

Article

AI Ekphrasis: Multi-Modal Learning with Foundation Models for Fine-Grained Poetry Retrieval

Muhammad Shahid Jabbar ¹, Jitae Shin ^{1,*} and Jun-Dong Cho ^{1,2,*}

¹ Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, Korea; eeshahid@skku.edu

² Department of Human ICT Convergence, Sungkyunkwan University, Suwon 16419, Korea

* Correspondence: jtshin@skku.edu (J.S.); jdcho@skku.edu (J.-D.C.)

Abstract: Artificial intelligence research in natural language processing in the context of poetry struggles with the recognition of holistic content such as poetic symbolism, metaphor, and other fine-grained attributes. Given these challenges, multi-modal image–poetry reasoning and retrieval remain largely unexplored. Our recent accessibility study indicates that poetry is an effective medium to convey visual artwork attributes for improved artwork appreciation of people with visual impairments. We, therefore, introduce a deep learning approach for the automatic retrieval of poetry suitable to the input images. The recent state-of-the-art CLIP provides a way for multi-modal visual and text features matched using cosine similarity. However, it lacks shared cross-modality attention features to model fine-grained relationships. The proposed approach in this work takes advantage of strong pre-training of the CLIP model and overcomes its limitations by introducing shared attention parameters to better model the fine-grained relationship between both modalities. We test and compare our proposed approach using the expertly annotated MiltiM-Poem dataset, which is considered the largest public image–poetry pair dataset for English poetry. The proposed approach aims to solve the problems of image-based attribute recognition and automatic retrieval for fine-grained poetic verses. The test results reflect that the shared attention parameters alleviate fine-grained attribute recognition, and the proposed approach is a significant step towards automatic multi-modal retrieval for improved artwork appreciation of people with visual impairments.

Keywords: image-based poetry retrieval; fine-grained attribute recognition; accessibility; multi-modal attention; cross-encoder



Citation: Jabbar, M.S.; Shin, J.; Cho, J.-D. AI Ekphrasis: Multi-Modal Learning with Foundation Models for Fine-Grained Poetry Retrieval. *Electronics* **2022**, *11*, 1275. <https://doi.org/10.3390/electronics11081275>

Academic Editor: George A. Tsihrintzis

Received: 12 March 2022

Accepted: 14 April 2022

Published: 18 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Poets often encompass sentiments, themes, and messages they intend to articulate implicitly through poetic verses. This implicit artistic conception by a poet is a unique feature of human-authored poetry as opposed to machine-generated poetry. Additionally, metaphor and symbolism are commonly employed for this type of poetry. Therefore, the message and feelings are characterized by symbolism, scenes, metaphor, activities, objects, and color tones rather than relying merely upon the objects in an image or the color tones. Existing solutions in image–text retrieval mainly focus on the concurrence of objects in an image from a verbal description of objects through image captioning or training on image captioning datasets. As a result of this, two matching poems from the candidate poetry dataset carrying the same notion but expressed differently may be regarded with distant retrieval rankings and vice versa, based on words matching intuition, for instance [1]. Fine-grained artwork and poetry-attribute recognition assume extensive domain knowledge, and, therefore, proper feature learning is a herculean task for conventional methods and classical CNN-based methods. Visually impaired visitors experience visual artwork appreciation limitations, such as a lack of sensory and cognitive access to exhibit artworks or replicas. The visual artworks appreciation opportunities

for people with visual impairments through various senses, such as auditory, tactile, and olfactory senses, already exist and are expanding [2–6]. The multi-sensory use of poetry to express visual artwork or images is an effective medium to convey visual artwork’s holistic content, as indicated by the existing research (Section 2.1). However, there is a research literature potential for automating the matching of visual data with suitable poetry. Moreover, this matching problem becomes more challenging due to the fine-grained features of poetry, in contrast to a mere description of objects and actions in image captions. This work is aimed at addressing the stated research literature gap.

This study aims to enable an improved appreciation experience for people with visual impairments by using automatically retrieved ekphrasis for artwork and to aid an abundant media art exhibition environment. These multi-sensory exhibits provide the users with a more immersive, realistic, and impressive experience. Moreover, they can potentially impart cognitive and emotional impacts on the appreciator. Thus, we present a fine-grained visual–poetry representation methodology on top of a general contrastive pre-training framework based on zero-shot, few-shot, and fully supervised learning. In our previous work, we discovered common semantic directivity through intermediate semantic adjective pairs for both artwork and poems and demonstrated the usability and user appreciation of including manually picked poems for color-coding in the artwork exploration of people with visual impairments [7]. In this work, we also aim to advance that by dissolving the intermediate common semantic directivity stage for fine-grained attribute recognition, automatic retrieval of poems, and expanding the poetic representation of a given image from colors to overall implicit artistic conception. The proposed architecture is presented in Figure 1, and the key contributions of this study are listed as below.

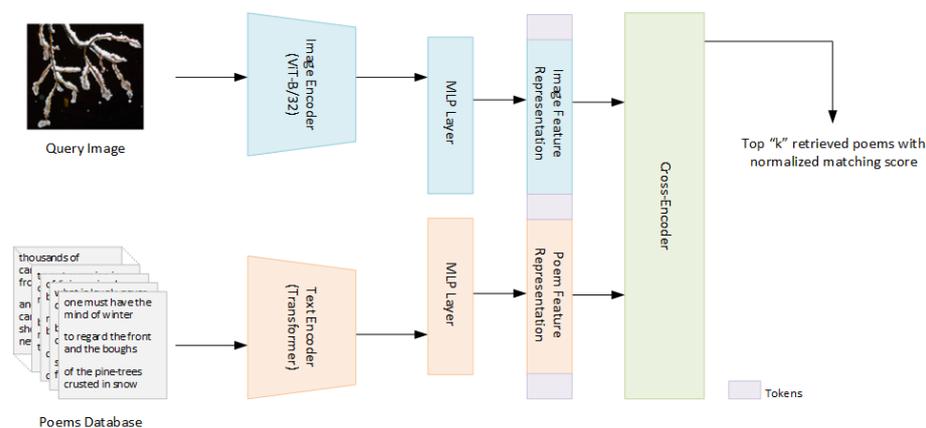


Figure 1. Flow diagram of the proposed approach. We use task agnostic transformers as image and poem encoders to obtain image and poem representations. Next, we concatenate tokens, images, and poem representations. Finally, we use a transformer model in the cross-encoder fashion, allowing shared attention parameters to learn discriminative image–text multi-modal information.

