Radboud University Nijmegen

Bachelor's Thesis in Artificial Intelligence

Synesthesia-inspired cross-modal learning of common representation using GANs

Author:
Phuong Trinh
s1005865

Supervisor:
Dr. Pablo Lanillos
Artificial Intelligence,
Faculty of Social Sciences,
Radboud University Nijmegen

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in

Artificial Intelligence
Department of Artificial Intelligence

January 19, 2021

Reading Committee: P. Lanillos & Y. Güçlütürk

Radboud University

Abstract

Synesthesia is a phenomenon in which the stimulation of one sensory modality simultaneously leads to the sensation in one another. A well-known type of synesthesia is grapheme-color synesthesia, i.e., letters and digits are consistently associated with specific colors. Understanding the way cross-modal perception in synesthesia works has broadened the research in the field of artificial intelligence (AI) and its applications in dealing with multimodal data. Here, we describe a novel application of the cross-modal generative adversarial networks (CM-GANs) approach in order to learn the cross-modal common representation enforced by the shared semantic classes between the visual letter grapheme modality and the color modality, as in grapheme-color synesthesia. In order to evaluate the effectiveness of the model, we perform two cross-modal retrieval tasks: bi-modal retrieval (i.e., retrieving the correct matching color instances using letters as queries) and all-modal retrieval (i.e., retrieving the correct matching letter and color instances using letters as queries). The experimental results, obtained from the cross-modal retrieval tasks, are shown to be relatively high, indicating that the shared semantics between two modalities have a cross-modal effect in common representation learning. Regarding multimodal representation learning, we investigate the effectiveness of the CM-GANs network and discuss the approaches to overcome its shortcomings. As for grapheme-color synesthesia, we assess the applicability of the model in mathematically modeling the cross-modal perceptual association experience.

Contents

1 Introduction

1.1 Grapheme-color synesthesia

Synesthesia (originated from the Greek *syn* for "together" and *aisthesis* for "perception") is a perceptual condition that characterizes the experience of a cross-modal sensory association. That is, the stimulation of one sensory modality (i.e., the inducer) simultaneously elicits an ancillary concurrent sensation in one another (i.e., the concurrent) [6]. Prominent theories with emerging evidence have been proposed regarding the neural basis of synesthesia. One of the most dominant theories is the cross-activation theory by Ramachandran and Hubbard [20; 21], which postulates that the cross-activation between adjacent cortical areas responsible for the processing of the inducer and the concurrent in the brain arises from the excess of direct neural connections, or cross-wiring.

One known common form of synesthesia is grapheme-color synesthesia, where letters and digits consistently evoke mental imagery of specific colors, that is, the letters and digits are the inducer whilst the color is the concurrent [23]. For instance, the presence of the alphanumeric symbol *3* may elicit a sensation experience of the color yellow for a synesthete. According to the cross-activation model, these synesthetic associations are then due to the high level of inter-regional connectivity between the grapheme area (or the visual word form area - VWFA) and the adjoining color occipital brain areas V4 [1]. Subsequently, with the support of converging evidence, visually presented graphemes may lead to the activation of the visual color processing area V4, which essentially elicits grapheme-color synesthesia [22].

These key findings do not only provide insights into the neural basis and nature of sysnesthesia but can also be a significant source of inspiration to other fields of research where the cross-activation between different modalities is of great importance. In particular, understanding how cross-modal perception (e.g., the interaction between vision and auditory cues) in human brain works has broadened the research horizon in constructing sophisticated artificial intelligence (AI) models, especially cross-modal deep learning models, which can perform complex tasks with high-dimensional multimodal data. Exemplary applications for this involve video classification, sentiment analysis or cross-modal information retrieval, which typically include different modalities such as audio, image, text or video [9].

1.2 Cross-modal deep learning models

Real-world information consists of a variety of simultaneous modality inputs: videos are composed of audio signals and visual images, or images come with semantic labels, e.g., tags or captions. In this context, the term 'modality' refers to an inde-

pendent way or mechanism of encoding information [9]. Generally, humans are well-capable of integrating and processing such multimodal, possibly high-dimensional data. In case of synesthesia, the processing of sensory input (i.e., the inducer) also employs the activation of the sensory information that is not present in the original sensory input signal (i.e., the concurrent). In the exigenc[illegible] exploiting useful and valuable information in nowadays world, it is of paramount importance to build bio-inspired approaches that enable artificial agents and robots to effectively conduct multimodal processing by integrating information from different modalities.

Each modality is characterized by a (mathematical) representation and may have different statistical properties. For instance, images are compositions of real-valued pixels whereas for text data, word vectors are the standard representation, where each word from a given vocabulary is mapped to a vector of real values. As a consequence, feature vectors of different modalities that represent similar objects or concepts would become too discrepan[illegible] Such discrepancy and inconsistency in the representation structure is called the *heterogeneity gap* between different modalities. This phenomenon does not only hinder the processing of multimodal data in machine learning modules, but also imposes a challenge for research in AI and neuroscience to model the underlying cognitive process in humans. For instance, in the case of synesthesia, it is intricately complicated to simulate as there might exist a large heterogeneity gap between the inducer and the concurrent.

An approach to solve such issue is to learn the joint distribution over the input of different modalities [17]. In a nutshell, heterogeneous data input can be mapped into a commonly shared inner product space. In this vector space, data of different modalities that describes similar semantic concepts can be represented by similar feature vectors. Two feature vectors are said to be similar whenever they are close w.r.t. the metric induced by the inner product. Within this common space, cross-modal similarities can be directly computed using the Euclidean distance or the cosine metric. The cosine measure is suitable for similarity measurements in an inner product space and thus widely used in information retrieval or measuring cluster cohesion in data mining [27]. Subsequently, the heterogeneity gap among different modalities is narrowed, which allows for more efficient and convenient ways to process and manage multimodal data, thus helps achieve comprehensive results in various applications. Amongst these applications is cross-modal information retrieval (i.e., with data of one modality, the data of the absent modality can be retrieved) in entity-based search engines that aim to perform automatic content parsing on multimodal data.

In the recent years, deep learning has shown its significance in cross-modal learning research and applications as it can be used to learn joint representations [16; 26; 5]. Along this direction of research, generative adversarial networks (GANs) has proven

its strong suit for modelling data distribution and learning discriminative representations [7]. Peng et al. [18] have proposed a model called Cross-modal Adversarial Networks (CM-GANs) that utilizes the power of GANs to model the cross-modal joint distribution between high-dimensional data of different modalities, and by doing so mitigate the existing heterogeneity gap. CM-GANs was shown to effectively exploited the correlation across modalities for common representation learning, which was shown to outperform several existing methods when performing cross-modal retrieval tasks between different modalities [18]. Therefore, it is of substantial interest to investigate whether such a cross-modal deep learning model could be used to learn the common representation between sensory input of different modalities in human cognitive process and the implications it has regarding simulating the perceptual experience of synesthesia where a unimodal sensory input can activate that of another modality.

1.3 Aim of the project

In this project, our main objective is to construct a deep learning model in order to learn the cross-modal common representation between the inducer and the concurrent in synesthesia, specifically, the experience of cross-modal sensory association in grapheme-color synesthesia using the aforementioned CM-GANs approach. More concretely, this project takes a twofold approach: 1) to construct a cross-modal generative adversarial network based on the approach proposed by Peng et al. [18] to learn the cross-modal common representation enforced by the shared semantic classes between letter images and colors as in grapheme-color synesthesia; 2) to perform two kinds of cross-modal retrieval tasks: bi-modal retrieval (i.e., retrieving the correct matching color instances using letters as queries) and all-modal retrieval (i.e., retrieving the correct matching letter and color instances using letters as queries) based on the calculated similarity scores from the common representation using the cosine distance. We propose a hypothesis that the shared semantics between two modalities have a cross-modal effect in cross-modal common representation learning.

1.4 Overview

This section provides an overall overview of the content entailed in each following section. In Section 2, we discuss relevant literature in the current field of research in both multimodal representation learning and in understanding grapheme-color synesthesia using mathematical simulation. Section 3 presents the methodology including the acquirement of the data with the processing steps (Section 3.1), the model architecture as well as the training and optimization procedure of the model (explicated in Section 3.2). In Section 4, we describe the evaluation steps in order to assess the performance of the model using mean average precision (mAP) score

as evaluation metric. For the purpose of evaluation, we perform two cross-modal retrieval tasks, namely bi-modal retrieval and all-modal retrieval, for which the results are presented in Section 4.2. Section 5 provides a detailed discussion on the implications that the constructed model has as: 1) a method in cross-modal common representation learning and 2) a way to mathematically simulating the cross-modal perceptual association in grapheme-color synesthesia, followed by a brief conclusion in Section 6.

2 Literature review

In this section, relevant literature and research are discussed from two perspectives: deep learning based methods in learning the cross-modal association in synesthesia, and deep learning based methods in common representation learning.

First, in regard to synesthesia-inspired deep learning methods, Yamaguchi et al. [31] focused on modelling the cross-modal characteristics of synesthesia by constructing a multimodal model which consists of two deep neural networks (DNNs): one for image compression and one for audio-visual sequential learning. This model was shown to be able to reconstruct one modality from another modality analogously to synesthesia after being trained with multimodal data acquired from an experiment of a synesthesia study. On the same line of research, Bock [2] developed a GAN model that can simulate the perceptual experience of grapheme-color synesthesia by generating a colored version of the achromatic grapheme, i.e., the inducing stimulus from a given statistical distribution. The central principal of GANs, proposed by Goodfellow et al. [7], is to establish an adversarial process with a generative model G that aims to generate samples $G(z)$ drawn from the prior distribution $p_z(z)$ as if they have been drawn from the distribution $p_{data}(x)$ over the real data; and a discriminative model D that tries to discriminate between the real data and the generated, or "fake" data. This is considered a so-called 'minimax' game in which model parameters are optimized to solve the objective function $V(G, D)$, that is, to solve:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{data}(x)}[log(D(x))] + \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z)))].$$

Similar to the previously discussed model by [2], conventionally, the main focus of GANs-based research has mostly been on the problem of generating new data through adversarial training (see [19; 12]).

