In: Proceesings of the International Joint Conference on Neural Networks (IJCNN/WCCI), pp. 5314-5321, 2018

Expectation Learning and Crossmodal Modulation with a Deep Adversarial Network

Pablo Barros[1], German I. Parisi[1], Di Fu[2,3], Xun Liu[2,3], and Stefan Wermter[1]
[1]Knowledge Technology, Department of Informatics, University of Hamburg, Hamburg, Germany
[2]CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, China
[3]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

***Abstract*—The human brain is able to learn, generalize, and predict crossmodal stimuli which help us to understand the world around us. Some characteristics of crossmodal learning inspired some computational models but most of the solutions only go as far as to implement strategies for early or late crossmodal fusion. In this paper, we propose the use of two mechanisms from behavioral psychology to enhance the capabilities of a deep adversarial network to learn crossmodal stimuli: the unity assumption modulation and expectation learning. We use real-world data to train and evaluate our model in a set of experiments and demonstrate how these mechanisms affect the learning behavior of the model and how they contribute to making it learn crossmodal coincident stimuli. Our experiments show that the addition of these two mechanisms modulates the crossmodal binding capabilities of the model and improves the learning of unisensory descriptors.**

I. Introduction

Crossmodal processing is a crucial feature of the human brain which is necessary for understanding the world around us. The meaningful processing of crossmodal information allows us to enhance our perceptual experience [1] also for unisensory stimuli [2], to solve associative incongruence and conflicts [3], and to learn new concepts [4].

Computational models for crossmodal learning have been proposed in the past to enhance tasks such as classification, regression, and prediction. Most of these models propose solutions for crossmodal fusion at an early [5] or late stage [6], [7], e.g., by using crossmodal representations to increase the level of abstraction for a perception task. However, these models typically rely on individual and independent mechanisms for processing unimodal representations where modalities do not influence each other [8], [9]. The development of computational models that use brain-inspired principles for crossmodal processing may lead to more robust perception and interaction mechanisms in complex crossmodal environments.

Neurophysiological findings show that different brain regions are activated and are communicating with each other also when processing unisensory information [10]. This has been explained as different modulations on perception, based on crossmodal integration. One of these modulations is known as the *unity assumption* [11], [12], [9], which is a behavioral psychology term that describes the modulation caused by knowing that different unisensory modalities belong to the same event. Different from causal inference [13], the concept of unity assumption is not to identify whether different stimuli are produced by the same source, but to assess how this assumption affects crossmodal integration [14]. Different factors lead to the unity assumption, among them the presence of crossmodal correspondences [15] (e.g. associating a small dog with a high-pitch barking) and semantic congruence [16] (learning to associate barks with dogs and meows with cats).

In addition to the unity assumption, the brain also fine-tunes information using what is known as the *expectation effect* [17]. While we are looking at an object, we are also estimating thousands of comparisons and clustering with similar and different objects that we have seen before. An even stronger effect of learning by expectation occurs with crossmodal information. When looking for a dog, one already expects to hear a bark [18]. This also causes an overfitting behavior when we have experienced very few examples in our lives: if we grew up near an opera house and never heard any other music style, every time we see a live show we would expect the singer to sing an opera. This is an important effect of learning as once we realize that there is an incongruence between what we expect and what really occurs, e.g. when the singer suddenly starts to sing rock music and not an opera, we learn a novel association [19]. Such a learning process, referred to as learning by expectation, makes us experts in associating concepts in an unsupervised way and using the difference of what was expected and what was perceived as a modulatory effect for learning new concepts.

In this paper, we introduce the use of a deep neural architecture for crossmodal associative learning which encodes some aspects of the *unity assumption* to enhance the learning and segregation of crossmodal stimuli. We also propose the use of *expectation learning* to make the model learn crossmodal representations from unimodal stimuli and to be adaptable to learning novel concepts. Our model is based on an adversarial autoencoder architecture, and thus, is able to learn in an unsupervised manner, and to encode and generate unimodal and crossmodal stimuli.

The proposed model has two auto-encoding channels, one to process each specific modality, the channels are composed of a series of convolution layers able to learn a high-level encoded representation of the data and to reconstruct the input stimuli. Additionally, we use a generative channel for each modality which is used to create modality-specific stimuli and to train the auto-encoder architecture in an adversarial way. Different from common adversarial autoencoders, we use a

self-organizing layer to learn how to associate the encoded representation of coincident modalities.

With such an architecture, we are able to model important modulations for the unity assumption, such as crossmodal correspondence and semantic congruence and to generate high-level stimuli based on the expected information. We evaluate our model using real-world data from the AudioSet database collection [20] and the Imagenet large-scale visual recognition corpora [21]. These corpora contains richly annotated audios and image, respectively, separated into different contextual classes. We use four of these classes: "Meowing", "Barking", "Oinking" and "Cooing", representing respectively cats, dogs, pigs, and pigeons. Our model learns from real-world data how to associate audio and visual information from these animals.