Contributions

1. The developed solution provides multi-modal representation learning about fine-grained poetry for matching images with poetry.
2. We combine the advantages of CLIP’s [8] strong pre-training and the shared attention parameters learning for multi-modal image–poem data. This improves the context awareness of our model by liaising among feature representations of image and poem sequences.
3. Our proposed model leverages the state-of-the-art pre-trained CLIP model and outperforms its zero-shot, few-shot, and fully supervised poetry retrieval performance for the image–poetry retrieval task.
4. The proposed solution considers fine-grained attribute recognition for matching the most relevant poems to a query image, contemplating the mutual association of scenes, sentiments, and objects under ekphrasis considerations of symbolism and metaphors.

This enables automatic poetry retrieval for the visual artwork appreciation of people with visual impairments.

The rest of this work is organized as follows. Section 2 describes the existing literature relevant to this work. Being one of the pioneer works for automatic image-based poetry retrieval for visually impaired peoples' artwork appreciation, a diverse range of relevant studies are discussed in the sub-sections. They cover, in order, a brief introduction to visual artwork appreciation through poetry and multi-sensory methods for visually impaired people; the role of transformer-based models in natural language processing pertaining to our transformer-based proposed method; the notable works involving deep-learning-based methods for images and poetry, where most existing works deal with Chinese poetry generation; and the existing deep-neural-network-based works on the matching of visual and textual data. The adopted baseline deep learning model and the proposed method are presented in Section 3. In Section 4, we detail the dataset utilized in this work along with the experimental setup of our proposed method, including its implementation details and evaluation metrics to measure its performance. We present the evaluation results of the proposed method and discuss its performance in comparison to the baseline method in Section 5. Finally, Section 6 concludes this paper and highlights some potential future work directions.

2. Related Studies

2.1. Multi-Sensory Artwork Poetry Exploration for People with Visual Impairments

Lately, contemporary art has been advancing beyond the mere visual appreciation of artworks. Moreover, the progress in multi-sensory interaction techniques is greatly influencing arts, culture, and exhibitions. It has ultimately provided artwork appreciation solutions through senses other than sight, such as smell [9], touch [10,11], and hearing [12], for people with visual impairments. Cho J.D. et al. [7] presented a multi-sensory color-coding system through the combination of music and poetry, such that manually picked poems represented the primary and secondary colors dimensions of warm and cool for conspicuous colors in a given artwork. They performed an implicit association test to discover common semantic directivity for color dimensions between artwork and candidate poems in order to pick a best-suited poem from a database of candidate poems. They confirmed, through system usability and user tests, that poetry can be effectively used to supplement and enhance visually impaired people's artwork exploration experience.

2.2. Transformers for Natural Language Processing

Transformer-based models [13] have dominated natural language processing (NLP) tasks and applications [14]. First, these foundation models are pre-trained on large text corpora, which can further be fine-tuned for downstream tasks [15]. The masked language modeling (MLM) employed in BERT [16] and permuted language modeling employed in XLNet [17] are two major pre-training objectives. Masked language modeling (MLM) masks some tokens with a masked symbol [MASK] and predicts the masked tokens based on the rest of the tokens. For instance, if tokens x_2 and x_5 are masked in a sequence $x = (x_1, x_2, x_3, x_4, x_5)$, the masked sequence is represented by $x = (x_1, [MASK], x_3, x_4, [MASK])$. This encourages the MLM models to learn and extract better representations of x_2 and x_5 . The MLM can consider the position information for the entire sentence but is unable to learn the complicated semantic relationship well among the predicted tokens due to its inability to model the dependency among them. With permuted language modeling (PLM), a sequence is randomly permuted and a token prediction is returned in an auto-regressive manner in the right part (predicted part). If a given sequence $x = (x_1, x_2, x_3, x_4, x_5)$ is permuted into $(x_1, x_3, x_4, x_5, x_2)$, PLM auto-regressively conditioned on (x_1, x_3, x_4) predicts x_5 and x_2 . In PLM, predicted tokens' dependence is modeled with auto-regressive prediction, but it cannot consider the entire sentence position information, which results in mismatches between pre-training and fine-tuning, since downstream tasks take into account the entire sentence position information. The MPNet model [18] unifies the non-predicted part of

PLM and MLM. For example, if the given sequence $x = (x_1, x_2, x_3, x_4, x_5)$ is permuted into $(x_1, x_3, x_4, x_5, x_2)$, it selects (x_2, x_6, x_5) tokens to the right as the predicted tokens. It then forms the non-predicted part as $(x_1, x_3, x_4, [\text{MASK}], [\text{MASK}], [\text{MASK}])$ by masking and $(p_1, p_3, p_4, p_2, p_6, p_5)$ is the corresponding position information. Instituting this output dependency and input consistency bridges the advantages of both MLM and PLM while avoiding their limitations.

2.3. Multi-Modal Image-Inspired Poetry Generation with Neural Networks

Typical approaches for automatic poetry generation are based on recurrent neural networks (RNN) [19], i.e., sequence-to-sequence encoder–decoder networks, auto-encoders [20], and attention models [13]. Most of these poetry generation systems rely on template-based methods. The content of these generated poems is restricted by the template in these methods, therefore generating monotonous poems. The existing works that look at multi-modal image–poetry using deep learning are mainly focused on poetry generation. Moreover, these related works for image-based poetry generation are developed for particular genres of poetry, predominantly in the Chinese language.

