Multi Stage Common Vector Space for Multimodal Embeddings

Sabarish Gopalakrishnan
Rochester Institute of Technology
sxg8458@rit.edu

Premkumar Udaiyar
Rochester Institute of Technology
pxu4114@rit.edu

Shagan Sah
Rochester Institute of Technology
sxs4337@rit.edu

Raymond Ptucha
Rochester Institute of Technology
rwpeec@rit.edu

***Abstract*—Deep learning frameworks have proven to be very effective at tasks like classification, segmentation, detection, and translation. Before being processed by a deep learning model, objects are first encoded into a suitable vector representation. For example, images are typically encoded using convolutional neural networks whereas texts typically use recurrent neural networks. Similarly, other modalities of data like 3D point clouds, audio signals, and videos can be transformed into vectors using appropriate encoders. Although deep learning architectures do a good job of learning these vector representations in isolation, learning a single common representation across multiple modalities is a challenging task. In this work, we develop a Multi Stage Common Vector Space (M-CVS) that is suitable for encoding multiple modalities. The M-CVS is an efficient low-dimensional vector representation in which the contextual similarity of data is preserved across all modalities through the use of contrastive loss functions. Our vector space can perform tasks like multimodal retrieval, searching and generation, where for example, images can be retrieved from text or audio input. The addition of a new modality would generally mean resetting and training the entire network. However, we introduce a stage-wise learning technique where each modality is compared to a reference modality before being projected to the M-CVS. Our method ensures that a new modality can be mapped into the M-CVS without changing existing encodings, allowing the extension to any number of modalities. We build and evaluate M-CVS on the XMedia and XMedianet multimodal dataset. Extensive ablation experiments using images, text, audio, video, and 3D point cloud modalities demonstrate the complexity vs. accuracy tradeoff under a wide variety of real-world use cases.**

***Index Terms*—retrieval, indexing, multimedia, multimodal, deep learning.**

I. Introduction

Neural network architectures have shown tremendous capabilities in tasks like image classification, image segmentation, language translation and object detection. The flow of knowledge from one data modality to another is a challenging task. In this work, we propose a Multi Stage Common Vector Space (M-CVS) that takes in data belonging to multiple modalities and projects it into a common embedding space. In the M-CVS, similar objects and concepts are clustered together whereas dissimilar objects and concepts are well separated from each other, irrespective of their modality. For instance, the representation of the images, text, audio, video and 3D point cloud of a giraffe will lie close to each other and far away from the representations of cars. Figure 1 shows the overall architecture of the M-CVS. Each of the modalities is mapped into the M-CVS using separate encoders. The fully connected layers help bring the different modalities into the M-CVS.

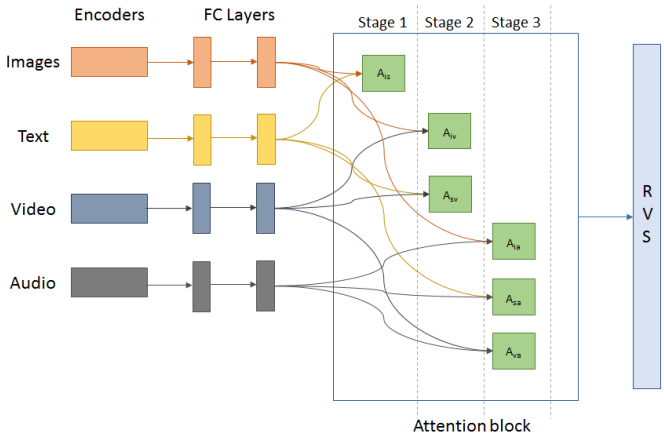

Fig. 1: The M-CVS model. The input modalities pass through their own encoding networks and then through fully connected layers before going in to the attention layer. The output of the attention layer is the M-CVS embedding of the input data.

The task of cross modal retrieval has been tackled in [4], [12] with different loss functions and architectures. Most cross modal tasks restrict themselves to two modalities. We are introducing a new training technique which enables adding of new modalities without changing the existing trained model.

