### **CMA-CLIP: CROSS-MODALITY ATTENTION CLIP FOR TEXT-IMAGE CLASSIFICATION**

Jinmiao Fu<sup>\*</sup> Shaoyuan Xu<sup>\*</sup> Huidong Liu Yang Liu Ning Xie Chien-Chih Wang Jia Liu Yi Sun Bryan Wang {jinmiaof, shaoyux, liuhuido, yliuu, xining, ccwang, hliujia, yisun, brywan}@amazon.com Amazon Inc.

#### ABSTRACT

Multi-modal learning with both text and images benefits multiple applications, such as attribute extraction for e-commerce products. In this paper, we propose Cross-Modality Attention Contrastive Language-Image Pre-training (CMA-CLIP), a new multi-modal architecture to jointly learn the fine-grained inter-modality relationship. It fuses CLIP with a sequencewise attention module and a modality-wise attention module. The network uses CLIP to bridge the inter-modality gap at the global level, and uses the sequence-wise attention module to capture the fine-grained alignment between text and images. Besides, it leverages a modality-wise attention module to learn the relevance of each modality to downstream tasks, making the network robust against irrelevant modalities. CMA-CLIP outperforms the state-of-the-art method on Fashion-Gen by 5.5% in accuracy, achieves competitive performance on Food101 and performance on par with the state-of-the-art method on MM-IMDb. We also demonstrate CMA-CLIP's robustness against irrelevant modalities on an Amazon dataset for the task of product attribute extraction.

*Index Terms*— Multimodal, attention, NLP, computer vision

# 1 Introduction

Modern Web systems such as e-commerce and social media contain rich contents expressed in text and images. Leveraging information from both modalities can improve the performance of downstream tasks, such as classification and recommendation. Existing multi-modal learning methods can be classified into two main categories: one-stream methods [1, 2, 3, 4, 5] and two-stream methods [6, 7, 8, 9, 10, 11, 12, 13, 14]. One-stream methods directly feed the features of different modalities into a Transformer [15], while twostream methods process the text and images using separate single-modality networks, such as Transformer [15] and vision transformer (ViT) [16]. A common challenge shared between both methods is the inter-modality gap, i.e., different modalities need different levels of processing due to their inherent complexity. To bridge this gap, research from both categories of methods have been focusing on pretraining the network with different datasets and tasks. For example, VL-BERT [2] pretrains the network on the Conceptual Captions dataset [17] for the task of Masked Language Modeling with Visual Clues and Masked RoI Classification with Linguistic Clues where certain text tokens and image patches are randomly masked, and the network is pretrained to reconstruct the masked components. However, these pretraining tasks typically involve complex settings. For example, in Masked RoI Classification with Linguistic Clues, the object categories of the masked image patches are required for the pretraining. In comparison, new multi-modal architectures to handle the inter-modality gap have been rarely explored.

Several two-stream architectures [11, 12, 13, 14] have been developed to explicitly bridge the inter-modality gap. For example, CLIP [12] trains a text encoder and an image encoder using contrastive learning, so that the text and image features from the same pair are as close as possible. Although such methods can align text and images at a global level, they do not incorporate the fine-grained relationship between text tokens and image patches, which is critical for downstream tasks such as fine-grained classification. Another challenge which is rarely explored is that, the text or image modality could be irrelevant to the downstream tasks. For example, for a product being sold on an e-commerce website, a product title "short-sleeve men's casual t-shirt" is irrelevant to the downstream task of extracting color attribute. In such case, naively leveraging both modalities will jeopardize the model performance, as the irrelevant modality provides nothing but noise.

To tackle the aforementioned challenges, in this paper we propose CMA-CLIP, a new multi-modal architecture which fuses both one-stream and two-stream methods. It leverages the pretrained CLIP, a two-stream method, to close the inter-modality gap at the global level. Subsequently, we add a sequence-wise attention module, which is a transformer as used in most one-stream methods, to capture the finegrained relationship between text tokens and image patches. Moreover, we propose a modality-wise attention module to learn the relevance of each modality to the downstream tasks, which significantly improves the network's robustness to irrelevant modalities. The major contributions of our works are as follows:

<sup>\*</sup>These authors contributed equally to this research work.

