Multi-modal Label Retrieval for the Visual Arts: The Case of Iconclass

Nikolay Banar[1,2], Walter Daelemans[1] and Mike Kestemont[1,2]
[1]*Computational Linguistics and Psycholinguistics Research Center, University of Antwerp, Belgium*
[2]*Antwerp Centre for Digital Humanities and Literary Criticism, University of Antwerp, Belgium*

Keywords: Iconclass, Information Retrieval, Multi-modal Matching, Cross-lingual Matching, Multi-lingual Matching.

Abstract: Iconclass is an iconographic classification system from the domain of cultural heritage which is used to annotate subjects represented in the visual arts. In this work, we investigate the feasibility of automatically assigning Iconclass codes to visual artworks using a cross-modal retrieval set-up. We explore the text and image branches of the cross-modal network. In addition, we describe a multi-modal architecture that can jointly capitalize on multiple feature sources: textual features, coming from the titles for these artworks (in multiple languages) and visual features, extracted from photographic reproductions of the artworks. We utilize Iconclass definitions in English as matching labels. We evaluate our approach on a publicly available dataset of artworks (containing English and Dutch titles). Our results demonstrate that, in isolation, textual features strongly outperform visual features, although visual features can still offer a useful complement to purely linguistic features. Moreover, we show the cross-lingual (Dutch-English) strategy to be on par with the monolingual approach (English-English), which opens important perspectives for applications of this approach beyond resource-rich languages.

1 INTRODUCTION

Iconclass (Vellekoop et al., 1973; Brandhorst, 2019) is a well-known iconographic classification system which is used to describe and retrieve content in artworks. The ontology is adopted across various institutions in the GLAM sector (Galleries, Libraries, Archives and Museums). Iconclass offers a hierarchy of unique codes, associated with keywords and definitions, to encode the presence of objects, people, events and ideas depicted in visual artworks, such as paintings. Assigning Iconclass codes (or other class labels coming from other iconographic thesauri) is a complex interpretive task that is typically carried out by highly-trained subject experts. Assigning an Iconclass code to an artwork is an especially challenging task because of the large number of available labels. Hence, the annotation process is time-consuming and requires the (expensive) intervention of skilled experts.

Recent advances in deep learning (LeCun et al., 2015; Schmidhuber, 2015) increasingly find real-world applications in the cultural heritage domain (Fiorucci et al., 2020). Iconclass, for instance, was recently used to improve the quality of neural machine translation, specifically for artwork titles (Banar et al., 2020). In spite of the growing availability of relevant datasets, however, the automatic assignment of Iconclass codes to artworks has yet not attracted the scholarly attention which this challenging task deserves. A related study into the automatic classification of artworks into (just) 10 Iconclass categories (Milani and Fraternali, 2020) recently demonstrated the considerable difficulty of this task.

We aim to move beyond the state of the art in this area through exploiting the latest advances in information retrieval in order to reliably match artworks with suitable interpretive metadata, such as Iconclass codes. Importantly, we propose a multi-modal approach that is able to jointly capitalize on various data sources, including (multilingual) textual information as well as visual characteristics available for these artworks. First, we describe the extraction of the linguistic features from Dutch and English artwork titles, as well as the visual feature extraction from the artwork images. We start by investigating the feasibility of cross-modal (image-to-text) matching using the image and text branches of a recently proposed architecture, called "Self-Attention Embeddings" (SAEM, (Wu et al., 2019a)). Subsequently, we compare this matching strategy to a text-to-text mapping, only using the text branch of SAEM. In these experiments, we additionally report on the feasibility of cross-language matching. Finally, we propose a

Banar, N., Daelemans, W. and Kestemont, M.
Multi-modal Label Retrieval for the Visual Arts: The Case of Iconclass.
DOI: 10.5220/0010390606220629
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 1*, pages 622-629
ISBN: 978-989-758-484-8

simple extension of the model that is able to simultaneously exploit multi-lingual metadata and visual features to match Iconclass codes to a work of art.

The structure of this paper is as follows. We first review the related work on cross-modal matching in Section 2. Then, we describe the proposed methods in more detail in Section 3. We further describe the representative cultural heritage dataset used in our experiments and discuss the experimental settings adopted in Section 4. The quantitative results of this work, together with a more interpretive discussion, are offered in Section 5. Finally, we summarize our contributions in Section 6 and propose worthwhile directions for future research.