We evaluate our model using two scenarios: first, the model is trained to evaluate the unity assumption. We present the model with a series of congruent and incongruent stimuli. We then evaluate how the unity assumption modulation behaves over time. In our second experiment, we evaluate the capability of the model to learn crossmodal representations from unisensory perception using the proposed expectation learning.

II. Crossmodal Stimuli Association Model

There are two components in the adversarial learning strategy: a generator and a discriminator. The generator (G) is trained to generate information which is as real as possible from a uniform distribution z, and the discriminator (D) is trained to distinguish between generated ($G(z)$) and real information (x). The idea of adversarial learning is that we train these two components in a competitive way with a final objective to make the generator produce information which is indistinguishable from the real information. The Generative Adversarial Network (GAN) objective is to use two neural networks: a discriminator and a generator. In recent years, the use of deep neural networks in adversarial learning tasks became successful on images and audio generation, however, most of the solutions suffer from instability during training and/or generation of low-fidelity information. One of the proposed solutions which achieved good performance on high-level vision tasks was the Boundary Equilibrium Generative Adversarial Network (BEGAN) [22], which we use as the basis for our work.

Different from traditional GANs, the BEGAN uses an autoencoder instead of a discriminator structure. In a conventional generator/discriminator architecture, the model minimizes the Kullback-Leibler divergence between a real data distribution, (Px), and the generated data distribution, (Pg). One of the common problems is that the discriminator starts to send meaningless gradients to the generator too quickly, meaning that the discriminator trains faster than the generator is capable to produce real-like data. The BEGAN tries to solve this problem by removing the discriminator of the architecture and replacing it by an autoencoder and introducing an equilibrium term. Instead of using the loss of the discriminator to train the generator, the BEGAN uses the Wasserstein distance between the reconstruction loss of the real data and the generated data, thus not relying on a discriminator loss. By doing that, the model assumes that by matching the distribution of the reconstruction losses it can match the data distribution. That means that giving the real data x and the noise distribution z, the BEGAN loss can be calculated as:

$$\begin{aligned} L_D &= L(x) - K_t L(\dot{G}(z)), \\ L_G &= L(G(z)), \\ K_{t+1} &= K_t + \lambda L(\dot{G}(z))(\gamma . L(x) - L(G(z))), \end{aligned} \tag{1}$$

where L_D and L_G are respective losses for the discriminator (the autoencoder) and the generator. Here, L_G is only used to update the parameters of the generator and L_D is only used to update the parameters of the discriminator. The losses of the reconstruction of the real data ($L(x)$) and generated data ($L(G(z))$) are calculated via the autoencoder. The adaptive term, K_t, is used to balance automatically the training and it avoids that the generated reconstruction loss (which at first is higher) has a stronger impact while training the autoencoder. The learning rate λ controls the update of the adaptive term. Finally, the diversity ratio, γ, is important to ensure that the generator does not dominate the autoencoder. It is important for the training of the model that the autoencoder learns to act as a discriminator, that means, to produce a higher reconstruction loss when generated data is presented. This way, the strategy is to choose the diversity ratio always within an interval between 0 and 1. Higher values of γ lead to a higher focus on the autoencoding ($L(x)$), forcing the generator to produce higher-fidelity data but reducing diversity, and lower values of γ makes the generator have more influence on the adaptive term ($L(G(z))$), resulting in an increase of diversity but sacrificing data quality.

One important aspect of the BEGAN is a convergence measure (M), which is used to determine if the model is converging or collapsing. This measure can be calculated as

$$M = L(x) + |\gamma L(\dot{x}) - L(G(z))|, \tag{2}$$

and will approach 0 when the network is converging, meaning that the reconstruction error of the real data is similar to the reconstruction error of the generated data discounted by the balanced term.

A. Crossmodal Architecture

Our proposed architecture is based on the BEGAN, but with two modality-specific channels: one learning visual features and the other one learning auditory characteristics. Each of these channels has an autoencoder structure and a generator structure. The autoencoder structure is composed of an encoding architecture which generates an encoded representation (ZA) and a decoding architecture which, from the encoded representation, reconstructs the information. The generator architecture has the same topological architecture as the decoder, however, does not share the same weights, and uses as input its own encoded representation: ZG_v for the visual generator and ZG_a for the auditory one.