- We fuse the CLIP and the sequence-wise attention module to help the network close inter-modality gap, while enabling the network to capture the fine-grained correlation between text tokens and image patches.
- We design a new modality-wise attention module to improve the network's robustness against irrelevant modalities.
- We develop task-specific modality-wise attentions and multi-layer perceptron (MLP) heads that enable the network to perform multi-task classification.

## **2 Proposed framework**

The architecture of CMA-CLIP is depicted in Figure 1. Given a text-image pair, CLIP first converts them into a sequence of text embeddings and a sequence of image embeddings that are aligned globally. Subsequently, the sequence-wise attention module, a stack of transformer blocks, updates the two sequences of embeddings by incorporating the fine-grained correlation between text tokens and image patches. Then the resulting text embedding  $T'_{[CLS]}$  and image embedding  $I'_{[CLS]}$  are weighed by the modality-wise attention module based on their relevance to the downstream tasks. The weighted sum between  $T'_{[CLS]}$  and  $I'_{[CLS]}$  is used for the classification through an MLP. At last, adding task-specific modality-wise attentions and MLPs enables CMA-CLIP to perform multiple-task classification.



Fig. 1: The architecture of CMA-CLIP.

#### 2.1 Contrastive Language-Image Pretraining

CLIP consists of a text encoder and an image encoder. In this work, we use Transformer [15] as the text encoder and ViT [16] as the image encoder. For each text-image pair, the text encoder projects the text into a sequence of text embeddings  $T_1, ..., T_n$  corresponding to different text tokens, and a text embedding  $T_{[CLS]}$  of an artificial token appended at the end to represent the whole text. Similarly, the image encoder projects the image into a sequence of image embeddings  $I_1$ , ...,  $I_m$  corresponding to different image patches, and an image embedding  $I_{[CLS]}$  of an artificial patch prepended at the beginning to represent the whole image. CLIP is trained to maximize the cosine similarity between  $T_{[CLS]}$  and  $I_{[CLS]}$  from the paired text and image, and minimize the cosine similarity of the unpaired ones using WebImageText (WIT) Dataset, which contains 400 million text-image pairs collected from the Web.

#### 2.2 Sequence-wise attention

CLIP is able to train the text and image encoders, so that the corresponding embeddings  $T_{[CLS]}$  and  $I_{[CLS]}$  are aligned at the global level. However, the fine-grained correlations between text tokens and image patches are not learned. Such information is critical for certain downstream tasks, where only a subset of text tokens and image patches are informative, such as classifying the sleeve type of a t-shirt product on e-commerce websites. To address this limitation, we add a sequence-wise attention module. It takes the concatenated sequence of text and image embeddings  $\{T_{[CLS]}, T_1, ..., T_n, ..., T_n,$  $I_{[CLS]}, I_1, ..., I_m$  from CLIP as input. The sequence-wise attention module is a stack of transformer blockers as in BERT [18]. We omit the details of BERT in this section. At the end, the  $T'_{[CLS]}$  and  $I'_{[CLS]}$  generated by the sequence-wise attention module are considered as the updated text and image embeddings which incorporate the fine-grained correlation between text tokens and image patches.