2 RELATED WORK

In this section, we review related methods from the recent literature on cross-modal matching. Generally, the published approaches can be classified into 4 categories (Chen et al., 2020): 1) pairwise learning embeddings; 2) adversarial learning; 3) attribute learning; 4) interaction learning.

Pairwise learning methods focus on designing a cross-modal loss function to map image and text embeddings into one common space. The loss function is specifically designed so as to reduce the distance between positive pairs, while increasing the distance between negative ones. Zhang and Lu (2018) proposed a Cross-Modal Projection Matching loss, which uses the Kullback-Leibler divergence between image-to-text matching probability and normalized ground-truth probability. Jian et al. (2019) proposed a similar method, adopting a softmax cross-entropy loss and a bi-triplet loss.

Adversarial learning methods use Generative Adversarial Nets (GANs, (Goodfellow et al., 2014)) for cross-modal matching. GANs were seminally applied in this context by Wang et al. (2017a). Their method is based on mini-max strategy applied to the generator and discriminator in GANs. Sarafianos et al. (2019) have extended this framework to obtain modality-invariant representations. Liu et al. (2019), finally, proposed a deep adversarial graph attention convolution network which exploits textual and visual scene graphs.

Attribute learning methods exploit high-level semantic attributes instead of the basic image features and text features. The Attribute-Guided Network (Ji et al., 2019) utilizes zero-shot learning and hashing retrieval for this purpose. With this approach, the attribute vectors are mapped onto hash codes. The aim of this mapping is to automatically obtain clusters, across different modalities, in a common space.

Interaction learning methods, the fourth and last category, aim to transfer information between the text and image branches in a model, before mapping them jointly into a common space. The Multitask learning approach for Cross-Modal Image-Text Retrieval (Luo et al., 2019) uses a relation-enhanced cross-modal auto-encoder. The cross-modal auto-encoder correlates the hidden representations of two uni-modal auto-encoders, before mapping image and textual features into a common latent space. Through stacked cross-attention, Lee et al. (2018) build attended vectors for each image region from the salient parts of the sentence. Then, the similarity is calculated between each of the attended vectors and the corresponding image region; subsequently, the dimensionality of the similarity matrix is reduced by spatial pooling. The SAEM framework (Wu et al., 2019a) adopts a self-attention mechanism (Vaswani et al., 2017) to process image regions. In this work, we resort to the image and text branches of SAEM, as it has achieved excellent results in cross-modal matching (Chen et al., 2020). Moreover, a reference implementation of the model is available online.[1]

3 METHODS

In this section, we describe the SAEM architecture in more detail. The framework generally consists of two branches (see Figure 1): we present the image branch in Section 3.1 and the text branch in Section 3.2. In Section 3.3, finally, we propose a simple extension of the architecture that allows to consider multiple data sources simultaneously in the matching task at hand.

3.1 Image Branch

This branch is responsible for extracting visual features from an arbitrary image. It implements a bottom-up-attention mechanism (Anderson et al., 2018) to extract salient regions from the image and a self-attention layer (Vaswani et al., 2017) to encode these regions. It deserves emphasis that the bottom-up-attention network is not fine-tuned in the training process. The bottom-up-attention mechanism brings specific advantages over the simple extraction of features from the last pooling layer of a convolutional neural network (CNN). Such CNN feature extractors divide an image into equal-size spatial blocks and preserve spatial information about the image in the process. This process will ignore the semantic structure of the image but might devote too much of its