In the M-CVS, a new modality can be added without changing the existing architecture which significantly reduces the time taken to train the model. The main contributions of

978-1-7281-4732-1/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: Carleton University. Downloaded on September 21,2020 at 02:02:05 UTC from IEEE Xplore. Restrictions apply.

this work are as follows:

- Develop a low dimensional vector space that accommodates data originating from multiple modalities.
- Demonstrate the ability of the M-CVS to seamlessly extend to more than two modalities without affecting existing inference.
- Perform cross modal retrieval on different multimodal datasets.
- Implement stage wise training of neural networks to see effect on performance of retrieval and inference time.
- Compare different training strategies and compare the pros and cons of each.

II. Related Work

Deep learning has been effective in tasks like image classification [7], [9], [10], [25], semantic segmentation [2], [11], object detection [19], [20], and language translation [6]. The interaction between text and vision in particular has made some great strides. For example, Sah *et al.* [22] trained a network that summarized several hour-long videos to short paragraphs. Using multiple modalities of data to perform retrieval is referred to as cross modal retrieval. Most work in the multimodal space restricts itself to only two modalities. Zhang *et al.* [32] developed a text encoding framework that generated a vector representation of a text while [23] shows how multiple captions can be used to generate images using common vector representations.

Deng *et al.* [3] performed cross modal hashing using a triplet based hashing network and Wu *et al.* [27] showed the use of adversarial training by using cycle consistency loss function. The task of real-value based retrieval can be modeled in different ways. Qi *et al.* [18] used the triplet loss function to form triplets (anchors, positive, negative) and train a model whose objective function ensured that the encoded vector representation of anchors and their corresponding positive samples lie close in the latent space, while the encoded vector representation anchors and negative samples are pushed far away. More recent works like Xu *et al.* [29] used adversarial loss functions and built an architecture that converted a 4096-dimensional vector representation to a 200 dimension vector using fully connected layers. Zhen *et al.* [34] used a combination of a linear classifier, modality invariance loss, and intermodal loss to perform retrieval. Peng *et al.* [15] created an architecture that was able to retrieve images from their outline sketches and vice-versa. Works in [8], [17] show the use of retrieval of data from one modality to the other at the granular level where the input query is able to retrieve its corresponding ground-truth sample of the other modality.

III. Proposed Methodology

A. Network Architecture

The architecture is shown in Figure 1. The model has five different modalities of data that are mapped onto the M-CVS using fully connected neural network layers. Each of the modalities have their own encoding networks that extract the individual vector representations of the data. The encoders then pass through two fully connected layers into the M-CVS through the attention layer. We use the $tanh$ activation function in between the fully connected layers. Using the fully connected layers, we bring all the data into a 512 dimensional vector representation. The vectors then pass through the aligned attention layer. The output of the attention layer is the final M-CVS representation of the data.

As new modalities are introduced, they are compared with all existing modalities, and their weights are update such that when mapping into M-CVS, we should not be able to tell which modality the data came from. In this comparison, the weights of all the modalities already present in the M-CVS are frozen and only the weights of the new modality are updated.

We compare our implementation of the M-CVS with the CVS model [21], [23], a non-stage wise implementation of the M-CVS. Conceptually, both the CVS and the M-CVS are similar but the way in which they are trained is quite different. In the CVS, we need $\binom{n}{2}$ attention blocks between each pair of modalities before the data can be projected in to the common embedding space. Another limitation of the CVS is that the addition of a new modality into the CVS alters the existing inference of all the other modalities. That is to say that the CVS needs all the data before being built and is difficult to scale.

The M-CVS requires n additional blocks of attention whenever we add a new modality. We constrain our model to update only the weights corresponding to the new modality.

The training takes place in a stage wise manner. In stage 1, we train the image and text networks. From stage 2 onward, the weights corresponding to the image and the text branch are not updated. In stage 2, we map the video to the M-CVS by comparing it with the image and the text modality. Only the weights corresponding the video branch are updated. Similarly, in stages 3 and 4, the audio and 3D branches are added respectively. By deploying such a stage wise training technique, we can add any number of newer modalities without affecting the existing model and its inference. The stage wise training also reduces the time taken to add the new modality as only a fraction of the total weights have to be learned by the model.