#### 2.3 Modality-wise attention

A common challenge in practice is that, the text or image inputs could be irrelevant to downstream tasks. For example, for a product being sold on e-commerce websites, a product title "short-sleeve men's casual t-shirt" is completely irrelevant to the downstream task of extracting color attribute. In order to dampen the impact of irrelevant modality, we propose a modality-wise attention module to learn the relevance of each modality, so that we can scale the text and image embedding by their relevance before aggregating them for classification. Specifically, we use a learnable parameter vector w to project  $T'_{[CLS]}$  and  $I'_{[CLS]}$  into two scalars to reflect their relevance as  $e_T = w^T T'_{[CLS]}$  and  $e_I = w^T I'_{[CLS]}$ .

relevance as  $e_T = w^T T'_{[CLS]}$  and  $e_I = w^T I'_{[CLS]}$ . Subsequently, we aggregate  $T'_{[CLS]}$  and  $I'_{[CLS]}$  as  $\lambda T'_{[CLS]} + (1 - \lambda)I'_{[CLS]}$ , where  $\lambda$  is the normalized relevance of the text embedding as  $\lambda = \frac{exp(e_T)}{exp(e_T) + exp(e_I)}$ . For any classification task, an MLP head is added on top

For any classification task, an MLP head is added on top of the aggregated feature. For multitask classification, we add task-specific modality-wise attention and MLP for each task separately. This is because the relevance of modality is dependent on the task.

### **3** Experiments

#### 3.1 Dataset

We compare CMA-CLIP with state-of-the-art multi-modal learning methods on three public datesets, Fashion-Gen [19], Food101 [20] and MM-IMDb [21]. We also demonstrate

CMA-CLIP's robustness again irrelevant modalities on APA (Amazon Product Attribute), a dataset collected from Amazon.com for the task of product attribute extraction. Due to legal concern, we cannot provide reference to this dataset.

#### 3.1.1 Fashion-Gen

This dataset contains 293,008 fashion images. Each image is paired with a text describing the image. It contains 121 subcategories, such as "SHORT DRESSES" and "LEATHER JACKETS". We use the same data as used in [22] for training and testing. The number of training data is 260,480, and the number of testing data is 32,528.

### 3.1.2 Food101

This dataset contains 101 food categories. The goal is to classify each text-image pair to a food category. We download the preprocessed images and texts from the Kaggle competition<sup>1</sup>. In the processed data, 67,971 images are in the training set, and 22,715 images are in the testing set. During training, we randomly split 80% of the data in the training set for training and the rest 20% data for validation.

#### 3.1.3 MM-IMDb

We use MM-IMDb to test CMA-CLIP's performance on multi-label classification. This dataset consists of the movie plots and the corresponding movie posters for 25,888 movies. The goal is to classify each movie into one or more of the 23 genres such as "Action" and "Horror". We assign 15,510 movies in training data, 2,599 movies in validation data, and 7,779 movies in test data.

### 3.1.4 APA

The text-image pairs in the above-mentioned three public datasets are all relevant to their corresponding tasks. Therefore, we cannot demonstrate CMA-CLIP's effectiveness against irrelevant modalities. To address this limitation, we collect the product title-image pairs for 6 million dress products from Amazon.com. The objective is to classify two product attributes, color and pattern. Color has 17 classes such as black and white, and pattern has 12 classes such as graphic and plain. We crawl the catalog system to fetch the labels of color and pattern attributes for those 6 million products. In addition, we manually annotate the color and pattern attributes for another 600 image-title pairs as the validation and test set, which are used for hyper-parameter tuning and performance evaluation respectively. Due to the limited auditing resource, we only annotate 600 products.

### 3.2 Implementation

#### 3.2.1 Experiment settings

Same as CLIP, the text encoder of CMA-CLIP is a 12-layer 512-width Transformer with eight heads used in [15], and

the image encoder of CMA-CLIP is a 12-layer 768-width ViT-B/32 [16] with twelve attention heads. The sequencewise attention transformer is a 12-layer 512-width model with eight attention heads. In all the experiments, we use an AWS p3.16xlarge instance with 8 GPUs for model training. The batch size is set to 1024, weight decay of Adam is set to 1e-4, and the learning rate is set to 1e-5.