Due to its immense abilities and significance in current deep learning applications, GANs also belong to the class of deep learning based methods in common representation learning. In a recent study of Peng et al. [18], GANs were utilized not for the

classical purpose of generation, but to learn the cross-modal common representation between different modalities. Specifically, Peng et al. [18] proposed a cross-modal GANs (CM-GANs) approach for common representation learning, and thus provided a mean to narrow the so-called heterogeneity gap. This particular CM-GANs architecture was shown to effectively model the joint distribution over multimodal data simultaneously (see Figure 1). In this framework, the generative and discriminative models compete against each other to boost cross-modal correlation learning. Moreover, this model employs two parallel generative models, one for each modality where cross-modal convolutional autoencoders with weight-sharing constraint are used to exploit the cross-modal correlation and thus allows for better generation of the common representation. In this model, Peng et al. [18] also proposed a cross-modal adversarial training mechanism in which two different types of discriminative models: an intra-modality discriminator and an inter-modality discriminator are utilized. Here, the intra-modality discriminator aims to discriminate the reconstruction representation from the original representation within one modality whereas the inter-modality tries to discriminate the common representations between the two modalities.

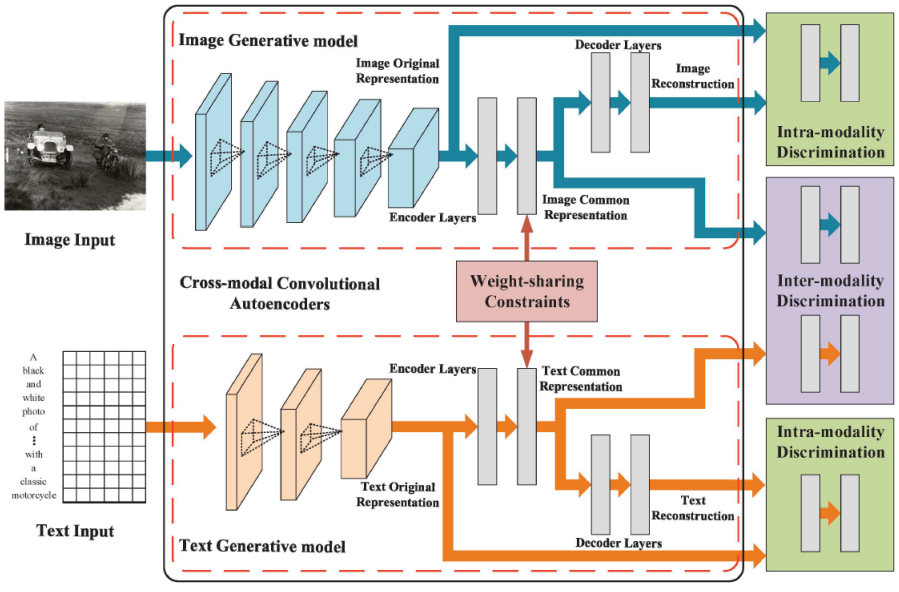


Figure 1: Schematic overview of the CM-GANs approach. Image take from [18].

In the current practice of modelling the joint distribution over data of different modalities for common representation learning, adversarial learning strategy is widely adopted. A few other prominent studies which also realized the advantages of adversarial learning are [29; 8; 30]. They were shown to achieve comparable results with

other existing state-of-the-art methods in common representation learning given the same datasets.

As specified earlier, the focus of this project lies in the adaptation of the CM-GANs model proposed in [18]. This was deemed to be the first model to utilize the power of GANs to learn cross-modal common representation learning and was shown to outperform 10 state-of-the-art methods in cross-modal retrieval for images and text. Understanding its strong abilities, we are interested in the possibility to generate a novel application of this approach in simulating the cross-modal phenomenon of synesthesia. The modifications made in this project are presented in the model shown in Figure 4. First of all, in our model, common representation learning is conducted on the multimodal dataset that consists of letters and colors, which belong to the letter modality and color modality respectively. Secondly, in the CM-GANs model by [18], the original representations of both modalities are extracted as feature vectors from the well-known VGG-Net (for images) [25] and Word2Vec (for text) [15]. For our model, two convolutional neural networks (CNNs) are constructed from scratch and pre-trained with the corresponding input type from our multimodal dataset, from which the original representation of the letters and colors are extracted. The complete architecture of our model is described in detail in Section 3.

3 Methods

A typical implementation of a machine learning module when dealing with multi-modal data involves three essential steps [9]: 1) Modality-specific feature extraction; 2) Multimodal representation learning and 3) Reasoning (e.g., clustering or classification). In this project, we attempt to address all three steps. Section 3.1 and Section 3.2 describe the initial data processing steps in detail, followed by the description of the feature extraction steps and the cross-modal common representation learning process, which are presented in Section 3.2.3. The last step (i.e., reasoning) belongs to Section 4 where we perform two cross-modal retrieval tasks (i.e., retrieving the corresponding color instances for the bi-modal retrieval task and retrieving both letter and color instances for the all-modal retrieval task given a letter as a query) in order to evaluate the performance of the constructed model.

3.1 Data processing

Representation learning methods typically assume that data of considered modalities are often semantically correlated to some extent. For instance, audio clips and text from the same web page are usually complementary or supplementary content-wise. Thus, in order to ensure a stronger correlation between the two separate modalities, semantic labels are incorporated. However, this might not always be the case in grapheme-color synesthesia as the semantic association does not arise naturally, i.e., letter-color pairs are usually not semantically correlated. In order to circumvent this, we include synesthetic association information between letters and colors in the network. This can be achieved by arbitrarily assigning colors to letter classes and creating a class label for each particular pair of corresponding letter and color. This means that the letter and color instance which form a 'synesthetic' pair have the same class label. This is analogous to the consistent association between letters/digits and colors in grapheme-color synesthesia. Here, for the sake of simplicity, we selected letter instances of 10 random classes to construct the modelling sample. The chosen letter classes are A, B, C, E, H, K, M, O, P and U. Here, we identify two cases of modelling:

(1) When the synesthetic letter-color pairs are unique, that is, each class of letter can only have one color and no color is assigned to more than one letter class (see Table 1).

Letter	A	B	C	E	H	K	M	O	P	U
Color	Red	Green	Yellow	Cyan	Pink	Orange	Purple	**Brown**	Blue	Gray

Table 1: Assigned unique synesthesia letter-color pairs.

(2) When the synesthetic letter-color pairs are not unique - multiple letters can be associated with the same color (see Table 2).

Letter	A	B	C	E	H	K	M	O	P	U
Color	Red	Green	Yellow	Cyan	Pink	Orange	Purple	**Blue**	Blue	Gray

Table 2: Assigned non-unique synesthesia letter-color pairs, here O and P are assigned the same color.

Thus, for the purpose of modelling, we construct a dataset consisting of letter instances and color instances derived from the grapheme modality and color modality, respectively. The raw handwritten letter images were extracted from the EMNIST dataset [4], stored in a 28x28 pixel image format. The EMNIST Letters subset contains 145,600 samples, which comprises 26 balanced classes of the uppercase and lowercase letters. As previously mentioned, we select the letter instances from the following classes: A, B, C, E, H, K, M, O, P and U. As for the color modality, for the case of modelling with unique letter-color pairs, we generate color instances of 10 different colors, namely red, green, blue, yellow, cyan, pink, orange, purple, brown and gray. Each color is represented as a 3-channel (R, G, B) 28x28 pixel image (see Figure 2.

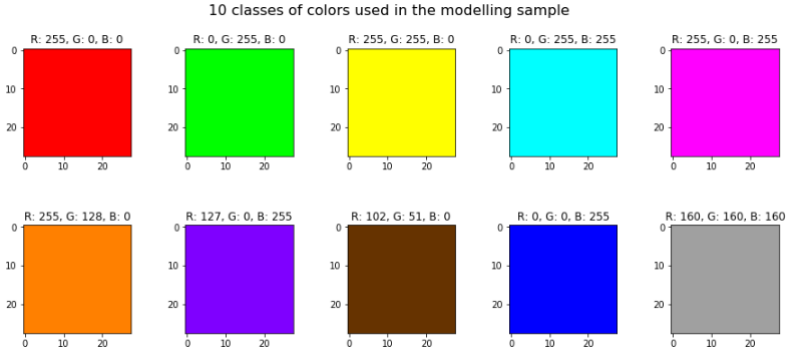


Figure 2: 10 classes of colors used in the modelling sample with their RGB values.

On the other hand, for the case of modelling with non-unique letter-color pairs, 9 different colors are generated for 10 classes of letter, namely red, green, blue, yellow, cyan, pink, orange, purple and gray, where the color blue is assigned to more than one class of letters (i.e., both letter O and letter P are blue).

This simplistic dataset is then used for the purpose of modelling and conducting experiments in common representation learning using the CM-GANs approach presented in Sections 3 and 4.

3.2 Cross-modal common representation learning

Our main focus lies in cross-modal representation learning for data consisting of letters and colors. In the following subsection, we first introduce the necessary mathematical notation.

3.2.1 Notation

The constructed dataset is denoted by $D = \{d_i\}_{i=1}^{N}$, with $d_i = (l_i, c_i)$, where $l_i \in \mathbb{R}^{d_l}$ denotes the i-th letter feature vector and $c_i \in \mathbb{R}^{d_c}$ denotes the i-th color feature vector. Note that d_l and d_c are the feature dimensions and usually $d_l \neq d_c$. In addition, each instance in the dataset comes with a class label, namely y_i^{letter} and y_i^{color} for the i-th letter instance and color instance in the dataset, respectively. The dataset is then split into the training dataset D_{train} and the test dataset D_{test} for the purpose of validation.