Our visual encoding structure is composed of four layers, each one implementing two subsequent convolution operations. We use a 128x128x3 image as input (X_v). The first layer has 16 filters, the second has 32 filters, the third has 48 and the fourth 64 filters. All of the convolutions implement filters with size (3x3), elu activation functions and the last convolution operation of each layer has a stride of (2x2). After the last layer, a dense layer with 64 units and a linear activation are used to encode the visual representation (ZA_v).

The decoding layer receives as input the encoded representation (with a dimension of 64 units) and has the same topology as the encoding but in the reverse order: the first convolution layers have 64 filters and the last ones have 16. The decoder uses a stride of 1 in each convolution operation, and after each layer, it applies an upsampling operation with a factor of (2,2). The last layer of the decoder is a single convolution layer with 3 filters which outputs an image with the same shape as the encoding input: 128x128x3.

The auditory channel receives as input the spectral representation of the audio. We calculate the FFT with a Hamming window of 1024 and a stride of 512. We use 3s of audio as input, which produces a spectrogram with size (520,96). We pre-process the audio by applying a pre-emphasis filter with a coefficient of 0.95. Our auditory encoding structure is also composed of 4 layers, with 2 convolution operations in each one. The first layer has 16 filters, the second has 32 filters, the third has 48 filters and the fourth has 64 filters, all of them with dimensions of (3x3) and implementing elu activation functions. All of them have a stride of (2x2). The last layer is a fully connected layer with 64 units, representing the encoded auditory representation (ZA_a). The auditory decoding structure is the same as the encoding but reversed, and with strides of (1x1) in all convolutions. After each layer, we use an upsampling operator with a factor of (2x2).

We then propose a self-organizing layer which is connected to the encoded representations (ZA_v and ZA_a). When presenting coincident stimuli the network, the neurons of the self-organizing layer encode crossmodal stimuli. This layer has a topology of 30x30 neurons, and each neuron has a dimension of 128 units representing audio (64 encoded units) and visual encodings (64 encoded units). This self-organizing layer is trained only when the convergence measure M of both modalities is smaller than a certain threshold. That avoids that this layer learns to encode meaningless representations in the beginning of the adversarial training. By observing the behavior of our network during our experiments, we empirically choose this threshold to be 0.06. Figure 1 illustrates the autoencoder structure of our model and the self-organizing layer.

III. Unity Assumption

The unity assumption can be defined as the belief that two or more unisensory stimuli belong together [14]. Over the past years, many researchers measured the effect of the unity assumption as a modulator for crossmodal integration [11], [23], [24]. They found evidence that the unity assumption is part of how the brain solves crossmodal binding problems. When assuming that different stimuli have the same event as the source, the subjects displayed a faster association learning mechanism when compared, which provided a more accurate crossmodal association behavior.

The unity assumption is correlated with some factors, among them semantic congruence and crossmodal correspondence. In diverse experiments, semantic congruence shows to be important for the unity assumption, as explained in the review of Chen and Spence [9]. They show that in different psychological experiments (spatial and temporal ventriloquist effect and McGurk effect), the fact that the two stimuli have a semantic congruence (a picture of a dog and a barking sound, for example), lead the participants to assume that they were produced by the same object. The same occurred when there was crossmodal correspondence, for example, a small dog, and a higher-pitch barking. While using crossmodal correspondence and semantic congruence to assume unity, the participants ended up creating a prior bias for new perceived stimuli, and this ended up improving their performance on recognizing and categorizing congruent stimuli.

Based on these findings, we introduce here the concept of unity assumption (UA) in our model. The unity assumption is calculated as a modulator for the autoencoder and generator losses and will help the model to penalize incongruent associations during training. That means that the model will learn how to generate and reconstruct unisensory information based on the crossmodal association. Also, it enables the model to identify the congruent and incongruent stimulus.

Given an encoded visual (ZA_v) and auditory (ZA_a) representation, we calculate the best matching unit (BMU) that is closer to each specific modality (WV as the visual BMU and WA as the auditory one) using only the modality part of the encoded representation. That means that each BMU has a concatenated representation of audio and visual modalities, represented by WV_v and WV_a for the visual BMU and WA_v and WA_a for the auditory one. We then calculate the crossmodal correspondence (CC):

$$CC_m = e^{-(|ZA_m - W_m|)^2}, \tag{3}$$

where m represents the modality ($m = v$ for visual and $m = a$ for auditory) and W represents the specific BMU (WV for visual and WA for auditory). The visual crossmodal correspondence (CC_v) measures the similarity of the perceived vision stimulus (WA_v) and the vision representation associated with the crossmodal encoding of the perceived auditory stimulus (WA_v). This means that if the perceived vision stimulus has features which would represent a small dog, it should be associated with a high-pitch sound as learned by coincidence during training and it is represented by (WA_v). Same occurs for the auditory crossmodal correspondence. The crossmodal correspondence will be higher if the Euclidean distance between the stimuli is closer, meaning that the perceived dog had indeed a high-pitch barking. Given that we do not introduce any prior bias in the model, the crossmodal correspondence