B. Loss Functions

There are two sets of loss functions that we define in the M-CVS. We use a modified version of the triplet loss function described in [24]. We define two data points belonging to the same class as positive pairs and data points belonging to different classes as negative pairs. Our triplet loss is a combination of intermodality triplet loss and intramodality triplet loss which calculate losses across two modalities and within the same modality respectively. For any given such triplet, the loss value is incremented as per (1) where α_1, α_2 and α_3 are the margins and $(f_a^m - f_p^m)$ is the distance between the reference anchor and the positive pair belonging to modality m modality and $(f_a^m - f_n^m)$ is the distance between the reference anchor and the negative pair belonging to modality m. The γ_1, γ_2 and γ_3 are the weights of the three

Authorized licensed use limited to: Carleton University. Downloaded on September 21,2020 at 02:02:05 UTC from IEEE Xplore. Restrictions apply.

individual losses. x and y refer to the different modalities that belong to either of images, text, audio, video or 3D point clouds.

$$L_t = \sum_{x,y}^{i,t,v,a,3d} max(0, |f_a^x - f_p^x| - |f_a^x - f_n^x| + \alpha_1 + max(0, |f_a^y - f_p^y| - |f_a^y - f_n^y| + \alpha_2 + max(0, |f_a^x - f_p^y| - |f_a^x - f_n^y| + \alpha_3 \quad (1)$$

In addition to the triplet loss, we deploy the cross entropy loss to classify the samples into their respective categories.

$$L_c = \sum_{i}^{N} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i) \quad (2)$$

The total loss is given by the linear combination of (1) and (2).

C. Aligned Attention

Attention mechanisms are able to emphasize more on certain parts of the data at different times. Vaswani *et al.* [26] showed its application in language translations and Xu *et al.* [28] showed how attention is used for image captioning.

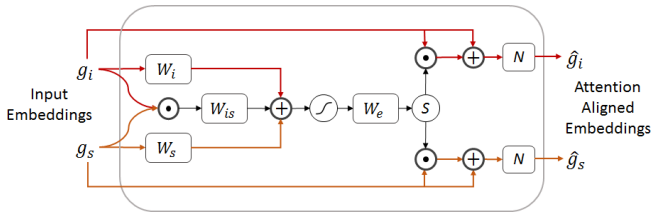

Fig. 2: Attention mechanism that aligns the new modality being added to the reference modality in the M-CVS.

The attention mechanism we develop takes in two input embeddings and transforms them to the M-CVS embedding. The attention mechanism that we use computes e_{is} where i and s are the two modalities. This is shown in (3).

$$e_{is} = W_e tanh(W_i g_i + W_s g_s + W_{is} g_i \cdot g_s + b) \quad (3)$$

The resulting vector from (3) is normalized to obtain the attention vector α_{is}.

$$\alpha_{is} = \frac{exp(e_{is})}{\sum exp(e_{is})} \quad (4)$$

The attention vector from (4) is passed through a residual block inspired from [7] as shown in (5) and (6) respectively.

$$\hat{g}_i = \alpha_{is} \cdot g_i + g_i \quad (5)$$

$$\hat{g}_s = \alpha_{is} \cdot g_s + g_s \quad (6)$$

$\hat{g}_i$ and $\hat{g}_s$ are further l_2 normalized before treating them as the final M-CVS representation of the data. Such an attention mechanism allows us to add newer modalities without changing existing models.

IV. Datasets

We evaluate our model on XMedia [15], [30] and XMediaNet [13], [14]. These datasets contain data belonging to five different modalities - image, text, audio, video, 3D points and is ideal to demonstrate the capabilities of the M-CVS. The XMedia dataset has 20 different categories and each sample of all the modalities belong to one of these categories. Similarly, the XMediaNet has 200 categories. Table I and II shows the number of samples and the different encoders for the XMedia and the XMediaNet dataset and their dimensions respectively.

TABLE I: Multimodal XMedia dataset [31] statistics.

Modality	#Train	#Test	Feature dim. (Method)
Image	4000	1000	4096 (CNN)
Text	4000	1000	3000 (BoW)
Video	400	100	4096 (C3D-CNN)
Audio	800	200	29 (MFCC)
3D Model	400	100	4700 (Light Field)

TABLE II: Multimodal XMediaNet dataset [13] statistics.