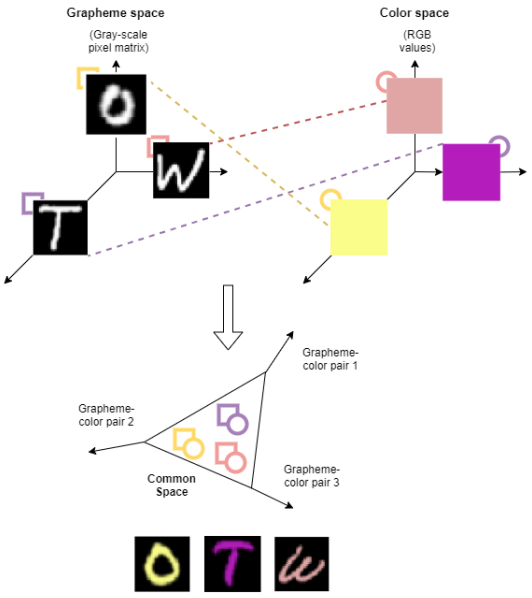


Figure 3: A schematic overview of the joint distribution learning (adapted from [9]), where heterogeneous data of different modalities is mapped into a common space, and the data of different modalities with similar semantics is represented by similar vectors.

Denote $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$ as the set of all letter instances and let $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ be the set of all color instances. Due to the heterogeneity gap between the letter features $\mathcal{L}$ and the color features $\mathcal{C}$, their cross-modal similarity cannot be computed directly for the purpose of cross-modal retrieval. Therefore, the challenge lies in

creating a commonly shared space $\mathcal{S}$ in which the letter features $\mathcal{L}$ and the color features $\mathcal{C}$ can be mapped to the letter common representation $\mathcal{S}_\mathcal{L}$ and color representation $\mathcal{S}_\mathcal{L}$ using the mapping functions $f_\mathcal{L} : \mathcal{L} \to \mathcal{S}_\mathcal{L}$ and $f_\mathcal{C} : \mathcal{C} \to \mathcal{S}_\mathcal{C}$, respectively (see Figure 3). The cross-modality similarity between the common representation $\mathcal{S}_\mathcal{L}$ and $\mathcal{S}_\mathcal{C}$ can then be directly calculated by adopting a distance metric for which we use the cosine similarity measure. Having obtained the vectors of common representation $\mathcal{S}_\mathcal{L}$ and $\mathcal{S}_\mathcal{C}$, the cosine similarity is defined as the cosine of the angle θ between them:

$$\text{cross-modal similarity} = cos(\theta) = \frac{\mathcal{S}_\mathcal{L} \cdot \mathcal{S}_\mathcal{C}}{\|\mathcal{S}_\mathcal{L}\|\|\mathcal{S}_\mathcal{C}\|} = \frac{\sum_{i=1}^{N} \mathcal{S}_{l_i} \cdot \mathcal{S}_{c_i}}{\sqrt{\sum_{i=1}^{N} \mathcal{S}_{l_i}^2}\sqrt{\sum_{i=1}^{N} \mathcal{S}_{c_i}^2}}, \tag{1}$$

where $\mathcal{S}_\mathcal{L} = \{\mathcal{S}_{l_1}, \mathcal{S}_{l_2}, \ldots, \mathcal{S}_{l_n}\}$ and $\mathcal{S}_\mathcal{C} = \{\mathcal{S}_{c_1}, \mathcal{S}_{c_2}, \ldots, \mathcal{S}_{c_n}\}$ are the sets of all letter common representations and color common representations, respectively.

Adopting the aforementioned CM-GANs approach, the model under investigation needs to meet two main requirements: 1) The underlying cross-modal correlation between heterogeneous data must be retained, and 2) The semantic consistency within each modality must be preserved. In this framework, a generative model is built with the objective to ensure that the first condition is satisfied. In addition, a discriminative model with an element of intra-modality discrimination is constructed, which plays a role in preserving the semantic consistency within each modality. The particulars regarding how this is conducted in the CM-GANs approach are explicated in the following subsections.

3.2.2 Model architecture

Employing the characteristics of GANs, the CM-GANs network architecture consists of two main models, namely the generative model $\mathcal{G}$ and the discriminative model $\mathcal{D}$. The generative model $\mathcal{G}$ aims to learn the joint distribution by modelling the cross-modal correlation and the reconstruction of information within each modality whereas the discriminative model aims to discriminate the information both between the two modalities and within each modality in order to boost the common representation learning. The detailed network structure is illustrated in Figure 4. On a high level, the generative model is constructed by two cross-modal convolutional autoencoders with weight-sharing whereas the discriminative model accounts for both intra-modality and inter-modality discrimination, each of which is handled by a corresponding discriminator. We shall now discuss the construction of both models in detail.

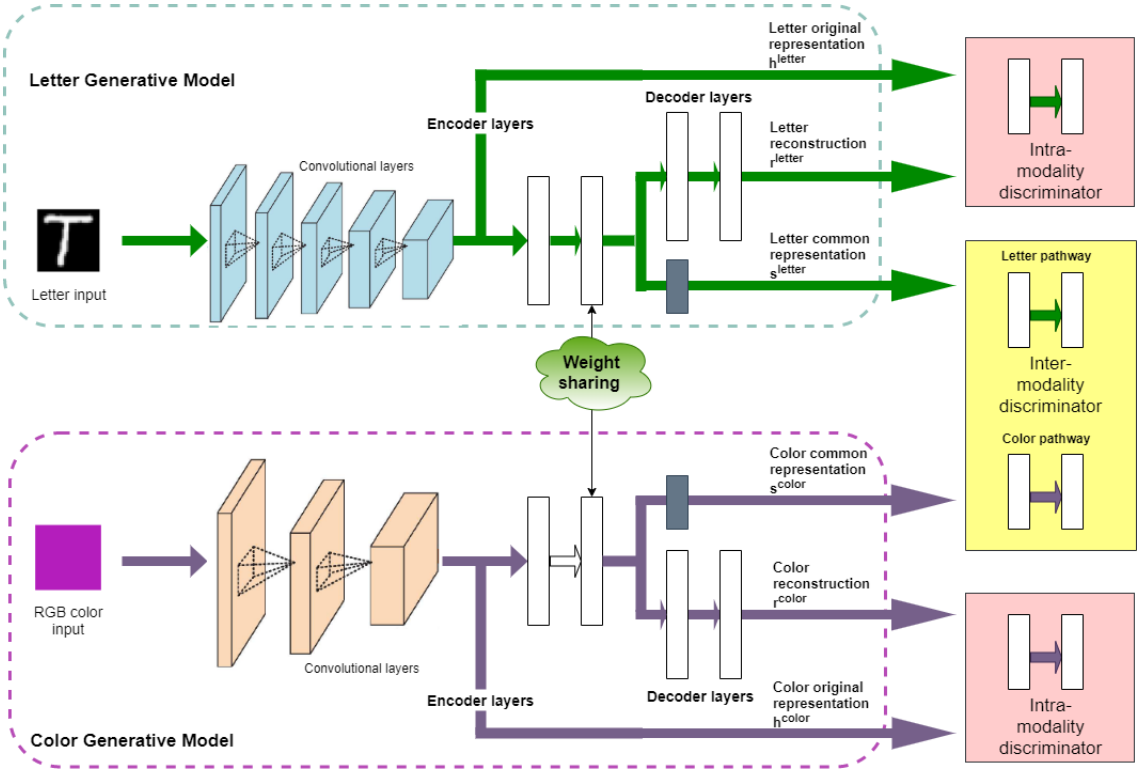


Figure 4: A schematic overview of the adapted CM-GANs model with two parallel generative models and two types of discriminative models: an intra-modality discriminative model for each modality and a two-pathway inter-modality discriminative model.

1) Generative model The network contains two parallel generative models, one for each modality. We denote with $\mathcal{G}_{\mathcal{L}}$ the generative model for the letter modality and similarly, let $\mathcal{G}_{\mathcal{C}}$ denote the generative model for the color modality. Essentially, each generative model is a cross-modal convolutional autoencoder comprising several encoder layers and decoder layers.

a) Encoder layers

Denote with $\mathcal{G}_{\mathcal{L}enc}$ the encoder for the letter modality and similarly, let $\mathcal{G}_{\mathcal{C}enc}$ denote the encoder for the color modality. Each encoder contains two sub-networks: a pre-trained convolutional neural network that learns the semantic information for the corresponding modality, followed by several fully-connected layers whose weights are shared with those of the other encoder to learn the cross-modal correlation between the two modalities.

As convolutional neural networks (CNNs) have emerged as a prominent technique

that can learn highly abstract and contextual image features with low computation cost [10], image features extracted from a CNN are often used as image representation in various studies and applications in image processing [29]. Hence, in this model, we use the extracted features to represent the images as well, which we refer to as the original representation of the image input. In order to obtain the feature vectors, for each modality, we construct a CNN. This is essentially the modality-specific feature extraction step in a typical multimodal representation learning study. Specifically, for letter data, each letter instance l_i of size 28x28 is first fed into the corresponding CNN, which is pre-trained on the same training data set of letter instances $\mathcal{L}_{train} \in D_{train}$. The letter CNN learns the features with letter images as input data using 2D convolutional layers. The network structure of this letter CNN is described in detail in Figure 5, followed by a visualization shown in Figure 6. We extract the 196-dimension feature vector, i.e., the output vector from the MaxPool2d-8 layer as the original representation of the letter input data, which we denote with h_i^{letter}.