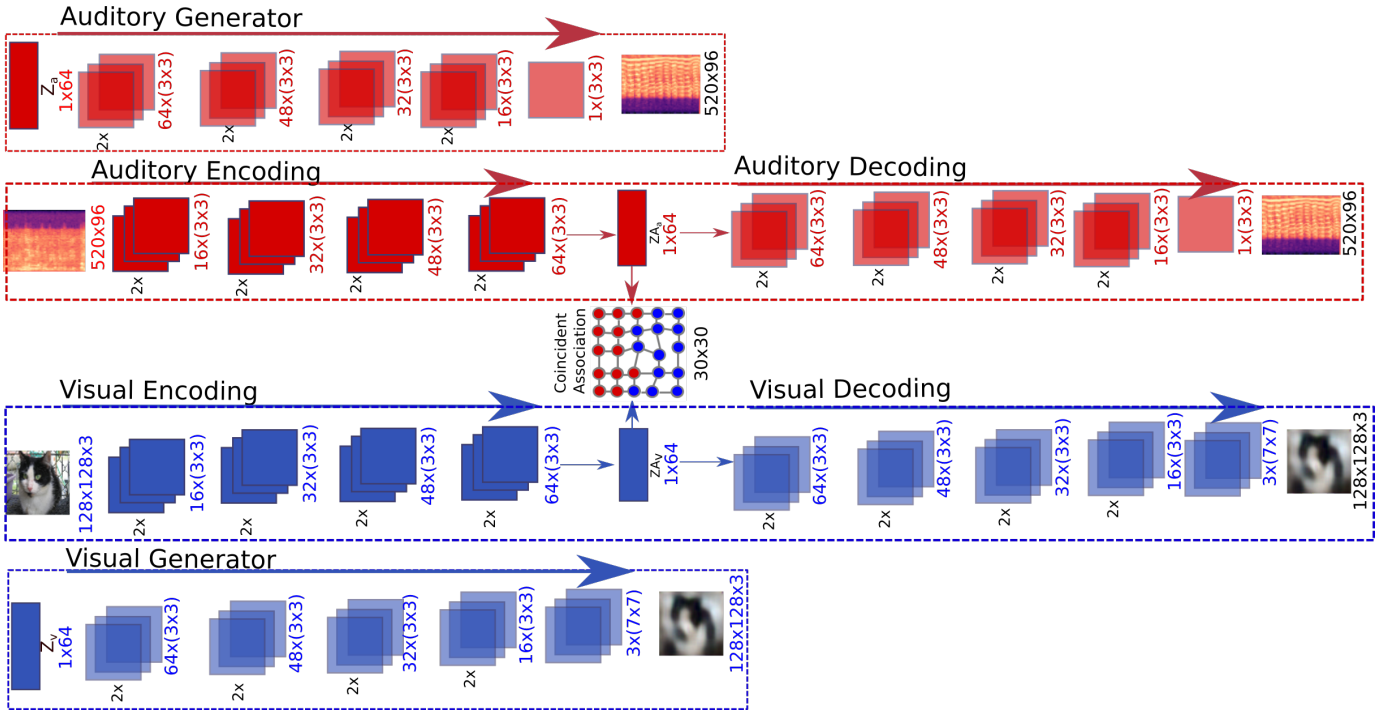


Fig. 1: The encoding, decoding and generator structures of our model. During the adversarial training, each unimodal autoencoder acts as a discriminator, while a specific generator for each modality learns how to generate high-fidelity information. The coincident association layer encodes two unimodal stimuli, which happen at the same time, as a crossmodal encoding.

is highly sensitive to coincident stimuli and thus can only be measured accurately after the network has learned to associate crossmodal stimuli.

The semantic congruence (*SC*) gives us a measure if the perceived stimulus is congruent or incongruent and is calculated as:

$$SC = e^{-(|Wv - Wa|)^2}. \tag{4}$$

The *SC* is calculated based on the topological location of the encoded crossmodal representation. Each unimodal stimulus will have an associated representation (*WV* for visual and *WA* for auditory), and the *SC* gives us the similarity of the stimulus, 1 for fully congruent and 0 for fully incongruent. Different from the *CC*, the *SC* uses the crossmodal information to identify semantic congruence, so if the vision represents a dog and the audition perceives a cat, the *SC* will indicate highly incongruent stimuli.