A recurrent neural network (RNN)-based approach for Chinese poetry generation is proposed in [21]. This method leverages the titles to generate quatrain poems, which are pieces of verse consisting of four rhymed lines. Given an image input, a language model is first applied for the generation of the first line of a poem, then a relevant theme is picked by the Latent Dirichlet allocation (LDA) for title generation. Finally, the hierarchy-attention sequence-to-sequence model is applied for the generation of rest of the three lines of the quatrains. The resultant similarity between pairs (2-grams) of generated and ground-truth data is close to 30 by BLEU-2 score. Liu Y. et al. [22] addressed the problem of semantic inconsistency and topic drift in generated Chinese poetry by incorporating abstract and concrete information from input images. They use abstract information embedding and explicitly infill concrete keywords into each line of a generated poem. Their deep learning model is based on a gated recurrent unit (GRU) [23] encoder, an attention mechanism, and a GRU decoder, which generates line-by-line poems. Wu L. et al. [24] focused on image-based-poetry-generated challenges of image–poem semantic consistency, avoiding topic drift and the repetition of words in generated poems. They have employed visual semantic vector construction from images and temporal and depth LSTMs in their topic-aware poetry generation model. Liu L. et al. [25] proposed the Image2Poem model, which considers image streams for the generation of the classical genre of Chinese poetry. It first chooses a representative image from the image stream, and the LSTM-based [26] poetry decoder further generates poem characters by adaptively considering previously generated target poetry characters or input image streams.

Liu B. et al. [27] present a novel approach to the image-based poetry generation problem, by incorporating a CNN-based deep-coupled visual–poetic embedding model for the object, sentiment, and scene features. This is followed by RNN-based adversarial training with multi-discriminators as rewards for policy gradient. Wu C. et al. [28] used an image-based Chinese poem generation network to generate quatrains. Their proposed method includes content, sentiment, and theme extraction of images. The sequence-to-sequence poem theme and style control module finally generates poem quatrains. A Chinese poetry generation approach for the classical genre is proposed by Liu Y. et al. [29]. They first fed the input image into an open-source image annotation service (Clarifai API), and the generated image annotations were then used to retrieve relevant phrases from another open-source poetic phrase taxonomy. A self-attention neural network-based generated further poetry based on embedding vectors of the retrieved poetic phrases, where prefixes and number of words per line may be fixed as a user-defined input. Finally, a beam-search-based screening mechanism screens the generated poetry output based on word repetition and rhyming.

Zhang, D. et al. [30] proposed a recommendation system that takes images at the input and returns recommendations from a poetry database. They posed this problem as

recommending classical Chinese poetic descriptors to go with photos on social media. The problem formulation for this work is based on maximizing object, theme, and sentiment consistency among input images and recommended poems. Their proposed method includes three modules: (1) The conception-aware heterogeneous information network (CaHIN) for modeling the semantic relationships between the sentiments, themes, objects, and metaphors in both images and classical poems; (2) The poetic visual analyzer (PVA) for extraction of notable objects and their descriptions from input images; and (3) The ranking module for latent representation learning and poetry recommendations.

2.4. Deep Neural Networks for Visual and Textual Data Matching

The unified textual and visual attention mechanism for multimodal reasoning and matching through Dual Attention Networks is proposed by Nam H. et al. [31]. Their model for visual question answering infers the answers collaboratively from images and texts. On the other hand, the multi-modal matching model uses separate visual and textual attention memories and leverages the joint training of both modalities to learn shared semantics. Lee K.H. et al. [32] propose a stacked cross-attention mechanism (SCAN) to map the multi-modal embeddings for image–text matching. The SCAN method measures a multi-modal similarity score by determining the most relevant image region. They point out that the particular image regions are mainly responsible for image captions, as image descriptions refer to certain objects and their attributes in an image. This approach is not suitable for our task, as it is pertinent to convey the overall semantic directivity of images through poems.

The visual semantic reasoning network (VSRN) is proposed in [33] for image–text matching. The method first employs Faster-RCNN [34] to extract region-wise features and recognize objects. The graph convolution network is further used to generate semantic relationship features by modeling region relationship reasoning. The VSRN then performs global semantic reasoning, selects discriminative features, and generates input image representations. These representations are mapped to text captions by jointly optimizing them using the Gated recurrent unit (GRU)-based [23] text decoder. Moreover, recent work in [35] has argued that visual artworks' fine-grained attribute recognition can be achieved by fine-tuning the contrastive learning framework foundation models.

The existing works, including Section 2.3, have devised multiple solutions for the synthesis of new machine-generated poetry. Similar to other forms of artists' creations, human-crafted poems come with their own aesthetic forms. They carry a deeper essence unparalleled by machine-synthesized poetry. Advances in deep learning have led to the development of artificial intelligence systems for multi-modal similarity or matching systems, such as image and text matching (Section 2.4). Most of these have demonstrated their functionality with image captioning datasets, where a text caption simply used to try to verbally describe objects in a given image. Art, on the other hand, relies on conveying indirect inferences of sentiments, rather than direct descriptions. These inferences can be ascertained by aspects such as color temperatures, scenes, and objects for visual artworks and through symbolism and metaphor for poems.

3. Approach

3.1. Baseline Model: Contrastive Language-Image Pre-Training (CLIP)

The Contrastive Language-Image Pre-Training (CLIP) as shown in Figure 2 [8] consists of one image encoder and one text encoder, where the text encoder is based on a transformer [13], with few architectural modifications as suggested in [36]. The vocabulary size of its tokenizer is 49,152 words, in uncased configuration. The transformer consists of 8 attention heads, with a 12-layer model of width 512. The image encoder of CLIP has two architecture variants: a CNN-based ResNet-50 [37] and a transformer-based Vision Transformer (ViT) [38]. The text and image encoders take tokenized text and images as their inputs and return their feature embeddings to be projected to the multi-modal embedding space. During training, the objective of the CLIP model is to minimize the cosine similarity

of the image and text embedding pairs and maximize it for unpaired embedding vectors. At the inference time for image-based text retrieval, the cosine similarity scores of query image feature embedding are computed with feature embeddings of text instances in a given database. These values can then be sorted to return top-K matched texts with the highest cosine similarity scores. The pre-training of CLIP is performed on a private WIT dataset with 400 M image–text pairs drawn from the internet. This strong pre-training of the foundation model enables zero-shot classification tasks, where a classification task is performed on images or datasets without the need to train the model. CLIP’s zero-shot inference results exceed the ResNet50 linear probe’s fully supervised results for a variety of datasets including StanfordCars [39], Country211 [8], Food101 [40], and UCF101 [41].