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1           [-1, 4, 28, 28]              40
       BatchNorm2d-2           [-1, 4, 28, 28]               8
              ReLU-3           [-1, 4, 28, 28]               0
         MaxPool2d-4           [-1, 4, 14, 14]               0
            Conv2d-5           [-1, 4, 14, 14]             148
       BatchNorm2d-6           [-1, 4, 14, 14]               8
              ReLU-7           [-1, 4, 14, 14]               0
         MaxPool2d-8             [-1, 4, 7, 7]               0
            Linear-9                  [-1, 10]           1,970
       LogSoftmax-10                  [-1, 10]               0
================================================================
Total params: 2,174
Trainable params: 2,174
Non-trainable params: 0
```

Figure 5: Model summary of the letter CNN. The hidden layers consist of two convolutional layers, followed by a batch normalization layer. The activation function is a ReLU layer, followed by a max pooling layer. The output feature (which is for extraction) then goes through a fully-connected layer and log softmax layer for the purpose of classification.

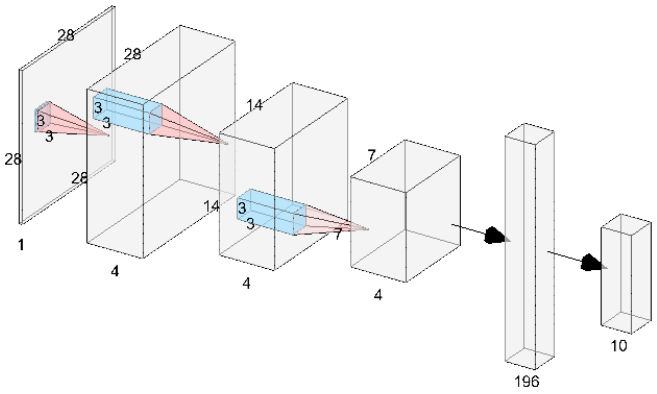


Figure 6: Visualization of the letter CNN model pretrained for letter feature extraction.

Similarly, for the color modality, a CNN for extracting the color feature vectors is constructed and pre-trained on the training color instances $\mathcal{C}_{train} \in D_{train}$. The configuration of this network is similar to that of the letter CNN in terms of the layer settings for each convolutional layers. The only slight difference lies in the fact that here, the color CNN handles 3-channel RGB images instead of grayscale images to generate the color original representation h_i^{color}, which is also extracted from the MaxPool2d-8 layer (see Figure 7 and 8).

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1           [-1, 4, 28, 28]             112
       BatchNorm2d-2           [-1, 4, 28, 28]               8
              ReLU-3           [-1, 4, 28, 28]               0
         MaxPool2d-4           [-1, 4, 14, 14]               0
            Conv2d-5           [-1, 4, 14, 14]             148
       BatchNorm2d-6           [-1, 4, 14, 14]               8
              ReLU-7           [-1, 4, 14, 14]               0
         MaxPool2d-8             [-1, 4, 7, 7]               0
            Linear-9                  [-1, 10]           1,970
       LogSoftmax-10                  [-1, 10]               0
================================================================
Total params: 2,246
Trainable params: 2,246
Non-trainable params: 0
```

Figure 7: Model summary of the color CNN. The hidden layers are quiet similar to those of the letter CNN, which consist of two convolutional layers, followed by a batch normalization layer. The activation function is a ReLU layer, followed by a max pooling layer. The output feature (which is for extraction) then goes through a fully-connected layer and log softmax layer for the purpose of classification.

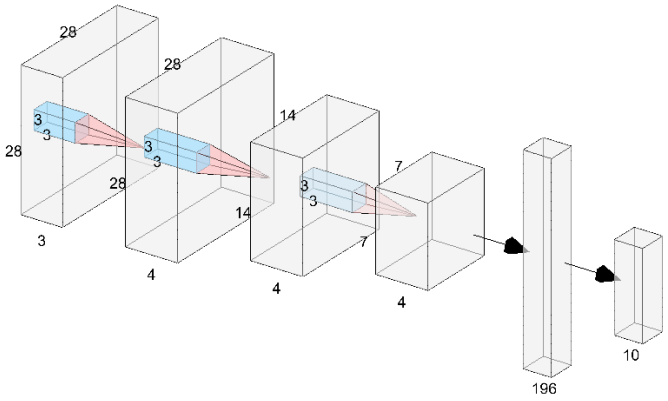


Figure 8: Visualization of the color CNN model pre-trained for color feature extraction.

Then, two fully-connected layers, which forms the second sub-network of the encoder, are employed to conduct common representation learning. Each layer is a linear layer, followed by a batch normalization layer that normalizes by re-centering and re-scaling and thus makes the network more robust against covariate shifts [11]. Then, a ReLU layer is added to increase the non-linearity of the images. From these encoder layers, we can obtain the common representation for the letters and the colors, denoted as s_i^{letter} and s_i^{color}, respectively. Here, there are 512 hidden units in the fully connected layer for both modalities, as such, the acquired common representation for each modality is essentially a feature vector with 512 entries.

Furthermore, it is important to emphasize that the weights of the second fully-connected layer for each modality are shared so as to effectively correlate the cross-modal common representation of the two modalities. The motivation behind this is the assumption we previously made, which indicates that the common representations for a synesthetic pair of corresponding letter and color should be as similar as possible. In practice, weight-sharing can be simplistically implemented by sharing the parameters of the second fully-connected layer between the two generative models. In addition, the weight-sharing layer is followed by a log softmax activation layer (which applies a softmax followed by a logarithm), which is colored gray in Figure 4, for further exploitation of semantic consistency within each modality [18].

b) Decoder layers

Let $\mathcal{G}_{\mathcal{L}dec}$ and $\mathcal{G}_{\mathcal{C}dec}$ denote the decoder for the letter modality and the color modality, respectively. The decoder layers intend to preserve the semantic consistency within

each modality by modelling the representation obtained from the corresponding CNN established above, namely h_i^{letter} and h_i^{color}. Specifically, the decoder layers aim to produce the reconstruction representation, denoted as r_i^{letter} and r_i^{color} from the common representation s_i^{letter} and s_i^{color}. In particular, the decoder layers simply consist of two fully-connected layers. The first layer has 512 dimensions whereas the dimension of the second layer is the same as the dimension of the original representation obtained by CNN, which is 196.

For the purpose of clarification, a summary of the representations generated by the generative models for both modalities is provided as follows.

Representation	**Notation**	
	Letter	Color
Original representation	$h_i^{letter} = \mathcal{G}_{\mathcal{L}conv}(l_i)$	$h_i^{color} = \mathcal{G}_{\mathcal{C}conv}(c_i)$
Common representation	$s_i^{letter} = \mathcal{G}_{\mathcal{L}enc}(l_i)$	$s_i^{color} = \mathcal{G}_{\mathcal{C}enc}(c_i)$
Reconstruction representation	$r_i^{letter} = \mathcal{G}_{\mathcal{L}dec}(s_i^{letter})$	$r_i^{color} = \mathcal{G}_{\mathcal{C}dec}(s_i^{color})$

Table 3: A summary of the representations generated by the generative models. Here, $\mathcal{G}_{\mathcal{L}conv}$ and $\mathcal{G}_{\mathcal{C}conv}$ denote the convolutional layers part - the first subnetwork of the encoder in the generative model for the letter modality and the color modality, respectively.

2) Discriminative model In this network, two types of discriminative models are presented in the cross-modal adversarial training: an intra-modality discriminator for the discrimination between the reconstruction representation r and the original representation h and an inter-modality discriminator that aims to discern from which modality the common representation s is from. The intra-modality discriminator and inter-modality discriminator are simultaneously utilized in the adversarial training procedure so as to enhance the cross-modal common representation learning.

a) Intra-modality discriminator

We define two intra-modality discriminators, i.e., the letter intra-modality discriminator, denoted as $\mathcal{D}_{\mathcal{L}}^{intra}$ and the color intra-modality discriminator $\mathcal{D}_{\mathcal{C}}^{intra}$. Each of them comprises a linear layer that transforms the input feature vector into a single-value score followed by a sigmoid activation layer. The aim of these intra-modality discriminators is straightforward - to discriminate the original representation h as the real data (labelled as 1) whereas the corresponding reconstructed representation r from the common representation s of the same modality is deemed as the generated, fake data (labelled as 0) in the training procedure of the minimax game.

b) Inter-modality discriminator

The inter-modality discriminator can be considered one of the key elements in cross-modal common representation learning as it helps boost the learning of the common representation by discriminating from which modality the generated common representation is from for both modalities. The inter-modality discriminator is essentially a two-path way network consisting of a letter pathway $\mathcal{D}_{\mathcal{L}}^{inter}$ and a color pathway $\mathcal{D}_{\mathcal{C}}^{inter}$. Each pathway is comprised of two fully-connected layers: the first fully-connected layer has 512 hidden units, followed by a batch normalization and a Leaky ReLU layer; the second fully-connected layer is a linear layer that takes the feature vector that is the output of the first layer and returns a single-value score, followed by a sigmoid layer. For the letter pathway, $\mathcal{D}_{\mathcal{L}}^{inter}$ discriminates the letter common representation s_i^{letter} as the real data (labelled 1) while the corresponding color common representation s_i^{color} and common representation of the mismatching letter instance $\hat{s}_i^{letter}$ are distinguished as the fake data (both labelled 0). For clarification, mismatching instance refers to an instance that is from the same modality but of a different class. In this network, mismatching instances are also exploited for better discrimination. Note that the inputs to the inter-modality discriminator are not just the aforementioned common representations but the concatenation of each common representation with its corresponding original representation, namely, $(s_i^{letter}, h_i^{letter})$ as the real data and $(s_i^{color}, h_i^{letter})$ with $(\hat{s}_i^{letter}, \hat{h}_i^{letter})$ as the fake data, where $\hat{h}_i^{letter}$ is the original representation of the letter mismatching instance and $\hat{s}_i^{letter}$ is the common representation of the letter mismatching instance. Using the concatenation of the common representation with the corresponding original representation allows for a more robust discrimination [18]. Similarly, the color pathway has the same structure but discriminates $(s_i^{color}, h_i^{color})$ as the real data and $(s_i^{letter}, h_i^{color})$ with $(\hat{s}_i^{color}, \hat{h}_i^{color})$ are the fake data instead, where $\hat{h}_i^{color}$ and $\hat{s}_i^{color}$ are the original representation and the common representation of the color mismatching instance, respectively.