The unity assumption (*UA*) is calculated taking into consideration the crossmodal correspondence and the semantic congruence:

$$UA = e^{-\frac{(|CC_v - CC_a|)}{SC^2}}, \tag{5}$$

where the value of *UA* is between the interval of 0, for assumed unrelated stimuli, and 1 for full unity assumption. The *UA* gives the model an indication if the unisensory stimuli are feature and semantic level correlated. If the distances crossmodal correspondence is strong, meaning a smaller distance between the specific crossmodal correspondences, the unity assumption will be less influenced by the semantic congruence, however, if the crossmodal correspondence is weak, the unity assumption will depend mostly on the semantic congruence. The unity assumption is then used to modulate the learning of the unisensory descriptors. To not unbalance the learning between the generator and the autoencoder, we proceed to add the *UA* in both losses:

$$\begin{aligned} L_D &= L(x) - k_t L(\dot{G(z)}) + UA\dot{t}, \\ L_G &= L(G(z)) + UA.t, \end{aligned} \tag{6}$$

where t represents the impact rate of the *UA* and it is chosen as a value between 0 and 1. By choosing a small t, the impact of the unity assumption on the unisensory learning is small and thus the learned representation is based on each individual modality. When t has a higher value, the unity assumption has as a major role to correlate the learned representations within each individual modality. That means that, when assuming a higher impact rate, the model will be strongly penalized when incongruent stimuli are presented.

The unity assumption works based on two factors: crossmodal correspondence and semantic congruence. Both factors can only be obtained if a prior-knowledge on the basic unisensory description is presented, i.e., to associate a dog and a bark, the model has to know how to describe a dog and a bark. Thus, we only use the unity assumption modulation when each unisensory descriptor of the model is trained properly. During our experiments, we empirically choose to activate

the unity assumption modulation only when both unisensory channels had a convergence measure, M, of 0.06.

IV. Expectation Learning

The expectation effect [17] is one of the important mechanisms of the brain to fine-tune crossmodal bindings. While perceiving an event in one modality, our brain is automatically reconstructing information about this event in different modalities. So, when you hear a siren, you can associate this with an ambulance being able even to visualize it. This mechanism allows us to expect, when perceiving a unisensory stimulus, how it will be described with other modalities.

The expectation effect helps to fine-tune our unisensory description capabilities. We can update our visual descriptors based on audio signals and vice-versa. That means that, once I learned that a dog will bar,k when I see a wolf, I would expect it to produce a sound similar to barking, and not to a bird singing. When the wolf starts to howl, I can easily adapt to the new sound as it is closer to the one I expected. This also helps to learn new crossmodal associations: when I see a hyena in a zoo, I expect it to produce a sound close to barking. When it produces a completely different sound, I perceive that there is a strong incongruence and a new crossmodal association is learned [19].

Giving our model the capability to learn by expectation, we can make use of visual information to fine-tune the auditory information and vice-versa. To make it possible, we first train our model using the generator/autoencoder structure and the adversarial learning strategy, so we guarantee that the model has enough prior-knowledge to describe high-level concepts. Then, we start with the expectation training: instead of using the encoded representations Z_v and Z_a as input for the decoding structure, we take the BMU for each individual modality, NV for vision and NA for audio. That means that instead of training the model based on the reconstruction of the real image, we train the model based on the difference between the real image and a prior-concept encoded in the coincident association layer. Figure 2 illustrates the expectation learning mechanism.

The expectation learning can be applied in two learning scenarios: to learn crossmodal associations based on unisensory perception and to learn new crossmodal associations. In the first scenario, we send an unisensory stimulus to the model, and we use the coincident association layer to reconstruct the associated crossmodal representation. That means that by seeing a novel dog species, it will expect it to bark, and thus, associate it with barking sounds. In the second scenario, by presenting often novel crossmodal stimuli to the model, the coincident association layer will update the model's own crossmodal associations. By updating the coincident association layer in an online manner, the model can adapt to novel information whenever it is presented.

V. Experimental Methodology

To evaluate our model we use a combination of two datasets: the AudioSet collection [20] and the ImageNet large-scale

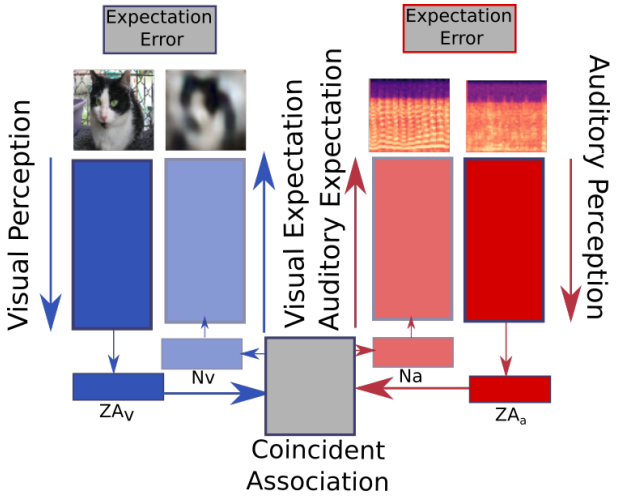


Fig. 2: Expectation learning strategy. We find the coincident association (WA for audio and WV for vision) for each stimulus and use it to create the expected stimulus. We then use the expectation error between the reconstructed signal and the perceived stimulus to train the network.