[1] https://github.com/yiling2018/saem

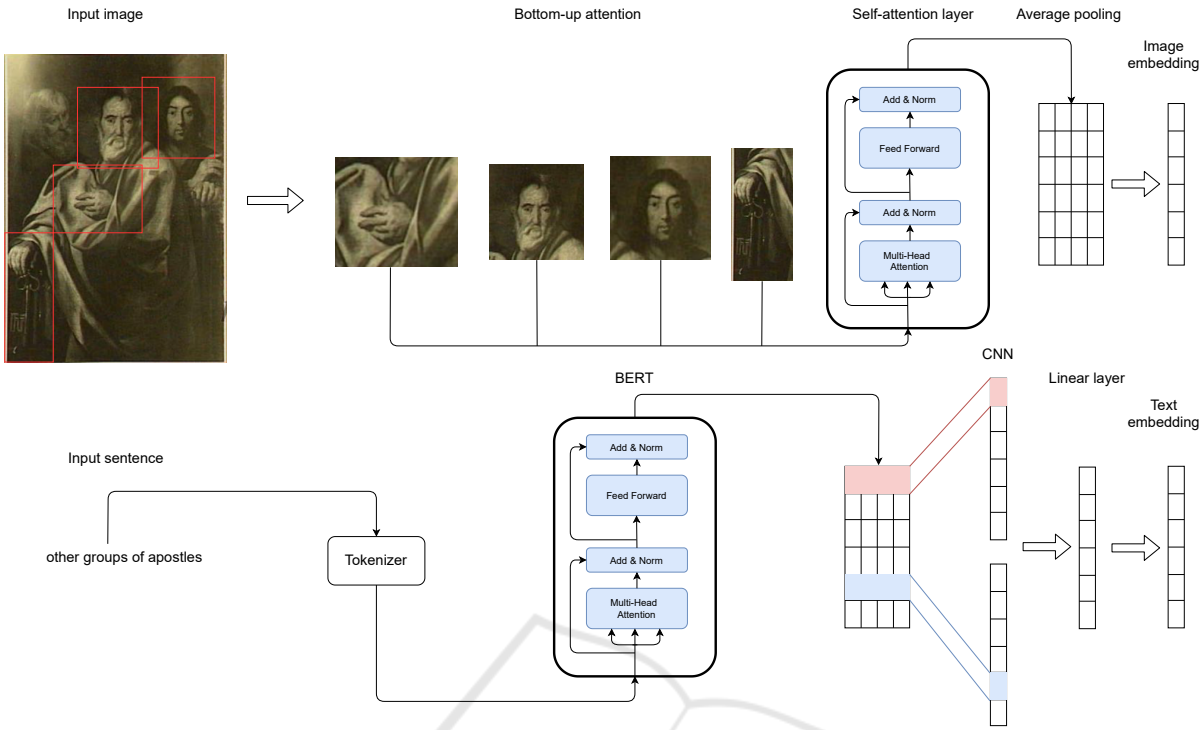


Figure 1: The scheme of the SAEM framework. The upper branch processes visual information and the lower branch is responsible for textual information.

representational capacity to image fragments containing redundant information. The bottom-up-attention mechanism, however, will boost salient image regions of the different sizes, which allows to emphasize the semantic information in the image and avoid unnecessary computations.

The bottom-up-attention mechanism employs the Faster R-CNN model (Ren et al., 2015) with a ResNet-101 (He et al., 2016), pretrained on Visual Genomes (Krishna et al., 2017). Faster R-CNN is a mature object detection algorithm that involves a series of steps. First, the CNN extracts features from an image to build a feature map. Next, an independent Region Proposal Network uses the feature map to identify regions-of-interest (RoI) proposals of different sizes. These regions are then resized into equal-size vectors by the RoI pooling layer. The resized regions are finally used to predict the offset values for bounding boxes and to classify objects.

The current pipeline uses the resized regions processed by the RoI pooling layer. Next on, these regions are fed through a position-wise fully connected layer to combine them into a single feature matrix. Only at this stage, the self-attention layer is applied to encode the relationships that exist between these regions. This step is crucial, as the extracted regions do not have a fixed order. The self-attention layer is able to access all regions simultaneously, which enables it to extract useful information from the regions, despite their lack of a fixed order. Finally, the image embedding is build by average pooling over the feature matrix, before it eventually gets L2-normalized.

3.2 Text Branch

The text branch of the model uses Bidirectional Encoder Representations from Transformers (BERT, (Devlin et al., 2019)) with the WordPiece tokenizer to encode the textual information. Importantly, and analogously to the image branch, the BERT weights are also not fine-tuned in the training process. The well-known BERT architecture has been pretrained taking into account bi-directional context. Thus, it provides context-aware sentence embeddings, which is a major advantage over context-insensitive (word) embedding approaches, such a Word2Vec (Mikolov et al., 2013) or Glove (Pennington et al., 2014). BERT has been pretrained on two different tasks. In the first task, BERT has been pretrained to predict randomly masked input tokens in sentences. In the second task, BERT was additionally pretrained on the task of next-sentence prediction.

The embeddings obtained from BERT are processed by one-dimensional convolutions (for uni-grams, bi-grams and tri-grams), followed by max-

pooling to capture the local context. Further, the features are concatenated into a single vector and fed through a fully connected layer to obtain the final text embedding, which is L2-normalized in the end.

3.3 Multi-modal Branch

We propose a simple extension to the original SAEM approach through which we are able to exploit information coming from multiple sources. In our work, we have available both textual and visual sources for the matching task. We process the available textual information with the text branch and process the available visual information with image branch. Hence, we obtain multi-modal embeddings, which we concatenate into a single vector. We resize the resulting embedding by a fully connected layer, to re-obtain the original size of the target embeddings.