3.2.3 Model training

The established generative model and discriminative model are trained in an adversarial manner to compete against each other so as to learn the common representation for the two modalities of interest. In this section, the adversarial training procedure is explained in detail, followed by the particulars regarding the performance of our implementation. We first provide a brief summary of the objective of each model in Table 4.

Generative model	**Discriminative model**
Each of the two generative models ($\mathcal{G}_\mathcal{L}$ and $\mathcal{G}_\mathcal{L}$) generates three kinds of representations: • the original representation h_i^{letter} or h_i^{color} from the convolutional layers • the common representation s_i^{letter} or s_i^{color} from the encoder layers • and the reconstruction representation r_i^{letter} or r_i^{color} from the decoder layers to fit the joint distribution of the letter and color modality.	• The intra-modality discriminator $D_\mathcal{L}^{intra}$ discerns the original representation h_i^{letter} as the real data and the reconstruction representation r_i^{letter} as the fake data. The same holds for $D_\mathcal{C}^{intra}$. • The two-pathway inter-modality discriminator, each discriminates the common representation as the real data while its mismatching common representation and the corresponding common representation from the other pathway are the fake data. – $\mathcal{D}_\mathcal{L}^{inter}$ distinguishes $(s_i^{letter}, h_i^{letter})$ as the real data and $(s_i^{color}, h_i^{letter})$ and $(\hat{s}_i^{letter}, \hat{h}_i^{letter})$ as the fake data – $\mathcal{D}_\mathcal{C}^{inter}$ distinguishes $(s_i^{color}, h_i^{color})$ as the real data and $(s_i^{letter}, h_i^{color})$ and $(\hat{s}_i^{color}, \hat{h}_i^{color})$ as the fake data

Table 4: Summary of the objective of the generative model and discriminative model in CM-GANs.

The goal of mitigating the heterogeneity discrepancy is to have feature vectors of different modalities that represent the same concept, ideally as similar as possible. In the minimax game, the main objective of the generative model $\mathcal{G}$ is twofold: 1) to learn the similar common representation between data that comes from different modalities but belongs to the same class and 2) to generate a reconstruction representation that is as close to the original representation as possible so as to 'trick' the discriminator. On the other hand, the discriminator $\mathcal{D}$ tries to make a distinction between the two intereseted modalities by carrying out intra-modality and inter-modality discrimination. In a nutshell, the generative model $\mathcal{G}$ and the discriminative model $\mathcal{D}$ compete in this minimax game with the value function $V(\mathcal{G}, \mathcal{D})$:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} V(\mathcal{G}, \mathcal{D}) = GAN_1(\mathcal{G}_\mathcal{L}, \mathcal{G}_\mathcal{C}, \mathcal{D}_\mathcal{L}^{intra}, \mathcal{D}_\mathcal{C}^{intra}) + GAN_2(\mathcal{G}_\mathcal{L}, \mathcal{G}_\mathcal{C}, \mathcal{D}_\mathcal{L}^{inter}, \mathcal{D}_\mathcal{C}^{inter}), \tag{2}$$

where

$$GAN_1(\mathcal{G}_\mathcal{L}, \mathcal{G}_\mathcal{C}, \mathcal{D}_\mathcal{L}^{intra}, \mathcal{D}_\mathcal{C}^{intra}) = \mathbb{E}_{l\sim p_\mathcal{L}(l)}[\mathcal{D}_\mathcal{L}^{intra}(l) - \mathcal{D}_\mathcal{L}^{intra}(\mathcal{G}_{\mathcal{L}dec}(l))] + \mathbb{E}_{c\sim p_\mathcal{C}(c)}[\mathcal{D}_\mathcal{C}^{intra}(c) - \mathcal{D}_\mathcal{C}^{intra}(\mathcal{G}_{\mathcal{C}dec}(c))],$$

and

$$\begin{aligned} GAN_2(\mathcal{G}_\mathcal{L}, \mathcal{G}_\mathcal{C}, \mathcal{D}_\mathcal{L}^{inter}, \mathcal{D}_\mathcal{C}^{inter}) = \mathbb{E}_{l,c\sim p_{\mathcal{L},\mathcal{C}}(l,c)}[&\mathcal{D}_\mathcal{L}^{inter}(\mathcal{G}_{\mathcal{L}enc}(l)) - \frac{1}{2}\mathcal{D}_\mathcal{L}^{inter}(\mathcal{G}_{\mathcal{C}enc}(c)) \\ &- \frac{1}{2}\mathcal{D}_\mathcal{L}^{inter}(\mathcal{G}_{\mathcal{L}enc}(\hat{l})) + \mathcal{D}_\mathcal{C}^{inter}(\mathcal{G}_{\mathcal{C}enc}(c)) \\ &- \frac{1}{2}\mathcal{D}_\mathcal{C}^{inter}(\mathcal{G}_{\mathcal{L}enc}(l)) - \frac{1}{2}\mathcal{D}_\mathcal{C}^{inter}(\mathcal{G}_{\mathcal{C}enc}(\hat{c}))], \end{aligned}$$

with $\hat{l}$ and $\hat{c}$ denoting the mismatching letter instance and the mismatching color instance, respectively.

In this minimax game, for each modality, the intra-modality discriminator tries to maximize the log-likelihood for correctly discriminating the original representation h as the real data and the reconstruction representation r as the fake data by ascending the letter intra-modality discriminator's stochastic gradient:

$$\nabla\theta_{\mathcal{D}_\mathcal{L}^{intra}} \frac{1}{N}\sum_{i=1}^{N}[log(1 - \mathcal{D}_\mathcal{L}^{intra}(r_i^{letter})) + log(\mathcal{D}_\mathcal{L}^{intra}(h_i^{letter}))], \tag{3}$$

and by ascending the color intra-modality discriminator's stochastic gradient:

$$\nabla\theta_{\mathcal{D}_\mathcal{C}^{intra}} \frac{1}{N}\sum_{i=1}^{N}[log(1 - \mathcal{D}_\mathcal{C}^{intra}(r_i^{color})) + log(\mathcal{D}_\mathcal{C}^{intra}(h_i^{color}))]. \tag{4}$$

As for the inter-modality discriminator, it aims to maximize the log-likelihood for correctly distinguishing the common representation of the interested modality as the real data, while the common representation of the other modality and the common representation of its mismatching instance are fake. Thus, $\mathcal{D}_\mathcal{L}^{inter}$ and $\mathcal{D}_\mathcal{C}^{inter}$ can be updated by ascending their stochastic gradients, which are:

$$\begin{aligned} \nabla\theta_{\mathcal{D}_\mathcal{L}^{inter}} \frac{1}{N}\sum_{i=1}^{N}[&log(\mathcal{D}_\mathcal{L}^{inter}(s_i^{letter}, h_i^{letter})) + \frac{1}{2}log(1 - \mathcal{D}_\mathcal{L}^{inter}(s_i^{color}, h_i^{letter})) \\ &+ \frac{1}{2}log(1 - \mathcal{D}_\mathcal{L}^{inter}(\hat{s}_i^{letter}, \hat{h}_i^{letter}))], \end{aligned} \tag{5}$$

and

$$\nabla\theta_{\mathcal{D}_{\mathcal{C}}^{inter}}\frac{1}{N}\sum_{i=1}^{N}[log(\mathcal{D}_{\mathcal{C}}^{inter}(s_i^{color},h_i^{color}))+\frac{1}{2}log(1-\mathcal{D}_{\mathcal{C}}^{inter}(s_i^{letter},h_i^{color}))+\frac{1}{2}log(1-\mathcal{D}_{\mathcal{C}}^{inter}(\hat{s}_i^{color},\hat{h}_i^{color}))]. \tag{6}$$

On the other hand, the generative model tries to minimize the objective function by descending its stochastic gradient. The letter generative model can be updated by descending the stochastic gradient:

$$\nabla\theta_{\mathcal{G}_{\mathcal{L}}}\frac{1}{N}\sum_{i=1}^{N}[log(\mathcal{D}_{\mathcal{C}}^{inter}(s_i^{letter},h_i^{color}))+log(\mathcal{D}_{\mathcal{L}}^{intra}(r_i^{letter}))]. \tag{7}$$

Similarly, by descending the following stochastic gradient, the color generative model can be updated:

$$\nabla\theta_{\mathcal{G}_{\mathcal{C}}}\frac{1}{N}\sum_{i=1}^{N}[log(\mathcal{D}_{\mathcal{L}}^{inter}(s_i^{color},h_i^{letter}))+log(\mathcal{D}_{\mathcal{C}}^{intra}(r_i^{color}))], \tag{8}$$

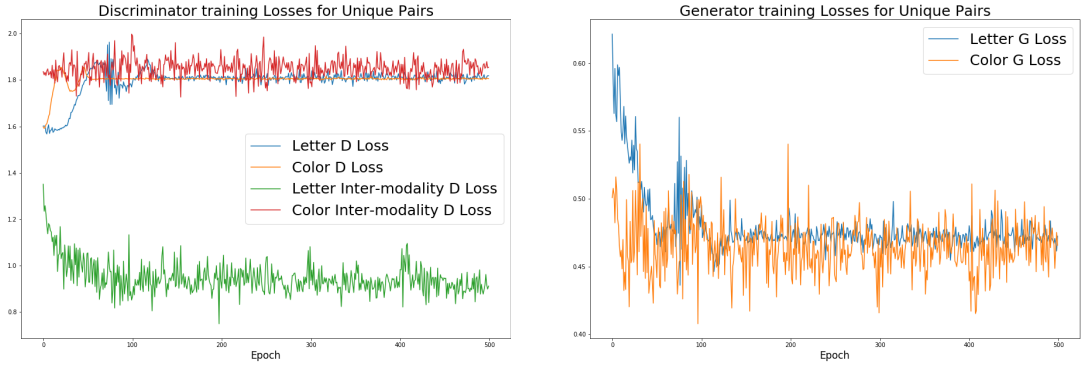
Following the minimax game defined in Eq. (2), the generative model and the discriminative model are pitted against each other in the adversarial training process which is discussed in detail in Algorithm 1 as provided below.