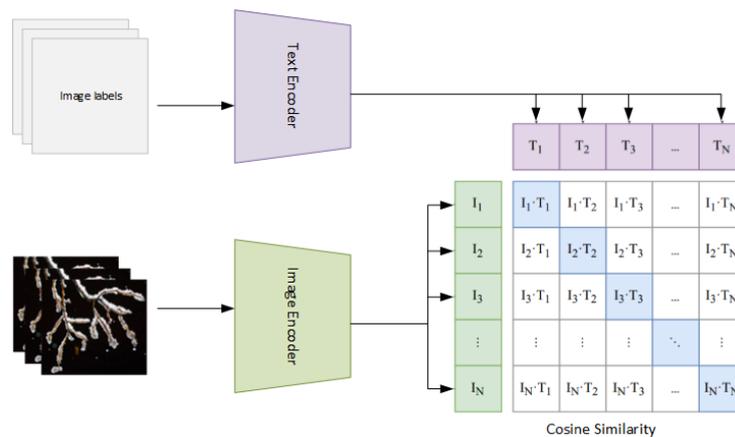


Figure 2. CLIP [8] model as a contrastive pre-training approach.

3.2. Cross-Encoder CLIP

Let A and B be two paired vectors. We can perform pair score and pair classification tasks using transformer [13] encoders in cross-encoder and bi-encoder configurations. Both vectors can simultaneously be passed through the cross-encoder transformer network. At the training stage, the output labels of “1” and “0” are assigned for similar and non-similar input vector pairs, respectively. At the inference stage, it returns output values ranging from 0 to 1 reflecting the similarity among input vector pairs. The cross-encoders do not provide vector embeddings, and individual vectors cannot be passed through them. In contrast to cross-encoders, bi-encoder transformer models provide vector embeddings for individual input vectors as shown in Figure 3. Consequently, vectors can be independently passed through bi-encoders to acquire vector embeddings. The similarity scores of these vector embeddings can be computed through vector similarity methods such as cosine similarity.

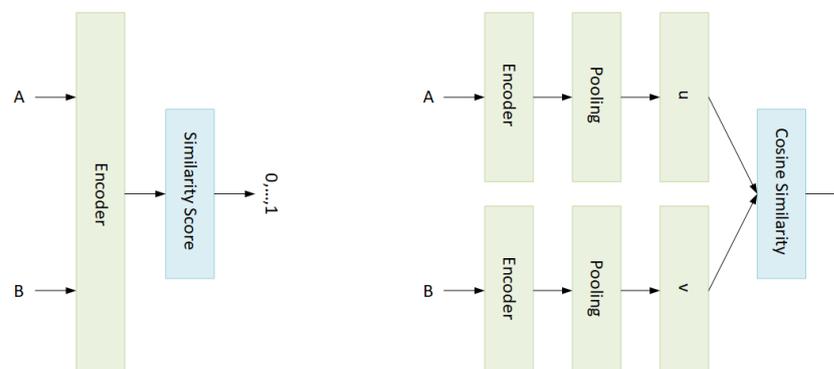


Figure 3. Cross-Encoder (left) versus Bi-Encoder (right) Configuration.

The transformer-based models have dominated NLP tasks and applications. However, architectures based on transformer models are now attaining excellent performances on

computer vision tasks, while utilizing significantly fewer computational resources for training [38,42]. Our proposed cross-encoder module combines visual and textual feature modeling into one transformer-based architecture. In this configuration, the image and poem extracted feature embeddings are mapped to obtain A and B as paired vectors, corresponding to input image and poem pairs. These paired vectors are concatenated as an input to the cross-encoder module. Our cross-encoder module is based on masked language modeling BERT [16] with modifications described in [18] for jointly exploiting permuted language modeling [17]. This architecture is chosen for its proven performance and increasing adoption. The input to the cross-encoder module consists of paired vectors as multi-modal feature representations instead of text sentences. Therefore, the use of a tokenizer to covert text inputs into numeric representations is no longer required. However, we provide token type IDs and mask tokens as per [18].

The vision transformer [38] and transformer model [13] with CLIP's contrastive pre-training [8] focuses on global feature extraction from images and poems. These extracted feature embeddings are mapped through one MLP layer each for an onward cross-encoder module input. The architecture of our cross-encoder CLIP (CE-CLIP) model is presented in Figure 1. In order to model the context-aware fine-grained features from CLIP-extracted embeddings for image patches and poem tokens, we capitalize on the shared attention mechanism in the cross-encoder module. This is particularly helpful for the image–poem pairs wherein the objects in given images are not depicted by words and descriptions. Instead, the essence of an image is represented through poetic symbolism and metaphor. The proposed algorithm with a training process is mentioned in Algorithm 1.

Algorithm 1 Training Process of Proposed Method.

Input: image–poem pairs (I_i, P_j) , and their Labels ($L = 1$ if $j = i$, else 0)

```

1: Training:
2: if Stage = Warm – up then
3:   Freeze Image_Encoder and Text_Encoder parameters update
4: else if Stage = Fine – tune then
5:   pass
6: end if
7: for epoch in range(MaxEpochs) do
8:    $I_{emb} = Image\_Encoder(I)$ 
9:    $P_{emb} = Text\_Encoder(P)$ 
10:  ▷ Image and Poem feature representation using Image and Text encoders from CLIP model
11:   $\hat{I}_{emb}, \hat{P}_{emb} = MLP(I_{emb}), MLP(P_{emb})$ 
12:  ▷ Respective Multi-Layer Perceptron (MLP) layers for Image and Poem feature representation
13:   $\hat{I}\hat{P} = Concatenate([CLS], \hat{I}_{emb}, [SEP], \hat{P}_{emb}, [PAD], [SEP])$ 
14:  ▷ Concatenation of tokens and feature representations, and get input IDs, token type IDs, and attention mask
15:   $\hat{L} = CE(\hat{I}\hat{P})$ 
16:  ▷ Apply cross-encoder to obtain prediction output
17:   $Loss = Loss\_Function(L, \hat{L})$ 
18:  ▷ Loss computation between ground truth label (L) and prediction ( $\hat{L}$ )
19:   $\theta_{epoch+1} = Optimizer(Loss, \theta_{epoch})$ 
20:  ▷ Update model parameters
21: end for