visual recognition collection [21]. The AudioSet collection contains 2 millions of human-labeled 10-seconds sound clips from Youtube videos separated into more than 600 classes, in total more than 5.8 thousand hours of audio. As the sound clips come from Youtube videos, they have a large variance on recording features: loudness, quality, background noise, and/or event duration among others. This makes the data from the AudioSet the closest we can get to real-world scenarios. For our experiments, we choose audios from four categories: "Meowing", "Barking", "Oinking" and "Cooing", representing cats, dogs, pigs, and pigeons respectively. There are around two thousand sound clips for each of these classes, each one with 10 seconds of duration.

To create the visual part of the crossmodal stimuli, we use the ImageNet collection to obtain images from the four categories we selected. For each category, we obtained around 1500 images with bounding boxes. Similar to AudioSet, these images are human-annotated and obtained from Youtube videos and Google images search. We then associate each image with an audio, creating multimodal corpora. As for some categories, there are more sound clips than images, we repeat some of the images while creating the audio-visual pairs. That means that for each 10s of audio, we have one associated image.

To evaluate our model, we proceed with three experiments: the baseline, the unity assumption, and the expectation learning. We first classify each of the categories using our model to serve as baseline and ground truth experiment for our experiments. For that, we train our generator/autoencoder structure for each modality individually and, after training, we add an extra classifier layer connected to each encoded representation (ZA_v for vision and ZA_a for audio). We then train only the classifier using a 4-fold cross-validation strategy

TABLE I: Mean accuracy, in percentage, and standard deviation of our baseline experiment: 4-fold crossvalidation with all the data for each category.

Modality	Cat	Dog	Pig	Pigeon
Half of the training data				
Audio	74.3(2.5)	72.4(1.5)	71.1(1.3)	72.2(2.2)
Vision	85.7(1.8)	81.2(1.7)	83.4(1.8)	84.1(2.3)
Entire training data				
Audio	83.5(2.2)	87.3(1.7)	80.2(1.9)	90.4(2.4)
Vision	91.2(1.5)	93.5(1.6)	94.2(1.2)	93.7(2.7)

and measure the mean accuracy of the model for each category. We perform this experiment with the entire training data and with half of the training data, which serves as a comparison for the next experiments.

To evaluate the unity assumption, we first train our model with half of the data in order to create prior crossmodal bindings and ensure that the descriptors are able to categorize the stimuli properly. We then train the network with the unity assumption modulation in two scenarios: in the first one we send to the network congruent stimuli, that means paired audio-visual stimuli are congruent, dog images and barks, and in the second we send to the network incongruent stimuli, and for that we shuffle the audio-visual pairs in a way that dogs images are never associated with barking sounds. We then calculate the crossmodal correspondence, semantic congruence, and unity assumption over 50 epochs. We also calculate the accuracy of the model for each scenario using the same evaluation strategy as our baseline.

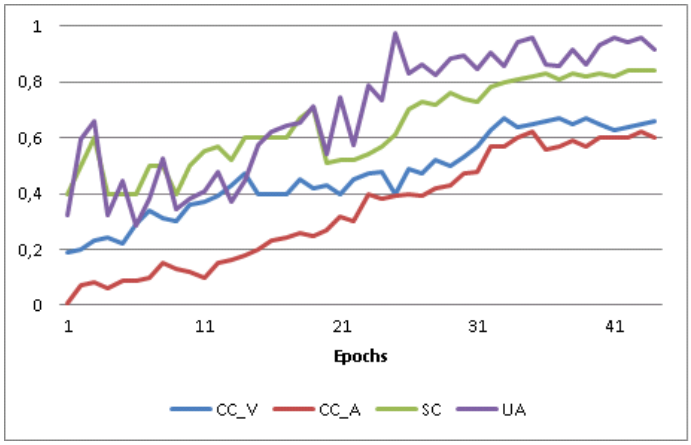
Finally, to evaluate the expectation learning, we repeat the same strategy as before, i.e. pre-training the model using half of the data to obtain a strong prior crossmodal association. We then evaluate the model in two steps: in the first one, we use the expectation learning to train the model with the remaining data and show some expected stimuli reconstructed by the model. In the second step, we train the model using only audio and only vision stimuli, evaluate the unity assumption behavior and calculate the accuracy of the model.