Algorithm 1 CM-GANs training process adapted from [18]

procedure TRAINING
  Inputs:
  Training set D_{train}, batch size N, learning rate α and the number of training steps K to train the generative model.
  repeat
    Sample letter and color pair $\{d_i = (l_i, c_i)\}_{i=1}^{N} \in D_{train}$ and mismatching instance for each of them.
    Generate h_i^{letter}, r_i^{letter} and s_i^{letter} from $\mathcal{G}_{\mathcal{L}}(l_i)$.
    Generate h_i^{color}, r_i^{color} and s_i^{color} from $\mathcal{G}_{\mathcal{L}}(c_i)$.
    Update $D_{\mathcal{L}}^{intra}$ by ascending its stochastic gradient in Eq. (3).
    Update $D_{\mathcal{C}}^{intra}$ by ascending its stochastic gradient in Eq. (4).
    Update $D_{\mathcal{L}}^{inter}$ by ascending its stochastic gradient in Eq. (5) and $D_{\mathcal{C}}^{inter}$ by ascending its stochastic gradient in Eq. (6).
    for K steps **do**
      Sample letter and color pair $\{d_i = (l_i, c_i)\}_{i=1}^{N} \in D_{train}$ and mismatching instance for each of them.
      Update $\mathcal{G}_{\mathcal{L}}$ by descending its stochastic gradient in Eq. (7).
      Update $\mathcal{G}_{\mathcal{C}}$ by descending its stochastic gradient in Eq. (8).
  until convergence

Here, with batch size $N = 128$, we train the CM-GANs model for 500 epochs and the generative model is trained with the number of training steps $K = 5$ in each epoch in order to ensure the balance between the discriminative model and the generative model. During the training phase, the convolutional layers of $\mathcal{G}_{\mathcal{L}enc}$ and $\mathcal{G}_{\mathcal{C}enc}$ are 'frozen', that is, the parameters of these layers remain constant as the main problem lies in cross-modal correlation learning but not classification. In addition, Adam optimization is used with the learning rate $\alpha = 0.0002$, which leverages the immense ability of adaptive learning rates methods to find individual learning rates for each parameter [13].

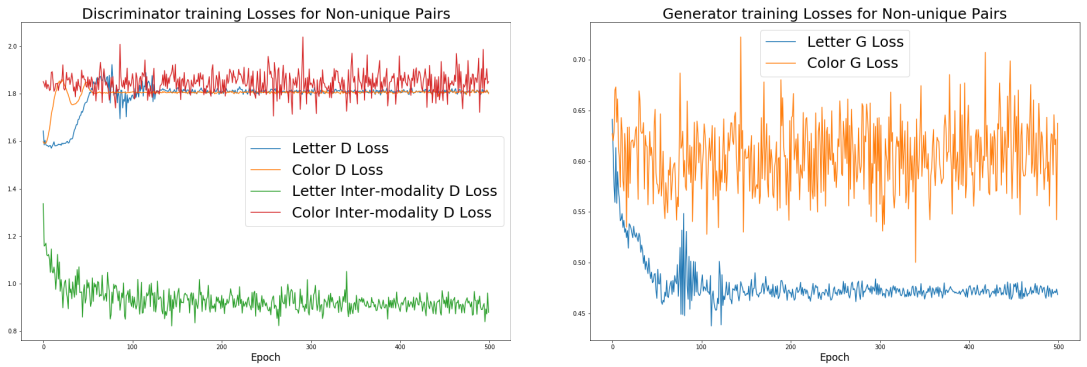
With regards to loss, during the training phase, the discriminator loss and generator loss over all samples are recorded after each epoch using binary cross-entropy loss function. The model is trained for the two modelling cases: 1) when the letter-color pairs are unique and 2) when the letter-color pairs are not unique on the corresponding dataset constructed in Section 3.1. These two cases are defined for cross-modal common representation learning so as to gain a better insight into the effectiveness of the CM-GANs approach. We hypothesized that in the second case where one color is assigned to multiple letter classes, semantic duplication or inconsistency might occur and thus, hindrances the performance of the model in cross-modal retrieval. For each case, the training loss of of the generative model and the discriminative model are plotted as a function of the number of epochs, as shown in Figures 9 and 10.



(a) Discriminative model.

(b) Generative model.

Figure 9: Training loss over 500 epochs of the generative model and the discriminative model for the case of unique letter-color pairings.



(a) Discriminative model.

(b) Generative model.

Figure 10: Training loss over 500 epochs of the generative model and the discriminative model for the case of non-unique letter-color pairings.

4 Results

4.1 Evaluation metric

As previously discussed, in the learned common space, ideally multimodal data with similar semantics is represented by similar feature vectors and we measure the cross-modal similarity with the cosine measure defined in Eq. (1).

In order to obtain a comprehensive evaluation of the model performance in common representation learning, we perform two cross-modal retrieval tasks in which letters are taken as queries to retrieve: the correct matching color instances (i.e., bi-modal retrieval) and the correct matchings of both letters and colors (i.e., all-modal retrieval) from the testing set D_{test} using the calculated cosine similarity scores. The evaluation process is straightforward, which consists of the three following steps:

(1) Common representation learning by sufficiently training the CM-GANs model as described in Section 3.

(2) Once the joint distribution has been learned, generate the common representation from the testing data using the learned generative models.

$$\mathcal{S}_{\mathcal{L}} = \mathcal{G}_{\mathcal{L}}(\mathcal{L}_{test}),$$
$$\mathcal{S}_{\mathcal{C}} = \mathcal{G}_{\mathcal{C}}(\mathcal{C}_{test}),$$

where $\mathcal{L}_{test}$ and $\mathcal{C}_{test}$ denote the letter instances and the color instances from the testing dataset D_{test}, respectively.

(3) Use the cosine distance metric established in Eq. (1) to compute the cross-modal similarity between the obtained common representations $\mathcal{S}_{\mathcal{L}}$ and $\mathcal{S}_{\mathcal{C}}$. This results in a distance matrix that is then used to perform the aforementioned cross-modal retrieval task.

(4) Apply the evaluation metric on the results of the cross-modal retrieval task.

The evaluation of the results of all queries (Step 4) is conducted in terms of the mean average precision (mAP) score, which is known as a classical performance evaluation criterion in cross-modal retrieval. As an evaluation metric, the mAP score for each letter-color class quantifies how good the model is at performing the queries. Specifically, the mAP is calculated as:

$$mAP = \frac{\sum_{i=1}^{Q} AvgPrecision(q)}{Q}, \tag{9}$$

where q is the i-th query in the set consisting of Q queries.

4.2 Performance comparison

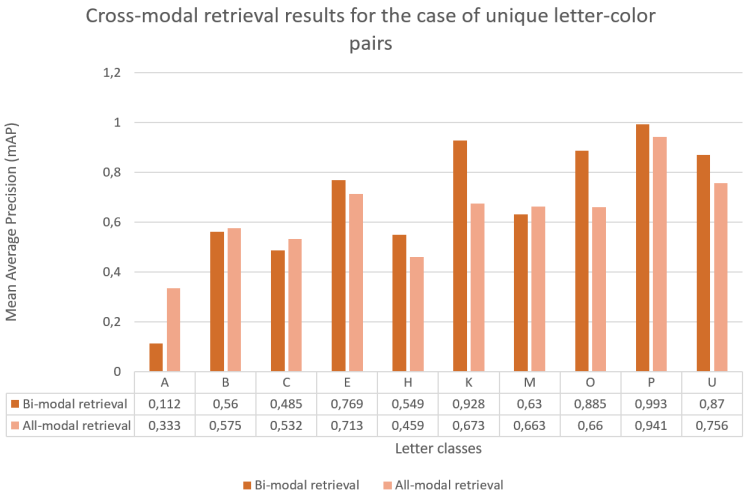
As previously established, we trained the model for two modelling cases in order to investigate further the effectiveness of the CM-GANs approach in cross-modal common representation learning using cross-modal retrieval. In particular, we perform two different cross-modal retrieval tasks, which are as follows.

- Bi-modal retrieval: Letters are taken as queries in order to retrieve the correct matching color instances from the testing dataset D_{test}.
- All-modal retrieval: Letters are taken as queries in order to retrieve both correct matching letter instances and color instances from the testing dataset D_{test}.

For each task, we conduct $16,000$ queries on the corresponding testing dataset D_{test} for the case of unique letter-color pairs and for the case of non-unique letter-color pairs, and compute the mAP scores for each synesthesia letter-color class.

4.2.1 Case 1: Unique letter-color pairs

This modelling case only concerns the letter-color pairs that are unique, as such, letter instances of one class cannot have more than one color and a color can only be assigned to one class of letters. The bi-modal retrieval results and all-modal retrieval results including the mAP scores for all classes in this case are represented in Figure 11.



	A	B	C	E	H	K	M	O	P	U
Bi-modal retrieval	0,112	0,56	0,485	0,769	0,549	0,928	0,63	0,885	0,993	0,87
All-modal retrieval	0,333	0,575	0,532	0,713	0,459	0,673	0,663	0,66	0,941	0,756

Figure 11: The mAP scores for each synesthesia letter-color class in the case of unique letter-color pairs in both the bi-modal retrieval task and the all-modal retrieval task.

