{m}}_j^1 \rangle + \langle \hat{\mathbf{m}}_i^2, \hat{\mathbf{m}}_j^2 \rangle)/2 \cdot \tau)}{\sum_{k=1}^{K} \exp((\langle \hat{\mathbf{m}}_i^1, \hat{\mathbf{m}}_k^1 \rangle + \langle \hat{\mathbf{m}}_i^2, \hat{\mathbf{m}}_k^2 \rangle)/2 \cdot \tau)} \tag{2}$$

The ultimate training loss $\mathcal{L}$ (3) is computed by combining the two losses ℓ_{ij} and ℓ_{ji} and averaging them over all pairs within each minibatch.

$$\mathcal{L} = \frac{1}{2 \cdot K} \sum_{i=1}^{K} \sum_{j=1}^{K} t_{ij} \cdot \ell_{ij} + t_{ji} \cdot \ell_{ji} \tag{3}$$

In LightCRL, the context modality identifier c_l offers two main benefits. Firstly, it allows a unified encoder (DFE) f to process embeddings from different pre-trained encoders (f_1 and f_2), enhancing semantic representation while reducing parameters. Secondly, the context modality vector $g_3(c_l)$ acts as an implicit prior, enriching cross-modal representation [11]. LightCRL's contrastive learning ensures efficient alignment across modalities, facilitating easy transfer to various tasks, thus enhancing its applicability.

3 Experiments

In this section, we apply LightCRL to the COCO Captions dataset [7], consisting of $330K$ images ($\mathbf{M}_1$) with 5 captions ($\mathbf{M}_2$) per image.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

	Accuracy				
Model	top-1	top-2	top-3	top-4	top-5
conVIRT	62.12	76.53	84.26	89.22	92.95
DFE-add (LightCRL)	**78.26**	**88.69**	92.5	94.87	96.48
DFE-dot (LightCRL)	75.62	87.37	**92.62**	**95.26**	**97.00**

Table 1: Assessment of zero-shot classification capabilities on the CIFAR-10 dataset. Here, DFE-add signifies the DFE using addition as the fusion method, and DFE-dot indicates the DFE employing scaled dot-product attention for the fusion process.

Parameter size gain: In our Lightweight Cross-Modal Representation Learning (LightCRL) framework, we use pre-trained BERT (12 layers, 110M parameters) and Vision Transformer (ViT) (12 layers, 86M parameters) to encode captions, keeping these models' parameters frozen to minimize the number of trainable parameters [2, 4]. Our training focuses exclusively on the Deep Fusion Encoder (DFE), a Transformer block with 4-head attention with just 1M parameters, employing two fusion methods: addition (DFE-add) and scaled dot-product attention (DFE-dot). This approach drastically reduces the computational burden compared to training all model parameters. The models are trained for 500 epochs with early stopping criteria.

Zero-shot image classification: Utilizes existing capabilities to classify images without prior training on specific classes. It involves computing feature embeddings for both images and text names of all classes within the CIFAR-10 dataset. Then, the cosine similarity of these embeddings is calculated and normalized into a probability distribution using softmax. Table 1 demonstrates the superior performance of LightCRL models (DFE-add and DFE-dot) on CIFAR-10 without retraining. These models, pretrained on COCO Captions and adapted to CIFAR-10, outperform the resource-intensive conVIRT approach, despite CIFAR-10 containing categories absent in COCO Captions.

Linear Classification: Using pre-trained ConVIRT and DFEs models, we apply a linear classifier to encode CIFAR-100 images. Only the linear classifier is trained for 100 epochs, while the pre-trained models remain frozen. This approach assesses the quality of extracted image features with the pre-trained DFEs. Validation on the test set occurs every 20 epochs during training. Table 2 displays the superior performance of LightCRL models (DFE-add and DFE-dot) for linear classification, echoing the findings from zero-shot classification. By utilizing pre-trained DFEs without image augmentation and keeping them frozen, LightCRL models outperform the conVIRT approach. This suggests that our method yields higher-quality image features, enhancing discrimination across categories and improving classification performance.

Fine-tuning: We follow a similar strategy as linear classification, but unlike in linear classification, we unfreeze the pre-trained ConVIRT and DFEs. This approach closely simulates real-world scenarios where pre-trained models weights are adjusted. We evaluate this technique on the Tiny ImageNet dataset,

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

	Epoch				
Model	20	40	60	80	100
conVIRT	55.67	59.22	61.09	62.14	62.71
DFE-add (LightCRL)	**64.63**	**66.78**	**67.93**	**68.88**	**69.37**
DFE-dot (LightCRL)	62.75	63.37	64.57	64.98	65.29

Table 2: Evaluating the precision of linear classification tasks on the CIFAR-100 dataset. Here, DFE-add signifies the DFE using addition as the fusion method, and DFE-dot indicates the DFE employing scaled dot-product attention for the fusion process.

	Epoch				
Model	20	40	60	80	100
conVIRT	59.00	62.71	64.45	65.37	65.89
DFE-add (LightCRL)	**64.49**	**66.31**	**67.60**	**68.12**	**68.91**
DFE-dot (LightCRL)	64.21	66.09	66.70	67.74	68.18

Table 3: Assessing the accuracy of fine-tuning outcomes on the Tiny ImageNet dataset. DFE-add denotes the DFE model that employs addition as its fusion technique, while DFE-dot signifies the DFE model using scaled dot-product attention for its fusion method.

training models for 100 epochs without using data augmentation. Validation on the test set is conducted every 20 epochs during training. Table 3 confirms the results from zero-shot and linear classification experiments. Models pre-trained with LightCRL method offer strong feature representation, serving as robust backbones across tasks.

4 Conclusion

We introduce LightCRL, a cost-effective method for multimodal representation learning. LightCRL's efficiency reduces reliance on extensive datasets and long training times. It utilizes frozen pre-trained models for encoding modalities, with training focused solely on the DFE. DFE employs uniform parameters for different modalities and incorporates a context modality identifier for relevant representation. LightCRL offers a versatile and robust cross-modal representation framework, demonstrated through experiments aligning text and image modalities.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Benjamin Steiner, Matthias Hein, Hugo Touvron, Antoine Hervieu, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Alice Johnson and et al. Cross-modal attention mechanisms for multimodal embedding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[6] Chen Li, Xiaodan Wei, Yuchen Zhai, Junyang Li, Jianfeng Gao, and Jiawei Chen. Visualgpt: Data-efficient image generation using guided variational autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9707–9715, 2021.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *Lecture Notes in Computer Science*, 8693:740–755, 2014.

[8] Indigo JD Orton. Vision based body gesture meta features for affective computing. *arXiv preprint arXiv:2003.00809*, 2020.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[10] John Smith and et al. Multimodal fusion techniques for image and text data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[11] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*, 2021.

[12] Lei Zhang, Miguel Á Carreira-Perpiñán, and Aron Lavie. Convirt: Contrastive visual representation learning for medical imaging. *IEEE Transactions on Medical Imaging*, 41(2):519–531, 2022.