```

In CE-CLIP, the sequence-wise attention features are shared among latent image and text representations in the cross-encoder module. Let d be the embedding dimension of the embedding vectors at the outputs of CLIP encoders followed by mappings through MLP layers and concatenated with tokens, and s be the length of the input sequence and the embedding vectors. Then, $X \in \mathbb{R}^{s \times d}$ represents the matrix for the sequence of embedding

vectors at the input of the cross-encoder. Additionally, the projection matrices $W_K \in \mathbb{R}^{s \times d}$, $W_Q \in \mathbb{R}^{s \times d}$, and $W_V \in \mathbb{R}^{s \times d}$ project each embedding in X to the key, query, and value spaces, respectively:

$$K = XW_K, Q = XW_Q, Q = VW_V. \quad (1)$$

The parameters of the embedding matrix X are updated as below [13,43];

$$Attention(K, Q, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (2)$$

For each pair in the feature embedding vector from X , the self-attention block learns a similarity matrix QK^T . The sequence embeddings are kept updated as the projected embedding average across all the value space embeddings weighted by their similarities. These sequence-wise shared attention parameters help to model the fine-grained relationships and dependencies among each image and the paired poem.

4. Experiments

4.1. Dataset

The MultiM-Poem dataset is a collection of 8,292 image and poem pairs [27]. It is a subset of a larger dataset (multiM-Poem-Ex, size = 26,161 pairs) collected from the internet, targeting human-written free-form poems to illustrate paired images. Considering the concordance of scenes, sentiments, and objects among image–poem pairs, the dataset was evaluated by five English literature major human judges to determine whether the poems are precisely inspired by paired images. After this evaluation, the irrelevant image–poem pairs were dropped, while the rest of the relevant pairs are considered to form the MultiM-Poem dataset. The poems in the MultiM-Poem dataset consist of an average of 7.2 lines per poem and an average of 5.7 words per line. We skipped the image–text pairs for which the corresponding images were unavailable from Web URLs, and the rest of the dataset is used in this paper. From this dataset, we randomly reserved 20% of the image–text pairs as a hold-out split for model evaluation. The rest of 80% of image–text pairs were utilized as train split. Some sample image–poem pairs from the MultiM-Poem dataset are provided in Figure 4. The dataset is publicly available and can be downloaded from [44].

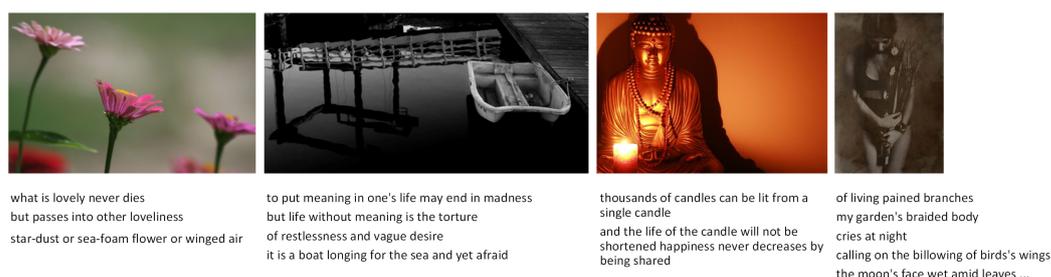


Figure 4. Sample image–poem pairs from MultiM-Poem Dataset.

4.2. CE-CLIP Training Objective

Poets often intend to express topics and sentiments explicitly, which is herein referred to as the artistic conception of poetry. For instance, smooth rivers may be used to represent peace and lamps to represent hope. However, the existing works on image–text matching focus on object relevance, where the matched text is selected based on commonly occurring objects in images and their names in text. Moreover, two poems may have similar keywords but can be arranged to convey entirely different sentiments and topics, and some poems may contain an entirely disjointed set of keywords expressing similar sentiments and topics. Another challenge involves the correct identification of poetic identity in a given image. For instance, multiple objects in an image, such as a sunset, river, trees, and a bird, might represent the topic of loneliness. In this case, existing methods may match with irrelevant

poetry based on object identification, where it was important to rank the relevance of objects based on their poetic value.

The MultiM-Poem dataset used in this work was judged by five human experts based on the sentiments, scenes, and object consistency in terms of poem inspirations of a given image. Considering the task of image–poem matching for visually impaired people’s artwork appreciation, the matching method should be able to consider metaphor and symbolism in poetry and should project sentimental value, in addition to object relevance, into image-based retrieved poetry. The objective of our CE-CLIP is to retrieve top-ranked K poems RP_K based on sentiment, scene, and object consistency from a database of N poems $LP_N = \{p_1, p_2, \dots, p_n\}$ for an input image. The sentiments, scenes, and object consistency in the MultiM-Poem dataset are considered as top-ranked ground-truth among image–poem pairs in the training dataset. Consequently, the objective of the optimization is essentially to maximize the sentiments, scenes, and object consistency among paired images and poems and to minimize it among unpaired images and poems.

4.3. Implementation Details

Same as CLIP, the image encoder of our model is a 12-layer 512-width ViT-B/32 [38] with 12 attention heads, and the text encoder of our model is a 12-layer 512-width transformer with 8 heads [13]. The multi-layer perceptron (MLP) layers used for mapping these encoders outputs onto cross-encoder input are of size 382. The sequence-wise shared attention transformer for the cross-encoder module is also a transformer architecture with 12 layers, it is of 768 hidden size with 12 attention heads. We assign the ground-truth label of “1” for given image–poem pairs from the dataset (positive samples), while the label “0” is assigned when an image is paired with a random poem from the dataset other than the given image–poem pair (negative samples). For each positive sample, we arrange three image–poem combinations of negative samples in our training dataset, thus extending the dataset size for training. We train the model with multiple negatives ranking loss [45] using mean-poolings, the similarity function of cosine similarity, and the scale of 20 and optimize the parameters using an AdamW optimizer. The longer poems are truncated, while shorter poems are padded to match the context vector size of the CLIP text encoder at input. In all the experiments, the batch size is set to 64, and the learning rate is set to 1×10^{-6} .