A. Baseline Recognition Results

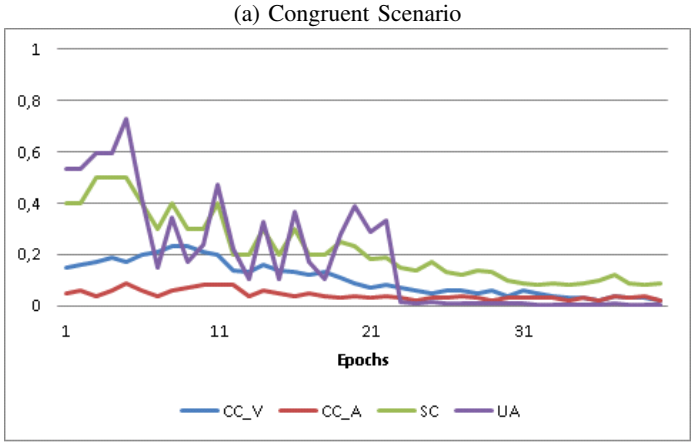
Table I exhibits the mean classification accuracy of each category for the audio and vision modalities when we train the model with half of the training data and with the entire training data. We see that the model can recognize each animal with a similar accuracy; however, it is having a better performance when using vision. This could be caused by the fact that the network learned better visual descriptors than auditory ones, which is consistent with state-of-the-art research in generative adversarial networks. Recent research shows that, to learn robust auditory representations, adversarial networks need much more data than for visual tasks [25]. With more training data, the network could learn better descriptors and it performs better when the entire training data is present.

B. Unity Assumption Results

The plot in Figure 3a illustrates the analysis of the model when trained with congruent stimuli. Over time, the unity assumption (*UA*) becomes higher indicating that the model is learning a strong crossmodal binding. The crossmodal correlations (CC_V for visual and CC_a for auditory) and semantic congruence (*SC*) also increase over time, indicating that the model is learning how to correlate the stimulus on these two different levels. The unity assumption (*UA*) is mostly affected by the semantic congruence in the beginning of the training, when the crossmodal correlations are higher and closer to each other, however, over time the development of crossmodal correspondence enhances the unity assumption.



(a) Congruent Scenario



(b) Incongruent Scenario

Fig. 3: Illustration of the model's crossmodal congruence (CC_v for visual, CC_a for auditory), semantic congruence (*SC*) and unity assumption (*UA*) when trained with the unity assumption modulation in two scenarios: with congruent and incongruent stimuli.

Training the model with incongruent stimuli produced the behavior illustrated in Figure 3b. The unity assumption has a decreasing trend until it finally goes to zero after a number of epochs. This happens when the crossmodal correlations reach a very low level.

We also calculate the accuracy of the model, exhibited in Table II, following the same strategy as our baseline, after training the model with the unity assumption in our two sce-

TABLE II: Mean accuracy, in percentage, and standard deviation of the model trained with unity assumption modulation in both scenarios: with congruent and incongruent stimuli.

Modality	Cat	Dog	Pig	Pigeon
		Congruent		
Audio	86.2(2.2)	91.2(2.8)	81.2(1.2)	94.3(2.1)
Vision	94.7(1.6)	96.3(1.2)	94.5(1.4)	96.7(1.3)
		Incongruent		
Audio	83.3(1.7)	86.7(1.2)	84.3(1.6)	91.4(2.7)
Vision	90.6(1.3)	94.2(1.7)	95.3(1.4)	92.9(2.0)

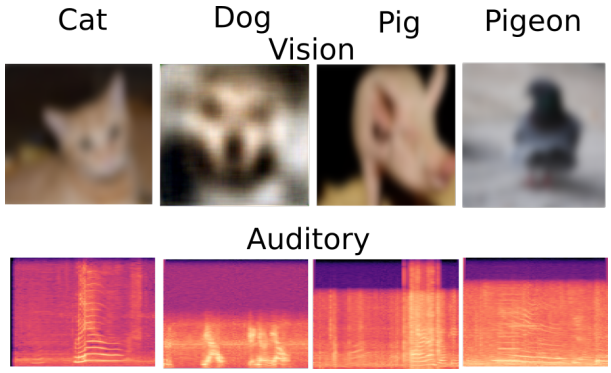


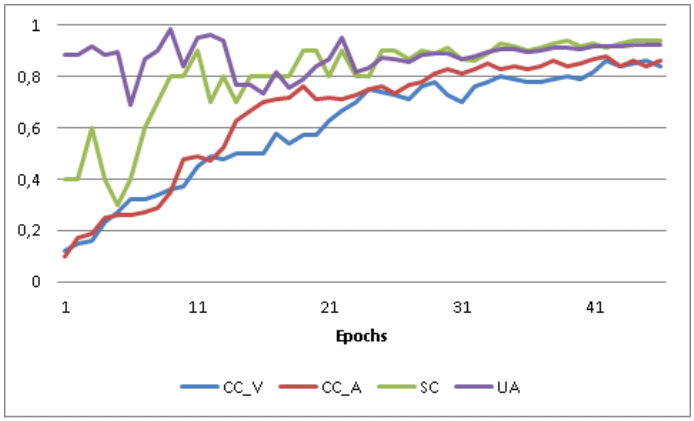
Fig. 4: Expected reconstructions per modality. By sending a sound to the network, it reconstructs the associated (above) image and vice-versa (below).