4 EXPERIMENTAL SETTINGS

In this section, we describe the datasets and discuss the experimental settings which we adopted.

4.1 Datasets and Preprocessing

4.1.1 Iconclass

Iconclass is an iconographic classification system used by cultural heritage intuitions to describe and retrieve content in the visual arts. An Iconclass code (see Figure 2) is a unique identifier assigned to an iconographic subject represented in an artwork. Iconclass includes 28,000 hierarchically ordered definitions (codes) and 14,000 keywords in multiple languages. As matching targets, we use the English definitions of the Iconclass codes in this work, as they are well presented in Iconclass compared to, for example, the Dutch definitions. Iconclass is divided into 10 main categories that are represented with a digit from 0 to 9: (0) abstract art; (1-5) general topics; (6) history; (7) Bible; (8) literature; (9) classical mythology and ancient history. Furthermore, an Iconclass code can be extended by the three options presented in Table 1. We have used the Iconclass Python package to obtain the Iconclass definitions associated with each individual code.[2] In our experiments, we have considered Iconclass codes with a depth of 5 and obtained 10,418 codes in this manner, which could be used as labels for the present matching task. Hence, we do not exploit the hierarchical structure of Iconclass codes in this work.

4.1.2 Dataset

To work with a representative collection of material from the domain of cultural heritage, we extracted the dataset from the database of the Netherlands Institute for Art History.[3] The dataset consists of 26,725 objects and their corresponding (manually assigned) Iconclass codes. Each object represents a unique visual artwork and includes the following metadata triplets: 1) Dutch-language artwork titles; 2) English-language artwork titles; 3) a photographic reproduction of the artwork under scrutiny. We set aside a random selection of 2,000 objects to be used as the development set and included another 2,000 objects for the test set. We randomly selected one Iconclass code per object in the development and test sets, while the training set has 1.41 ± 1.40 Iconclass codes per object. In our experiments, we use all unique combinations of the triplets to train and test the models (see Table 2).

4.2 Training and Inference Details

The SAEM framework is mainly implemented in PyTorch (Paszke et al., 2019) but it is dependent on a visual feature extractor implemented in Caffe (Jia et al., 2014). Instead of the cased English BERT, we use the uncased multi-lingual version of BERT to maximally benefit from our multi-lingual metadata. Throughout our work, we have maximally adhered to the default parametrization of the SAEM framework, with the obvious exception of our own extensions to the model. The models were trained by minimizing a combination of a bi-directional triplet loss (Wu et al., 2019b) and a bi-directional angular loss (Wang et al., 2017b), with hard negative mining using the Adam optimizer (Kingma and Ba, 2014) and a batch size of 64 sentences. The initial learning rate was set to 0.0001 but it was decayed by a factor of 0.1 after every 10 epochs. The models were trained for 30 epochs on a single GeForce GTX 1080 Ti with 11 GB RAM. The evaluation is conducted using the standard metric Recall@K (for K=1, 5, 10), which represents the portion of relevant items found in the top-K retrieved labels. Below, we report the performance of the models for (image, text or multi-modal) query versus the Iconclass definitions. For each experiment, we select the best performing model which had the opti-

[2]https://labs.brill.com/ictestset/

[3]https://rkd.nl/en/explore/images

Table 1: Extension of Iconclass codes: (1) a letter or digit increases specificity; (2) bracketed text adds the name of a specific entity; (3) bracketed text with a plus-sign introduces an additional 'shade of meaning'.

N	**Extension**	**Definition**	**Keywords**
1	73E8	Joseph's death and coronation	Joseph (St.), death
	73E8**1**	Joseph on his deathbed; Christ and Mary present	Mary (Virgin), deathbed
2	22C4	colours, pigments, and paints	colour, paint, pigment
	22C4**(GOLD)**	colours, pigments, and paints: gold	gold
3	31D111	infant, baby the ages of man	baby
	31D111**(+89)**	infant, baby the ages of man (+ nude human being)	Akt, baby, nackt, nu, nude, nudo

(a) 71H611 (b) 11I424 (c) 73A624

Figure 2: Examples of images assigned Iconclass codes (Posthumus, 2020) with the following definitions: (a) 'David communicating with God; David praying (in general)'; (b) 'angel (possibly with book) symbol of St. Matthew'; (c) 'Mary saluting Elisabeth, who kneels before her'.

mal (summed) performance of the metrics presented above on the validation set.