We use the CLIP’s pre-trained weights to initialize the weights of the image encoder and text encoder in our model. The weights in the cross-encoder module are initialized from a pre-trained MPNet model [18]. The open-source python implementations of these models with pre-trained model weights are available at [46,47], respectively. The pre-training was performed on 215M question–answer text pairs from diverse sources. The pre-trained MPNet model was designed for the semantic search task, which involves a single data modality of text, contrary to the multi-modal image–text data for our task. Therefore, we employ a multi-stage training strategy to train our model to avoid the under- or over-fitting of certain modules, which includes the warm-up and fine-tuning stages. In the warm-up stage, the weights of the image encoder and the text encoder are frozen while we train the cross-encoder module. Afterward, we unfreeze the weights of the image encoder and the text encoder and end-to-end train all the modules together. During the warm-up stage of our model, the weights of the cross-encoder module, including the MLP layers, are updated for a maximum of 100 epochs, while the weights of the CLIP modules are frozen. After the warm-up stage, the checkpoint with the highest validation accuracy is picked, and the weights of both the CLIP and cross-encoder modules are updated for a maximum 200 epochs in the fine-tuning stage.

For comparison of the proposed method at fine-grained poem retrieval task, we conducted experiments by fine-tuning the CLIP pre-trained model [8] with the ViT-B/32 [38] and transformer [13] as image- and text-encoder backbones, respectively. In our zero-shot, few-shot, and fully-supervised experiments, 0, 20, and 100 percent of data from the training set was used for fine-tuning the CLIP pre-trained model. The zero-shot here is essentially a baseline method, reflecting the results of the existing pre-trained CLIP model. The few-shot

learning model is task-oriented on making inferences or predictions based on a limited number of samples rather than the full training dataset. Conventionally, fully supervised learning models are trained or fine-tuned on a training dataset, where the goal of training is to generalize on training data features and be able to recognize them. On the contrary, few-shot learning involves a support set consisting of a small number of labeled samples, and the training involves merely learning to be able to recognize. Similar to an existing work based on the CLIP model [35], we only used a 20% subset of training data for few-shot fine-tuning of the pre-trained CLIP model, denoted by CLIP (Few Shot). These fine-tuned models are evaluated on the hold-out set of the MultiM-poem dataset, where 512 dimension text representations are extracted for all the poems in the hold-out set. Later, normalized pairwise cosine similarities are calculated between the extracted text representations and acquired image representations of the given image.

4.4. Evaluation Metrics

The performance of the fine-grained poetry retrieval task is evaluated in terms of the relevance and ranking of returned matches. Retrieval@K is the percentage of events with ground-truth poems associated with query images included in the top K retrieved poems at the output. Retrieval@K with K values of 5 and 20 are evaluated similar to [35], and, additionally, results for Retrieval@1 are provided, which is the percentage of exact matches by query images from all of the hold-out set poems. Ranking performances are computed by mean retrieval rank and median retrieval rank, which are mean and median rankings of ground-truth poems in the poems' retrievals from all the query images.

5. Results and Discussion

Our model was evaluated on the MultiM-Poem dataset, which is comprised of image–poetry pairs and is the largest publicly available dataset to the best of our knowledge. We compared the obtained performance with the pre-trained CLIP model as zero-shot, the pre-trained CLIP model fine-tuned on 20% of our dataset as few-shot, and the pre-trained CLIP model fine-tuned on the train set of our dataset as a fully supervised CLIP model. The poetry retrieval results were evaluated using the evaluation metrics explained in Section 4.4 and are provided in Tables 1 and 2. The CLIP backbone image and text encoders helped to yield the fine-grained image and poetry features. In our proposed scheme, these feature embeddings capitalize on strong pre-training on 400 million image–text pairs. Considering that the transformer models constructed from scratch are known to be data-hungry, the feature embeddings provide the representations leveraging the contrastive pre-training. However, it is evident that instead of using cosine similarity for matching, the shared-attention-feature learning improves the context awareness among images and fine-grained poems.

Existing works on image–text matching (Section 2.4) are mainly based on frameworks where ROIs (regions of interest) in images and their descriptive words are learned jointly. This ROI-word joint learning essentially breaks down the problem of image–text matching to object-feature detection before computing similarity. These methods, such as [31–33], attend to words in the text with respect to each ROI in the images, which is leveraged by datasets such as Flickr30k [48] and MS-COCO [49], as their image ROIs can be localized through respective bounding boxes and segmentation masks. The development of image–text matching datasets with intermediate annotations of labeled segmentation masks or bounding boxes, in addition to text descriptions of images, is a very expansive task. Therefore, applications of these frameworks on a variety of tasks and datasets are limited. Moreover, these methods rely on object relevance for image–text matching, and they are not suitable for our task as poetic verses do not necessarily describe objects from a given image. For the aforementioned reasons, these methods cannot be evaluated on the MultiM-Poem dataset, due to the unavailability and unsuitability of annotated ROIs. The related works involving images and poetry (Section 2.3) focus on the generation of new poetry and/or constrained Chinese poetry of a particular genre, so they cannot be compared with this

work. The CLIP model exceeds the performance of these image–text matching methods for caption retrieval as per their respective reported results. Therefore, we have demonstrated the effectiveness of the proposed CE-CLIP model in comparison with the CLIP model.

Table 1. Retrieval Results and Comparison of MultiM-Poem Dataset Hold-out Set. (↑: higher is better, ↓: Lower is better).

Method	Data (%)	Retrieval@K (↑)			Retrieval Ranking (↓)	
		K = 1	K = 5	K = 20	Mean	Median
CLIP (Zero Shot)	0	10.9	25.1	40.5	139.9	42
CLIP (Few Shot)	20	12.0	27.7	44.3	116.3	31
CLIP (Fully Supervised)	100	13.0	28.6	45.5	109.8	26
This Work	100	18.4	44.7	65.1	71.4	17

Table 2. Results and Comparison for Image-based Ground-Truth Poem Retrieval out of 100 Poems. (↑: higher is better, ↓: Lower is better).