#### 3.2.2 Training strategy

We use the pre-trained weights of CLIP as the initial weights of the text and image encoders in CMA-CLIP. We randomly initialize the weights in the sequence-wise attention module, modality-wise attention module, and MLP. As CMA-CLIP contains a mixture of pre-trained weights and randomly initialized weights, instead of training the model end-to-end which may cause under- or over-fitting of certain modules, we adopt a multi-stage training strategy:

- Warm-up stage: In this stage, the weights of the text and image encoders are frozen. We train the sequencewise attention, the modality-wise attention and the MLPs.
- End-to-end training stage: In this stage, we unfreeze the weights of the text and image encoders, and train all the components together.
- **Tuning stage:** This stage is only required for multi-task training. In this stage, we only train the modality-wise attentions and MLPs.

For Fashion-Gen, Food101 and MM-IMDb, the warm-up stage is trained for 100 epochs and the end-to-end training stage is trained for 300 epochs. Since these datasets only involve single-task classification, the tuning stage is not needed. At the end of each training stage, the model checkpoint with the lowest validation loss is used as the starting weights for the next stage. For APA, all three stages are trained for 20 epochs due to earlier convergence. The training time for all 3 public datasets is approximately 1 day and the training time for the APA dataset is approximately 3 days. Our code will be available on GitHub soon.

### 3.3 Results

#### 3.3.1 Fashion-Gen

On the Fashion-Gen dataset, we compare CMA-CLIP with multiple SOTA methods, including FashionBERT [23], ImageBERT [24], OSCAR [4], and KaleidoBERT [22]. Results are included in Table 1. CMA-CLIP achieves the highest accuracy of 93.6%, which improves over KaleidoBERT, the current SOTA method, by 5.5%. KaleidoBERT is pretrained on tasks including Aligned Masked Language Modeling, Image and Text Matching, and Aligned Kaleido Patch

<sup>&</sup>lt;sup>1</sup>https://www.kaggle.com/gianmarco96/upmcfood101

Method	Accuracy	Method	Accuraccy
FashionBERT	85.3	ViT [12]	81.8
ImageBERT	80.1	BERT [25]	87.2
OSCAR	84.2	CLIP [12]	88.8
KaleidoBERT	88.1	MMBT	92.1
CMA-CLIP	93.6	CMA-CLIP	93.1

Table 1: Accuracy (%)	Table 2:Accuracy (%)
on Fashion-Gen	on Food101

Modeling using the Fashion-Gen dataset. Then the network is fine-tuned for the Fashion-Gen classification. In comparison, CMA-CLIP does not require any specific pretraining on Fashion-Gen. The superior performance of CMA-CLIP indicates the strength of the fusion between one-stream and twostream methods.

#### 3.3.2 Food101

On the Food101 dateset, we compare CMA-CLIP with two single-modality baseline methods including BERT [18] and ViT [16], two multi-modality baseline methods including MMBT [25] and CLIP using a linear probe [12] with same ViT-B/32. Results are included in Table 2. CMA-CLIP achieves the best accuracy of 93.1%, which improves 1% over the a strong baseline method MMBT. MMBT is a one-stream method that uses a transformer to capture the fine-grained correlations between text tokens and image patches. The superior performance of CMA-CLIP indicates the strength of the fusion between one-stream and two-stream methods.

#### 3.3.3 MM-IMDb

On the MM-IMDb dateset, we compare CMA-CLIP with MMBT [25], the current SOTA multi-modal method. Since this is a multi-label classification problem, we compare the performance in terms of micro F1 and macro F1. As shown in Table 3 (the performance of MMBT is different from what is reported in [25] because we retrain MMBT by excluding movies that belong to some minority genres that are out of the 23 genres for a fair comparison with CMA-CLIP), CMA-CLIP has better micro-F1 but worse macro-F1. Overall, CMA-CLIP's performance is on par with MMBT.

Method	Micro-F1	Macro-F1
MMBT	63.7	55.5
CMA-CLIP	65.3	52.7

Table 3: Micro and Macro F1 (%) on the MM-IMDb.