5 RESULTS AND DISCUSSION

In this section, we present and discuss our experimental results. We present our results in three different sections. First, we discuss the cross-modal (image-text) matching and compare it to one-modal (text-text) matching in Section 5.1. Secondly, we present the results for the cross-lingual matching in Section 5.2. The results for the multi-modal approach follow in Section 5.3.

5.1 Cross-modal Matching

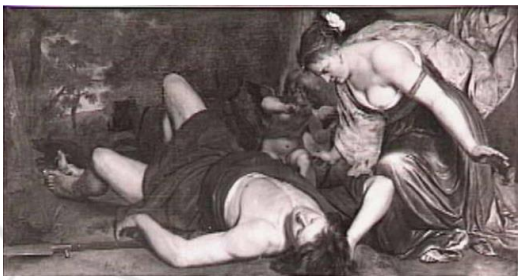
As shown in Table 2, model (a) for cross-modal matching demonstrates low results in comparison to the models (b, c) utilizing textual features for matching. Therefore, we conclude that the cross-modal matching does not deliver satisfactory results in the current setting. However, we should emphasize that this alone does not prove that cross-modal matching would not be feasible for this task. The bottom-up-attention mechanism used to extract the visual features is pretrained on the Visual Genomes dataset which obviously belongs to an entirely another domain. In computer vision, it is a well-known problem of domain mismatch. The common remedy to overcome this issue is transfer learning (Ribani and Marengoni, 2019), namely, fine-tuning weights of a neural network. However, the current framework is

Table 2: Results of the experiments with different matching sources.

	Source			Metrics			
	Image	EN Title	NL Title	Recall@1	Recall@5	Recall@10	Average
a	✓			13.10	19.60	23.20	18.63
b		✓		62.80	77.30	80.75	73.62
c			✓	66.45	77.05	80.55	74.68
d	✓	✓		67.85	79.20	82.30	76.45
e	✓		✓	66.60	78.80	81.80	75.73
f		✓	✓	68.95	80.30	82.90	77.38
g	✓	✓	✓	**70.05**	**80.35**	**83.10**	**77.83**

Table 3: Example of Iconclass code matching from the best performing model (g). The ground-truth label is highlighted in bold. The title of the artwork is 'Venus mourning the dead Adonis' ('Venus beweent de dode Adonis').

N	code	definition
1	92C42	love-affairs of Venus
2	92C49	**offspring, companion(s), train etc. of Venus**
3	92C46	suffering, misfortune of Venus
4	92C41	early life, prime youth of Venus
5	92C48	attributes of Venus
6	92C47	specific aspects, allegorical aspects of Venus [...]
7	92C44	aggressive, unfriendly activities [...] of Venus
8	24C19	Venus (planet)
9	92K22	Dione, mother of Venus
10	24C34	Venus representing copper



not suitable for fine-tuning as it is spread among PyTorch and Caffe. These parts should be merged in order to achieve this goal.

5.2 Cross-lingual Matching

From Table 2, we can see that the model (c) trained on the Dutch titles performs on par with the corresponding model (b) trained on the English titles. The model (c) outperforms the corresponding model (b) in Recall@1 by a large margin. However, the model (b) is moderately better in Recall@5 and Recall@10. This result demonstrates that BERT provides high quality multi-lingual embeddings for our task. Hence, there is no need to translate metadata to English to achieve strong results in the task of Iconclass codes matching. We conclude that the textual features are language-independent and extremely useful in our task.

5.3 Multi-modal Matching

From Table 2, we can observe that all models in the multi-modal scenario (d, e, f, g) outperform all models in the cross-modal (a) and one-modal (b, c) scenarios. Surprisingly, the visual features in the models (d, e) help to improve performance compared to the corresponding models (b, c). This result may be explained by the hypothesis that the visual features might overall have a lower quality, but that the information which they offer is orthogonal to the textual features, and thus a worthwhile addition to the model. The best performer in the scenario with 2 sources is the model (f) where the featured are extracted from both the Dutch and English titles. And finally, the best performing model overall is the model (g) that exploits all available information. Hence, we conclude that the multi-modal approaches (d, e, f, g) outperform cross-modal and one-modal methods and the increased amount of sources improves the quality of matching.

As can be seen from Table 3, Iconclass codes matching remains a challenging task due to the hierarchical structure and a high number of similar labels (codes). Not only, the number of similar labels grows significantly with the depth of Iconclass labels, but allowing a greater depth will also increase the number of very similar labels. In the example, the first seven labels have the same upper code (92C4) and are highly similar which makes the matching extremely difficult. In order to learn such subtle shades of meaning, the models would probably need even more properly annotated data, which is challenging in the cultural heritage domain.