Method	Data (%)	Retrieval@K (↑)			Retrieval Ranking (↓)	
		K = 1	K = 5	K = 20	Mean	Median
CLIP (Zero Shot)	0	31	51	74	15.5	5
CLIP (Few Shot)	20	34	59	81	12.1	4
CLIP (Fully Supervised)	100	32	63	84	9.3	3
This Work	100	53	78	90	8.2	1

Considering the task and scope of our application and the nature of fine-grained poetry data, the candidate method must account for sentimental correlation among images and poems for a fair comparison (Section 4.2). In practice, we have considered methods including an object detector to extract image keywords to be looked up in poems and image captioning models to generate input image descriptions to be rank poems based on text similarity. In these cases, the state-of-the-art object detectors and image-captioning models did not yield any predictions for about half of the test images pertaining to the nature of images carrying sentimental value. This, considering weak predictions for the case when they yield some, is not sufficient to look up keywords in poems or compute text-similarity for matching poems. It is also complicated to jointly train the combinations of; object detector in the pipeline followed by keyword lookup, and the image-captioning model followed by the text-similarity model, in the absence of such literature and intermediate labels for images. Another candidate pipeline may be based on matching extracted sentiments from both images and poems. However, existing sentiment classification methods with compatible outputs either classify emotions or return one-dimensional outputs, which is not sufficient to encapsulate a wide variety of sentiments for fine-grained image–poem matching.

Table 1 lists the evaluation results of the hold-out set, which is the dataset split unseen by the model during training and fine-tuning. This retrieval task is quite challenging, as each input image must pick the top match(es) from more than 1500 free-form poems. For instance, the random guess number for mean retrieval ranking is around 780. Zero-shot CLIP results demonstrate the CLIP model’s pre-trained features and matching capabilities. The CLIP model performance has minor retrieval@K improvements for few-shot and fully-supervised and indicative improvement for retrieval ranking. This reflects that the fine-tuning of the CLIP model on the training dataset contributes positively towards feature learning. The disparity among the rate of improvement for relevance and ranking metrics may be characterized by the abundance of suitable poems in the dataset for any given image. The improvements in ranking metric performance suggest that further learning of the model is contributing to moving the ground-truth poem upwards in its ranking against other suitable or similar poems. In essence, further learning may have an impact

on bringing the image representations closer to some clusters of similar poems, including ground-truth poems. Likewise, this trend can be observed in Table 2, when each input image must pick a ground-truth poem from 100 poems provided arbitrarily. Many other suitable poems exist, all while indicating that, in addition to ground-truth image–poem pair, pairing and matching images with just one poem each may not be sufficient during dataset construction and automatic retrieval. The retrieval results for both relevance and ranking significantly improve with our proposed method, reflecting the model-learning capability for fine-grained feature recognition. The harmony in the rate of improvement for relevance and ranking metrics performance may be characterized by the cross-encoder module and the shared-attention mechanism across both modalities.

6. Conclusions

To solve multi-modal image-based poetry fine-grained attribute recognition, modeling, and retrieval challenges, we presented a transformer-based novel approach in this work. In the absence of visual artworks and poetry one-to-one or one-to-many pairing public datasets, we have used the multiM-Poem dataset that contains a wide variety of images of sentimental value and poems, not constricted to any specific genre. We consider this the largest publicly available image–poem paired dataset with expert matching. The existing CLIP model can be used directly for image–text matching, without or with optimization for this problem. However, it lacks cross-modality shared attention parameters between visual and textual feature encoding pipelines, which constrains the attention query parameter to look up to the key and the value parameters for modeling the relationship between both modalities. Concurrently though, the CLIP model backbone encoders for images and texts are pre-trained on a private dataset containing 400 M image–text pairs, providing great global feature representations. Our proposed method inherits the benefits of both the CLIP model and cross-encoders and evades their limitations. The proposed approach leverages the global feature representations based on the CLIP model and the fine-grained feature representations and matching through the shared-attention mechanism. We employ a cross-encoder-based method for the shared attention mechanism, which intrinsically combines masked and permuted modeling. The improved results over the CLIP model demonstrate that the proposed approach is capable of better modeling fine-grained features in poems, such as symbolism and metaphor, and discovering the common semantic directivity among images and poems. The proposed deep learning approach may further be extensively explored in future for its zero-shot, few-shot, and fully supervised generalization capabilities on a diverse range of tasks, such as optical character recognition, geo-localization, and action recognition. Further work on this problem can focus on the development of a dataset for visual artwork and poetry matching, followed by user tests for performance evaluation of the multi-sensory matching system. These, together with this work, will provide a seamless way to convey the plausible sentimental portrayal of visual artworks through poems for people with visual impairments.

Author Contributions: Conceptualization, methodology, software, and writing—original draft preparation, M.S.J.; supervision, J.S.; writing—review and editing and project administration, J.S. and J.-D.C.; resources and funding acquisition, J.-D.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program(IITP-2022-2018-0-01798) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The MultiM-Poem dataset can be downloaded at the official website <https://github.com/researchmm/img2poem/tree/master/data> (accessed on 31 March 2022). The