# **4** Ablation study

To demonstrate CMA-CLIP's robustness against irrelevant modality, we first test the performance of CMA-CLIP with and without the modality-wise attention (MWA) by extracting color and pattern attributes using the APA dataset. After removing MWA, the recall at 90% precision drops from 61.1%

to 60.0% for color attribute, and from 76.3% to 67.9% for pattern attribute. We further remove the sequence-wise attention (SWA). The recall at 90% precision further drops from 60.0% to 57.3% for color attribute, and from 67.9% to 60.3% for pattern attribute. The impact on the pattern attribute is larger because the percentage of titles with no pattern information is 75%, higher than 33% for the color attribute. In Table 4 we randomly pick some product title-image examples that CMA-CLIP can produce correct classification whereas CMA-CLIP without MWA cannot. We can observe from these examples that, the product titles do not contain any tokens related to the attributes. Furthermore, after we insert the attribute related keywords to the titles, CMA-CLIP with and without MWA can both produce correct classification. This proves MWA's ability to filter out irrelevant modality.

Attribute Ima	age Title	Label	Prediction w/ & w/o MWA
Color	Portland T-Shirt Dress	Black	Black & Blue
Pattern	Women's Mini Dun- garee	Plain	Plain & Graphic

**Table 4**: Examples where CMA-CLIP is able to produce the correct attribute classification while CMA-CLIP w/o the modality-wise attention cannot.

# 5 Conclusion

In this paper, we propose CMA-CLIP, a new multi-modal architecture. It fuses the pretrained CLIP and the sequence-wise attention, which helps the network close inter-modality gap, while enabling the network to capture the fine-grained correlation between text tokens and image patches. We propose the modality-wise attention, which learns relevance of each modality to dampen the impact from the irrelevant one for downstream tasks. We add task specific modality-wise attentions and MLPs so that we can leverage a unified network for multi-task classification. We evaluate our method on the Fashion-Gen, Food101 and MM-IMDb datasets. It surpasses the SOTA method on the Fashion-Gen dataset by 5.5% in accuracy, achieves competitive performance on the Food101 dataset and performance on par with the SOTA on the MM-IMDb dataset. We also demonstrate CMA-CLIP's robustness against irrelevant modality using the APA dataset.

## **6** References

[1] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline

for vision and language," *arXiv*, vol. abs/1908.03557, 2019.

- [2] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," in *ICLR*, 2020.
- [3] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *ECCV*, 2020.
- [4] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., "Oscar: Object-semantics aligned pre-training for visionlanguage tasks," in *ECCV*, 2020.
- [5] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers," *ArXiv*, vol. abs/2004.00849, 2020.
- [6] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019.
- [7] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - a new approach to self-supervised learning," in *NeurIPS*, 2020.
- [8] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *arXiv preprint arXiv:2006.09882*, 2020.
- [9] K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations," in *CVPR*, 2021.
- [10] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *NeurIPS*, 2020.
- [11] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," *arXiv*, vol. abs/2010.00747, 2020.
- [12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.
- [13] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," *arXiv*, vol. abs/2102.05918, 2021.

- [14] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, and S. Marchand-Maillet, "Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 17, pp. 1–23, 2021.
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, vol. 30.
- [16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.
- [17] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in ACL, 2018.
- [18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, vol. abs/1810.04805, 2019.
- [19] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal, "Fashion-gen: The generative fashion dataset and challenge," *arXiv*, vol. abs/1806.08317, 2018.
- [20] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *ECCV*, 2014.
- [21] J. Arevalo, T. Solorio, M. M. y Gómez, and F. A. González, "Gated multimodal units for information fusion," 2017.
- [22] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao, "Kaleido-bert: Vision-language pre-training on fashion domain," in *CVPR*, 2021.
- [23] D. Gao, L. Jin, B. Chen, M. Qiu, P. Li, Y. Wei, Y. Hu, and H. Wang, "Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval," 2020.
- [24] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, "Imagebert: Cross-modal pre-training with largescale weak-supervised image-text data," *arXiv preprint arXiv:2001.07966*, 2020.
- [25] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv*, vol. abs/1909.02950, 2019.