From the experimental results, it can be observed that CM-GANs achieve relatively high accuracy on performing the cross-modal letter → color retrieval task with one exception for the synesthetic letter A with the color red, whose score is quite low. The calculated mAP score for the letter A in the case of non-unique letter-color pairs is 0.333 and in the case of unique letter-color pairs, ir seems to fall on the chance level (mAP = 0.112). It might be the case that the variance between the hand-written letters A has caused the confusion when performing the cross-modal retrieval task. As mentioned in Section 3.1, the hand-written letter images are extracted from the EMNIST Letters dataset, which consists of both lowercase and uppercase letters. The mix of the lowercase and uppercase letters has caused significantly high variance in the dataset, causing the problem of misclassification between the lowercase and uppercase versions of the same letters [4]. Among the 10 letter classes of interest, the letter A is most misclassified, as shown in the confusion matrix obtained from the letter CNN in Figure 12. Moreover, according to [4, Fig. 6], based on the resulting confusion matrix obtained from the OPIUM-based classifier for all 26 letter classes, the letter A is also one of the letters where most of the confusions occur. This was deemed to be due to the inconsistencies in the handwriting of individuals rather than mislabeling in the dataset. This has inevitably led to the confusion in the intra-modality discriminator, which appears to suffer from preserving the semantic consistency of the letter A within the letter modality.

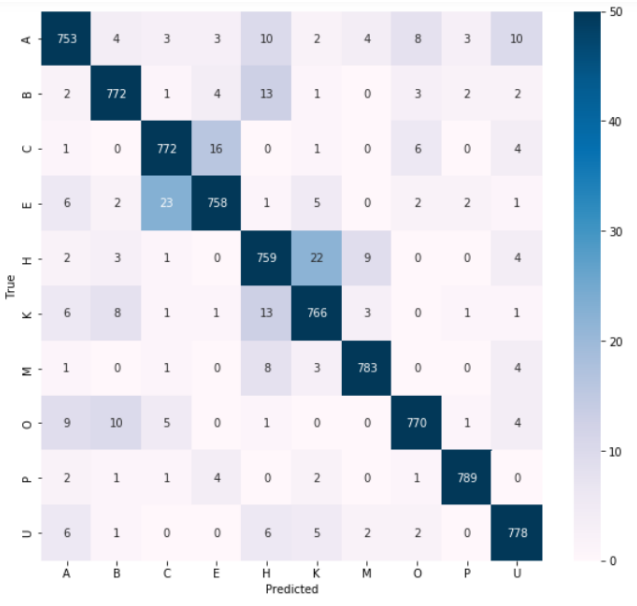


Figure 12: The confusion matrix for the 10 letter classes from the EMNIST Letters Dataset for constructed letter CNN. This shows that most of the confusions occur in the letter class A, which inherently contains ambiguity.

Furthermore, it can also be observed that the letter classes with high accuracy in classification also obtain higher score in the cross-modal retrieval task.

4.2.2 Case 2: Non-unique letter-color pairs

In this case, the uniqueness constraint is disregarded as one color can be assigned to more than one class of letter instances. Here, we are interested in the change of the model performance when such a case of semantic duplication is present. As mentioned in Section 3.2.3, we propose a hypothesis that semantic duplication, or inconsistency will worsen the model performance. As can be observed in Figure 11, the retrieval results for the letter O (mAP = 0.885 in the bi-modal retrieval task and 0.660 in the all-modal retrieval task) and P (mAP = 0.993 in the bi-modal retrieval task and 0.941 in the all-modal retrieval task) are quite high. Thus, in the case of non-unique letter-color pairs, we specifically investigate the change in the mAP scores for these two classes as letter instances from both the letter class O and P are now assigned the same color (i.e., blue). Similarly to the previous case, the bi-modal retrieval results and the all-modal retrieval results are shown for each letter-color pairs, as illustrated in Figure 13.

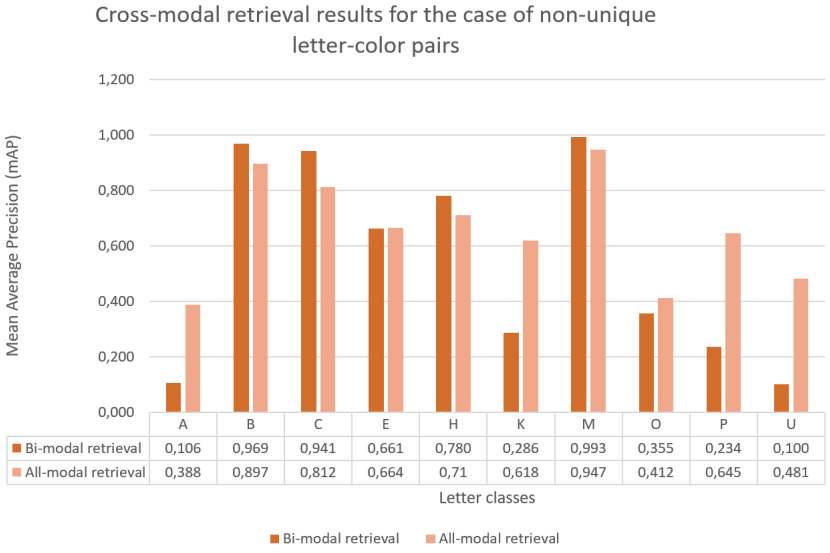


Figure 13: mAP scores for each synesthesia letter-color class in the case of unique letter-color pairs in both bi-modal retrieval task and all-modal retrieval task.

Based on the above experimental results, it can be observed that comparing to the first case, the model performance on retrieving the correct matching instances for queries of the letter A remains the same while there appears a to be a slight

improvement for the class of letters B and C. Remarkably, the mAP scores for both the letter O (mAP = 0.355 in the bi-modal retrieval task and 0.412 in the all-modal retrieval task) and P (mAP = 0.234 in the bi-modal retrieval task and 0.645 in the all-modal retrieval task) have decreased significantly, though still well-above the chance level. This evidence supports our aforementioned hypothesis - the presence of color duplication, which essentially leads to semantic inconsistency has caused the confusion for the model when learning the common representation between the two modalities of interest.

Nonetheless, the overall results suggest that the cross-modal correlation learning can be enhanced by the adversarial training process in the CM-GANs approach. It can be concluded that the cross-modal joint distribution between data of the letter modality and the color modality is captured and thus helps boost the learning of the cross-modal common representation. However, the model is still susceptible to issues such as semantic duplication and inconsistency.

5 Discussion

In this section, we discuss the implications of the constructed model from two perspectives: regarding cross-modal common representation learning and regarding modelling grapheme-color synesthesia. In particular, we reflect on the attainments and identify the limitations as well as potential hindrances, which provide directions to future work for further improvements of the model in both cross-modal common representation learning and modelling the cross-modal association phenomenon in synesthesia.

5.1 Implications regarding cross-modal common representation learning

The cross-modal retrieval results build upon the existing evidence that the CM-GANs approach could successfully learn the cross-modal common representation between two different modalities. The three main elements that contribute to the effectiveness of this approach are summarized as follows. First, the weight-sharing between the fully connected layers of the two generative models holds responsibility for learning the cross-modal correlation. Along with the preservation of semantic consistency handled by the decoder layers, this allows for mutual boosting of common representation learning. Secondly, the inter-modality discriminator also plays a major role in cross-modal representation learning. Succinctly, the inter-modality discriminator acts like a modality classifier that discriminates from which modality the common representations yielded by the generative models come from. The intra-modality discriminator is complementary to the inter-modality discriminator as it is for maintaining the modality-specific semantic consistency. Lastly, the adversarial learning strategy, which entails the competition between the generative model and discriminative model also boosts the cross-modal correlation learning process [18]. In order to verify the effectiveness of the three aforementioned aspects, one can conduct a baseline experiment where an aspect is removed and compare the differences in the model performance. Note that the inter-modality discriminator cannot be excluded due to its immense role in correlation learning during adversarial training. Conducting these baseline experiments is out of scope of this project as to avoid repetition in appraising the key aspects of this CM-GANs approach.

Even though the constructed model was shown to accomplish the goals established throughout this work, there are several hindrances for which further investigations are needed. Firstly, even though the overall results are shown to be relatively good, there is no baseline approach implemented to compare the model performance in cross-modal common representation learning, especially for data inputs from the letter modality and the color modality. Secondly, the reliability of the results might

be constrained by the chosen representation of the color. The colors are constructed as 3-channel RGB images and the color feature vectors are extracted from a convolutional neural networks for color images. This implicitly implies that the mathematical representation of the colors might be similar to that of the letters and thus the heterogeneity gap between the letter modality and the color modality is not as significant as other modalities. This might also explain the relatively high cross-modal retrieval results. The third limitation is with the generalizability of the results. The constructed model was only validated on a very small case of semantic conflict which only occurs between two classes of letters. More occurrence of semantic and noise, which are common issues in multimodal representation learning might hinder the performance of the model significantly. As for future work, a method for resolving such problems is to incorporate the reasoning ability into the cross-modal representation learning networks. In particular, using the reasoning mechanism would allow the network to actively select the sorely needed evidence, and thus helps mitigate the impact of the aforementioned common issues [9]. Another point of concern is among the implementation aspects of GANs, which is prone to training instability, usually reflects in mode collapse or convergence failure [24]. In order to circumvent such convergence issues during the training phase, we employ the transfer learning in which we pretrain the convolutional neural networks then freeze the convolutional layers in our model. In addition, in each training epoch, the generative model is trained K times so as to avoid the imbalance between the generative model and the discriminative model. All in all, this can give directions for improvements to further research in common representation learning using GANs-based methods.