Media	Text	Image	Video	Audio	3D
Training	32,000	32,000	8,000	8,000	1,600
Testing	8,000	8,000	2,000	2,000	400

V. Implementation and Results

A. Implementation Specifications

Our M-CVS is a 512 dimensional space and all modalities are mapped in to this space using two fully connected layers. For instance, the 4k image feature is connected through two fully connected layers, both of which contain 512 neurons.

All our models are trained using TensorFlow [1]. We use a batch size of 128 and train all models for 50 epochs. We use Adam as our optimizer. During inference, each test sample is evaluated against all instances of the other modality. This comparison returns a similarity matrix that is then used to calculate the mean average precision score.

B. Evaluation Metrics

The mean Average Precision (mAP) scores [5] is used to evaluate our model. The Average Precision (AP) is computed for every query on the first R top-ranked retrieved data samples:

$$AP = \frac{1}{N} \sum_{r=1}^{R} p(r) \cdot rel(r) \quad (7)$$

Authorized licensed use limited to: Carleton University. Downloaded on September 21,2020 at 02:02:05 UTC from IEEE Xplore. Restrictions apply.

where, N is the total relevance of the samples in the retrieval, $p(r)$ is precision at r, and $rel(r)$ is a flag for the relevance of a given rank (one if relevant and zero otherwise). If the class label of the query is same as that of the retrieved sample, then the relevance is said to be true. mAP is the average of AP across all queries. We report $mAP@50$ (R = 50) for all experiments as is standard practice in the literature.

C. Experiments and Analysis

We report the $mAP@50$ on the XMedia and the XMediaNet multimodal datasets. Our results clearly indicate that the use of the attention model is very effective in improving the results. The image and text modalities have more samples to train on and hence are able to generalize better as indicated by the high retrieval scores between images and text. The number of samples in both the XMedia and the XMediaNet is imbalanced and the dimensions of the audio features is very low. We believe having a larger dataset could improve results.

D. CVS versus M-CVS model

The CVS model performs better than the M-CVS model, but requires all modalities to be trained in unison. Adding a new modality to the CVS model would require re-training all modalities from scratch. Our M-CVS model overcomes this by performing the training in a stage-wise manner. This results in a robust model that performs nearly as well as the CVS model as evident from III and IV. We can see from Figure 3 that the time taken per iteration for the CVS model is much higher when it is trained with incremental number of modalities as compared to the M-CVS model.

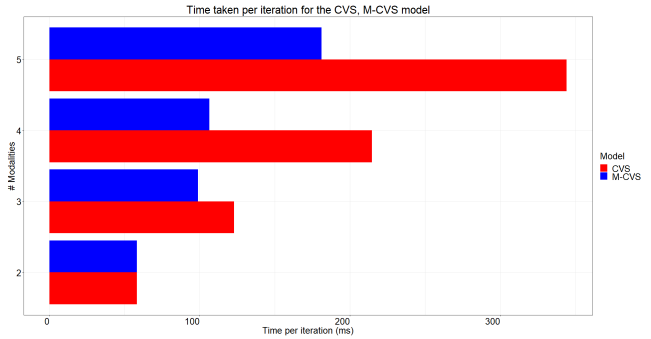

Fig. 3: The timing comparison between the CVS and the M-CVS model. The M-CVS

TABLE III: $mAP@50$ for XMedia dataset (I - image, T- text, A - audio, V - video, 3D - three dimension). Q is query and R is retrieval modality.

	Q \ R	I	T	A	V	3D
S2UPG [16]	I	–	0.270	0.265	0.264	0.394
	T	0.275	–	0.242	0.242	0.338
	A	0.274	0.244	–	0.207	0.363
	V	0.225	0.193	0.168	–	0.267
	3D	0.345	0.275	0.329	0.276	–
	Avg	0.273				
SCVM [33]	I	–	0.903	0.406	0.532	0.655
	T	0.889	–	0.507	0.549	0.722
	A	0.438	0.527	–	0.313	0.370
	V	0.553	0.580	0.302	–	0.450
	3D	0.603	0.679	0.370	0.426	–
	Avg	0.539				
CVS	I	–	0.908	0.708	0.801	0.731
	T	0.950	–	0.743	0.828	0.769
	A	0.416	0.477	–	0.341	0.420
	V	0.490	0.481	0.366	–	0.434
	3D	0.580	0.545	0.457	0.558	–
	Avg	**0.600**				
M-CVS	I	–	0.897	0.531	0.809	0.782
	T	0.943	–	0.538	0.840	0.822
	A	0.508	0.513	–	0.318	0.505
	V	0.521	0.211	0.202	–	0.282
	3D	0.616	0.627	0.457	0.558	–
	Avg	**0.567**				

TABLE IV: $mAP@50$ for XMediaNet dataset (I - image, T- text, V - video). Q is query and R is retrieval modality.