algorithm and methodological details that support the findings of this study have been presented in this manuscript along with original implementation resources.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
- Cho, J.D. A Study of Multi-Sensory Experience and Color Recognition in Visual Arts Appreciation of People with Visual Impairment. *Electronics* **2021**, *10*, 470 [[CrossRef](#)]
- Cho, J.D.; Jeong, J.; Kim, J.H.; Lee, H. Sound Coding Color to Improve Artwork Appreciation by People with Visual Impairments. *Electronics* **2020**, *9*, 1981. [[CrossRef](#)]
- Gilbert, A.N.; Martin, R.; Kemp, S.E. Cross-modal correspondence between vision and olfaction: The color of smells. *Am. J. Psychol.* **1996**, *109*, 335–351. [[CrossRef](#)] [[PubMed](#)]
- Iranzo Bartolomé, J.; Cho, J.D.; Cavazos Quero, L.; Jo, S.; Cho, G. Thermal Interaction for Improving Tactile Artwork Depth and Color-Depth Appreciation for Visually Impaired People. *Electronics* **2020**, *9*, 1939. [[CrossRef](#)]
- Lawrence, M.A.; Kitada, R.; Klatzky, R.L.; Lederman, S.J. Haptic roughness perception of linear gratings via bare finger or rigid probe. *Perception* **2007**, *36*, 547–557. [[CrossRef](#)]
- Cho, J.D.; Lee, Y. ColorPoetry: Multi-Sensory Experience of Color with Poetry in Visual Arts Appreciation of Persons with Visual Impairment. *Electronics* **2021**, *10*, 1064. [[CrossRef](#)]
- Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, PMLR, Online, 18–24 Jul 2021; pp. 8748–8763.
- Maric, Y.; Jacquot, M. Contribution to understanding odour–colour associations. *Food Qual. Prefer.* **2013**, *27*, 191–195. [[CrossRef](#)]
- Slobodenyuk, N.; Jraissati, Y.; Kanso, A.; Ghanem, L.; Elhajj, I. Cross-modal associations between color and haptics *Atten. Percept. Psychophys.* **2015**, *77*, 1379–1395. [[CrossRef](#)]
- Jabbar, M.S.; Lee, C.H.; Cho, J.D. ColorWatch: Color Perceptual Spatial Tactile Interface for People with Visual Impairments. *Electronics* **2021**, *10*, 596. [[CrossRef](#)]
- Kim, Y.; Jeong, H.; Cho, J. D.; Shin, J. Construction of a soundscape-based media art exhibition to improve user appreciation experience by using deep neural networks. *Electronics* **2021**, *10*, 1170. [[CrossRef](#)]
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [[CrossRef](#)]
- Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 Nov 2020; pp. 38–45.
- Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. *arXiv* **2021**, arXiv:2108.07258.
- Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
- Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 2019.
- Song, K.; Tan, X.; Qin, T.; Lu, J.; Liu, T.Y. MpNet: Masked and permuted pre-training for language understanding. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 16857–16867.
- Medsker, L.; Jain, L. C. *Recurrent Neural Networks: Design and Applications*; CRC Press: Boca Raton, FL, USA, 1999.
- Tschannen, M.; Bachem, O.; Lucic, M. Recent advances in autoencoder-based representation learning. *arXiv* **2018**, arXiv:1812.05069.
- Liu, D.; Guo, Q.; Li, W.; Lv, J. A multi-modal chinese poetry generation model. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8.
- Liu, Y.; Liu, D.; Lv, J.; Sang, Y. Generating Chinese Poetry from Images via Concrete and Abstract Information. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
- Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
- Wu, L.; Xu, M.; Qian, S.; Cui, J. Image to modern chinese poetry creation via a constrained topic-aware model. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2020**, *16*, 1–21. [[CrossRef](#)]
- Liu, L.; Wan, X.; Guo, Z. Images2poem: Generating chinese poetry from image streams. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 1967–1975.
- Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
- Liu, B.; Fu, J.; Kato, M.P.; Yoshikawa, M. Beyond narrative description: Generating poetry from images by multi-adversarial training. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 783–791.

28. Wu, C.; Wang, J.; Yuan, S.; Wang, L.; Zhang, W. Generate classical Chinese poems with theme-style from images. *Pattern Recognit. Lett.* **2021**, *149*, 75–82. [[CrossRef](#)]
29. Liu, Y.; Liu, D.; Lv, J. Deep poetry: A chinese classical poetry generation system. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 13626–13627. [[CrossRef](#)]
30. Zhang, D.; Ni, B.; Zhi, Q.; Plummer, T.; Li, Q.; Zheng, H.; Zeng, Q.; Zhang, Y.; Wang, D. Through the eyes of a poet: classical poetry recommendation with visual input on social media. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Vancouver, BC, Canada, 27–30 August 2019; pp. 333–340.
31. Nam, H.; Ha, J.W.; Kim, J. Dual attention networks for multimodal reasoning and matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 299–307.
32. Lee, K.H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked cross attention for image–text matching. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 201–216.
33. Li, K.; Zhang, Y.; Li, K.; Li, Y.; Fu, Y. Visual semantic reasoning for image–text matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 4654–4662.
34. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [[CrossRef](#)] [[PubMed](#)]
35. Conde, M.V.; Turgutlu, K. CLIP-Art: Contrastive Pre-Training for Fine-Grained Art Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3956–3960.
36. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
39. Krause, J.; Stark, M.; Deng, J.; Fei-Fei, L. 3d object representations for fine-grained categorization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Sydney, Australia, 2–8 December 2013; pp. 554–561.
40. Bossard, L.; Guillaumin, M.; Gool, L. V. Food-101—mining discriminative components with random forests. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 446–461.
41. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
42. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
43. Liu, H.; Xu, S.; Fu, J.; Liu, Y.; Xie, N.; Wang, C.C.; Wang, B.; Sun, Y. CMA-CLIP: Cross-Modality Attention CLIP for image–text Classification. *arXiv* **2021**, arXiv:2112.03562.
44. Researchmm/img2poem: [MM’18] Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training /Data. Available online: <https://github.com/researchmm/img2poem/tree/master/data> (accessed on 31 March 2022).
45. Henderson, M.; Al-Rfou, R.; Strophe, B.; Sung, Y.H.; Lukács, L.; Guo, R.; Kumar, S.; Miklos, B.; Kurzweil, R. Efficient natural language response suggestion for smart reply. *arXiv* **2017**, arXiv:1705.00652.
46. Openai/CLIP: Contrastive Language-Image Pretraining. Available online: <https://github.com/openai/CLIP> (accessed on 31 March 2022).
47. UKPLab/Sentence-Transformers: Multilingual Sentence & Image Embeddings with BERT. Available online: https://github.com/UKPLab/sentence-transformers/blob/master/docs/pretrained_models.md (accessed on 31 March 2022).
48. Plummer, B.A.; Wang, L.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 2641–2649.
49. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 740–755.