6 CONCLUSION AND FUTURE WORK

In this paper, we investigated different strategies for matching (metadata about) art objects with suitable Iconclass codes. We additionally proposed a simple method that utilizes multiple sources, through a linear mapping of the source embeddings. We utilized textual and visual features extracted from English and Dutch titles and artwork images, respectively. The experiments demonstrate that the cross-modal (image-text) matching using the visual features are not promising compared to the uni-modal (text-text) matching using purely textual features. We show that the cross-lingual matching using the Dutch-language artwork titles works as good as the matching that uses the English-language artwork titles. This finding will be meaningful to practitioners in the field, because it suggests that GLAM institutions around the world, including thus operating in a more resource-scarce context to use their metadata in local languages to match Iconclass codes without translating them first to English. And finally, the proposed method that uses all available information is the best performer. In this case, the visual features help to boost the performance.

The current pipeline has several disadvantages. First, the model uses the BERT and the bottom-up-attention to extract features without actual fine-tuning. It may explain low results for cross-modal matching due to the unsuitable feature representation. Secondly, some parts of the pipeline are implemented in different frameworks which makes model-wide fine-tuning difficult. Thirdly, the current dataset is comparatively small as we had only 22,725 objects in the training set for 10,418 possible labels. A larger dataset certainly would be useful. In future work, we would like to reimplement the entire pipeline in PyTorch in order to fine-tune the full model. In addition, exploiting the hierarchical structure of Iconclass codes remains an important desideratum.

REFERENCES

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Banar, N., Daelemans, W., and Kestemont, M. (2020). Neural machine translation of artwork titles using iconclass codes. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 42–51.

Brandhorst, H. (2019). A word is worth a thousand pictures: Why the use of iconclass will make artificial intelligence smarter. https://labs.brill.com/ictestset/ICONCLASS_and_AI.pdf.

Chen, J., Zhang, L., Bai, C., and Kpalma, K. (2020). Review of recent deep learning methods for image-text retrieval. In *IEEE 3rd International Conference on Multimedia Information Processing and Retrieval*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A., and James, S. (2020). Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133:102–108.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ji, Z., Sun, Y., Yu, Y., Pang, Y., and Han, J. (2019). Attribute-guided network for cross-modal zero-shot hashing. *IEEE transactions on neural networks and learning systems*, 31(1):321–330.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678.

Jian, Y., Xiao, J., Cao, Y., Khan, A., and Zhu, J. (2019). Deep pairwise ranking with multi-label information for cross-modal retrieval. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1810–1815. IEEE.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, K.-H., Chen, X., Hua, G., Hu, H., and He, X. (2018). Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216.

Liu, J., Zha, Z.-J., Hong, R., Wang, M., and Zhang, Y. (2019). Deep adversarial graph attention convolution network for text-based person search. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 665–673.

Luo, J., Shen, Y., Ao, X., Zhao, Z., and Yang, M. (2019). Cross-modal image-text retrieval with multitask learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2309–2312.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Milani, F. and Fraternali, P. (2020). A data set and a convolutional model for iconography classification in paintings. *arXiv preprint arXiv:2010.11697*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Posthumus, E. (2020). Brill iconclass ai test set. https://labs.brill.com/ictestset/.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Ribani, R. and Marengoni, M. (2019). A survey of transfer learning for convolutional neural networks. In *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, pages 47–57. IEEE.

Sarafianos, N., Xu, X., and Kakadiaris, I. A. (2019). Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5814–5824.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vellekoop, G., Tholen, E., and Couprie, L. D. (1973). *Iconclass : an iconographic classification system*. North-Holland Pub. Co., Amsterdam.

Wang, B., Yang, Y., Xu, X., Hanjalic, A., and Shen, H. T. (2017a). Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 154–162.

Wang, J., Zhou, F., Wen, S., Liu, X., and Lin, Y. (2017b). Deep metric learning with angular loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2593–2601.

Wu, Y., Wang, S., Song, G., and Huang, Q. (2019a). Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2088–2096.

Wu, Y., Wang, S., Song, G., and Huang, Q. (2019b). Online asymmetric metric learning with multi-layer similarity aggregation for cross-modal retrieval. *IEEE Transactions on Image Processing*, 28(9):4299–4312.

Zhang, Y. and Lu, H. (2018). Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 686–701.