	Q \ R	I	T	V
CVS	I	–	0.775	0.605
	T	0.685	–	0.504
	V	0.395	0.383	–
	Avg	**0.558**		
M-CVS	I	–	0.777	0.457
	T	0.673	–	0.349
	V	0.227	0.260	–
	Avg	**0.457**		

VI. Conclusion

The CVS model is an effective technique to map similar concepts close and dissimilar concepts far, irrespective of their input modality. During training of CVS architectures, all input modalities need to be available for the model to be built. Adding subsequent modalities into the CVS model require re-training all modality weights from scratch. To avoid this problem, we introduce the M-CVS model that performs stage wise addition of modalities into the embedding space. Our M-CVS model preserves existing modality weights as new modalities are introduced. Results show that our M-CVS architecture can achieve performance similar to CVS, but at a significant compute cost savings.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional

Authorized licensed use limited to: Carleton University. Downloaded on September 21,2020 at 02:02:05 UTC from IEEE Xplore. Restrictions apply.

nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.
[3] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903, 2018.
[4] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.
[5] F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16. ACM, 2014.
[6] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. 1999.
[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
[8] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia*, 18(7):1363–1377, 2016.
[9] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.
[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
[11] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
[12] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015.
[13] Y. Peng, X. Huang, and Y. Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on circuits and systems for video technology*, 28(9):2372–2385, 2017.
[14] Y. Peng, J. Qi, and Y. Yuan. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing*, 27(11):5585–5599, 2018.
[15] Y. Peng, X. Zhai, Y. Zhao, and X. Huang. Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):583–596, 2015.
[16] Y. Peng, X. Zhai, Y. Zhao, and X. Huang. Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):583–596, 2016.
[17] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.
[18] J. Qi, X. Huang, and Y. Peng. Cross-media similarity metric learning with unified deep networks. *Multimedia Tools and Applications*, 76(23):25109–25127, 2017.
[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2015.
[20] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.
[21] S. Sah, S. Gopalakrishnan, and R. Ptucha. Cross modal retrieval using common vector space. In *2018 Image and Vision Workshop at IEEE Computer Vision and Pattern Recognition*. IEEE, 2018.
[22] S. Sah, S. Kulhare, A. Gray, S. Venugopalan, E. Prud'Hommeaux, and R. Ptucha. Semantic text summarization of long videos. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 989–997. IEEE, 2017.
[23] S. Sah, C. Zhang, T. Nguyen, D. K. Peri, A. Shringi, and R. Ptucha. Vector learning for cross domain representations. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–5. IEEE, 2017.
[24] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.
[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
[27] L. Wu, Y. Wang, and L. Shao. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 28(4):1602–1612, 2018.
[28] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
[29] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web*, 22(2):657–672, 2019.
[30] X. Zhai, Y. Peng, and J. Xiao. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6):965–978, 2013.
[31] X. Zhai, Y. Peng, and J. Xiao. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6):965–978, 2014.
[32] C. Zhang, S. Sah, T. Nguyen, D. Peri, A. Loui, C. Salvaggio, and R. Ptucha. Semantic sentence embeddings for paraphrasing and text summarization. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 705–709. IEEE, 2017.
[33] H. Zhang, T. Wang, and G. Dai. Semi-supervised cross-modal common representation learning with vector-valued manifold regularization. *Pattern Recognition Letters*, 2019.
[34] L. Zhen, P. Hu, X. Wang, and D. Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019.

Authorized licensed use limited to: Carleton University. Downloaded on September 21,2020 at 02:02:05 UTC from IEEE Xplore. Restrictions apply.