5.2 Implications regarding grapheme-color synesthesia

In this project, even though the practical implementation of this model based on the CM-GANs approach lies the in area of multimodal representation learning, it is framed in the context of grapheme-color synesthesia. Thus, in this subsection, we discuss the implications our model has regarding simulating cross-modal association between modalities in the perceptual experience of grapheme-color synesthesia.

First, in grapheme-color synesthesia, the synesthetic association between letters/digits and colors are always consistent for an individual [28]. Our experimental results suggest that the constructed model is capable of learning the consistent association of the ten letter-color pairs shown in Table 1. This is in line with the synesthetic consistency in grapheme-color synesthesia. Secondly, the association in grapheme-color synesthesia is unidirectional, that is, the letter and digits induce a certain color but the colors but not vice versa [3]. In cross-modal common representation learning, the joint distribution of different modalities is essentially learned and thus, cross-modal retrieval can be implemented in the two following ways: 1) bi-modal

retrieval, including letters to colors retrieval task and colors to letters retrieval task and 2) all-modal retrieval task in which letters are queried in order to retrieve all correct matching letter and color instances. However, considering the unidirectionality in grapheme-color synesthesia, we only investigate the results on the bi-modal letters to colors retrieval task and the all-modal retrieval task.

Despite being able to learn the cross-modal common representation between the two modalities with our model, in grapheme-color synesthesia, in order to attain a deep understanding of this phenomenon, much more is required from a mathematical simulation than this model could provide. For the sake of simplicity, only 10 pairs of letter-color were studied and the digits were not taken into account. Leaving out classes of letters and digits have put a constraint on the generalizability of the results in the context of grapheme-color synesthesia. Moreover, the experience of grapheme-color synesthesia varies between synesthetic individuals, which is not accounted for in the construction of the model. In addition, in grapheme-color synesthesia, often multiple letters or digits can elicit the sensation of the same color. However, the results from the second experiment have shown that the model does not perform well in such cases due to semantic duplication, which is deemed to hinder the learning process. Thus, further investigations are needed to improve the part of the network architecture responsible for maintaining the consistency of the classes within one modality, e.g., consider using softmax loss. Furthermore, in order to provide a new insight or explain the neural basis of grapheme-color synesthesia and the levels of activation in the brain areas for a synesthetic individual, further analysis should be performed.

6 Conclusion

In this project, we addressed cross-modal common representation learning between the data of the letter modality and the color modality, analogously to the inducer (i.e., letters and digits) and the concurrent (i.e., colors) in the cross-modal perceptual association of grapheme-color synesthesia. In particular, we constructed an adversarial model based on the CM-GANs approach proposed by [18] so as to effectively learn the common representation between grapheme images and colors. The cross-modal GANs architecture was adapted to learn the joint distribution over the data of the letter modality and the color modality in an adversarial training process. Here, the weight-sharing mechanism as well as the preservation of semantic consistency within each modality played an immense role in learning the cross-modal correlation. In addition, the discriminative model accounts for both inter-modality and intra-modality discrimination, which mutually boost the process of common representation learning. Two experimental cross-modal retrieval tasks were carried

out in order to obtain a comprehensive evaluation of the model performance, which is done in terms of the mean average precision score for each of the ten classes of synesthetic letter-color pairs. The experimental results showed the model has successfully learned the cross-modal common representation between the modalities of interest. We also discussed several shortcomings regarding the generalizability of the model in cross-modal representation learning and its applicability in mathematically modelling the phenomenon of synesthesia. Depending on the research interest, which either lies in the problem of multimodal representation learning, or gaining insights into the perceptual experience that involves two modalities in synesthesia, further investigations are required. For multimodal representation learning, a reasoning mechanism can be integrated to mitigate the impact of semantic inconsistency and ambiguity on the model performance. With the reasoning mechanism, a system can actively select the evidence that is sorely needed. The incorporation of the reasoning mechanism in multimodal representation learning will endow machines with more effective and less prone to errors learning capabilities [9]. As for future work, an interesting open issue for cross-modal representation learning is to further adjust the CM-GANs approach to deal with other different modalities and for when there are more than two modalities, leaning towards equipping machine learning modules with better intelligent cognitive abilities.

References

[1] G. Bargary & K. J. Mitchell, "Synaesthesia and cortical connectivity", *Trends in neurosciences*, vol. 31, no. 7, pp. 335–342, 2008.

[2] J. Bock, "A Deep Learning Model of Perception in Color-Letter Synesthesia", *Big Data and Cognitive Computing*, vol. 2, no. 10, 2018.

[3] D. D. Cid, J. F. Awad, B. C. Hackney, J. Buenrostro, & S. A. Drew, "Are Synesthetic Perceptions a 2 Way Street?: A Study On The Bidirectionality of Grapheme-Color Synesthesia", *Journal of Vision*, vol. 16, pp. 466, 2016.

[4] G. Cohen, S. Afshar, J. Tapson & A. van Schaik, "EMNIST: An extension of MNIST to handwritten letters", *arXiv*, 2017.

[5] F. Feng, X. Wang, R. Li, "Cross-modal retrieval with correspondence autoencoder", *ACM Int. Conf. on Multimedia*, pp. 7–16, 2014.

[6] L. Gertner, I. Arend & A. Henik, " Numerical synesthesia is more than just a symbol-induced phenomenon", *Front Psychol*, vol 4, no. 860, 2013.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville & Y. Bengio, "Generative adversarial nets", *Advances in Neural Information Processing Systems (NIPS)*, pp. 2672-2680, 2014.

[8] J. Gu, J. Cai, S. Joty, L. Niu & G. Wang, "Look, imagine and match: improving textual-visual cross-modal retrieval with generative models", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7181-7189, 2018.

[9] W. Guo, J. Wang & S. Wang, "Deep Multimodal Representation Learning: A Survey", *IEEE Access*, vol. 7, pp. 63373-63394, 2019.

[10] S. Indolia, A. K. Goswami, S. P. Mishra & P. Asopa, "Conceptual Understanding of Convolutional Neural Network - A Deep Learning Approach", *Procedia Computer Science*, vol. 132, pp. 679-688, 2018.

[11] S. Ioffe & C. Szegedy, Christian, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", *arXiv*, 2015.

[12] T. Karras, T. Aila, S. Laine & J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation", *ArXiv*, 2018.

[13] D. P. Kingma & J. L. Ba, "Adam: A method for stochastic optimization", *arXiv*, 2014.

[14] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet: classification with deep convolutional neural networks", *Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, & J. Dean, "Distributed representations of words and phrases and their compositionality", *Advances in Neural Information Processing Systems (NIPS)*, pp. 3111-3119, 2013

[16] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee & A. Y. Ng, "Multimodal deep learning", *International Conference on Machine Learning (ICML)*, Bellevue, USA, 2011.

[17] Y. Peng, W. Zhu, Y. Zhao et al., "Cross-media analysis and reasoning: advances and directions", *Frontiers Inf Technol Electronic Eng*, vol. 18, pp. 44-57, 2017.

[18] Y. Peng, J. Qi, Y. Yuan, "CM-GANs: Cross-Modal Generative Adversarial Networks for Common Representation Learning", *ACM Trans. Multimedia Comput. Commun. Appl.*, Association for Computing Machinery, vol. 15, no. 1, pp. 1551-6857, 2019.

[19] A. Radford, L. Metz & S. Chintala, " Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", *CoRR*, 2015.

[20] V. S Ramachandran & E. M Hubbard, "Psychophysical investigations into the neural basisof synaesthesia" *Proceedings of the Royal Society Biological Sciences Series B*, vol. 268, no. 1470, pp. 979-983, 2001.

[21] E. M. Hubbard, D. Brang & V. S. Ramachandran, "The cross-activation theory at 10", *Journal of neuropsychology*, vol. 5, no. 2, pp. 152–177, 2011.

[22] E. M. Hubbard, V. S. Ramachandran, G. M. Boynton, "Cortical cross-activation as the locus of grapheme-color synesthesia", *Journal of Vision*, vol. 3, no. 9, pp. 621-621a, 2003.

[23] A. B. Safran & N. Sanda, "Color synesthesia. Insight into perception, emotion, and consciousness", *Current opinion in neurology*, vol. 28, no. 1, pp. 36–44, 2015.

[24] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford & X. Chen, "Improved techniques for training GANs", *Proc. Adv. Neural Inf. Process. Syst.*, pp. 2234-2242, 2016.

[25] K. Simonyan & A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *International Conference on Learning Representations (ICLR)*, 2014.

[26] R. Socher, A. Karpathy, Q. Le et al., "Grounded compositional semantics for finding and describing images with sentences", *Trans. Assoc. Comput. Ling*, vol. 2, pp. 207–218, 2014.

[27] P. -N. Tan, M. Steinbach & V. Kumar, "Introduction to Data Mining", *Addison-Wesley*, Ch. 8, pp. 500, 2005.

[28] J. Ward, "Synesthesia", *Annual Review of Psychology*, vol. 64], pp. 49-75, 2013.

[29] B. Wang, Y. Yang, X. Xu, A. Hanjalic & H. T. Shen, "Adversarial Cross-Modal Retrieval", *In Proceedings of the 25th ACM international conference on Multimedia (MM '17)*, pp. 154-162, 2017.

[30] X. Wang, D. Peng, P. Hu & Y. Sang, "Adversarial correlated autoencoder for unsupervised multi-view representation learning", *Knowledge-Based Systems*, Vol. 168, pp. 109-120, 2019.

[31] Y. Yamaguchi, K. Noda, S. Nishide, H. G. Okuno & T. Ogata,"Learning and association of synaesthesia phenomenon using deep neural networks," *Proceedings of the 2013 IEEE/SICE International Symposium on System Integration*, Kobe, pp. 659-664, 2013.