narios. It is possible to see a slight improvement of the results when the congruent stimuli are present, indicating that the unity assumption has an effect on the learned representations when congruent stimuli are used to train the network. When incongruent stimuli are used, the network presents a behavior which is similar to the baseline, indicating that although the network does not infer a strong unity assumption, it still can categorize individual unisensory stimuli.

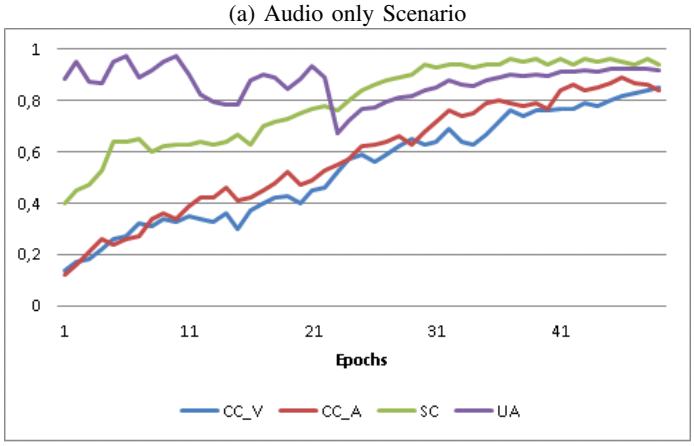
C. Expectation Learning Results

When training the model with expectation learning, it learns how to reconstruct expected stimuli. Figure 4 illustrates examples of reconstructions when the network, after having been trained, hears sounds or sees images from each of the four animals. As it is possible to see the reconstructions, although not having 100% fidelity with real data, they have general characteristics which resemble each of the animals.

When training the network with only unisensory stimuli, it produces a stronger unity assumption, as described in Figures 5a and 5b. It is possible to see that the crossmodal correspondences, semantic congruence, and unity assumption increase at a higher rate when the network is trained with unisensory stimuli when compared to when trained with crossmodal stimuli. This probably happens because, when using the coincident association layer to create an expected reconstruction, the network tends to limit the expected representation to a concept that it already knows, and thus, it is producing a closer crossmodal correspondence and semantic congruence. It is possible to see a very different behavior when training the



(a) Audio only Scenario



(b) Vision only Scenario

Fig. 5: Illustration of the model's crossmodal congruence (CC_a for visual, CC_a for auditory), semantic congruence (SC) and unity assumption (UA) when trained with the unity assumption modulation and the expectation learning in two scenarios: only using audio and only using video stimuli.

network with audio only or vision only: the network trained with audio only has a higher increasing trend, which could indicate that the auditory features learned by the network are not diverse enough, which makes it then associate known features with new concepts, and thus, present a stronger unity assumption.

When evaluating the performance of the model when only unisensory stimuli are used, as described in the results of Table III, it is possible to see that, when training with vision only, the vision accuracy behaves somehow similar to when training with both modalities. The same occurs with audio only. However, it is important to note that the performance of the network on the absent modality is only slightly lower than when trained with both modalities. This shows that the network can learn the absent modality representations with a similar performance as when they are present.

TABLE III: Mean accuracy, in percentage, and standard deviation of the model trained with expectation learning in two scenarios: only with auditory stimuli and only with visual ones.

Modality	Cat	Dog	Pig	Pigeon
Vision Stimuli				
Audio	83.1(1.4)	89.7(1.4)	83.1(1.1)	93.2(1.7)
Vision	95.1(1.3)	95.2(1.2)	94.3(1.2)	96.1(1.4)
Auditory Stimuli				
Audio	85.7(1.3)	92.3(1.6)	83.4(1.7)	95.2(2.2)
Vision	91.3(1.2)	92.1(1.4)	93.2(1.7)	90.3(2.5)

VI. Discussions, Conclusion and Future Work

Our model was evaluated with the implementation of two novel mechanisms which are known to participate in crossmodal binding in humans: the unity assumption modulation and the expectation learning. After our experiments we can see that these two mechanisms change the model's behavior on binding crossmodal concepts in different ways: while the unity assumption provides the model with a bottom-up modulation that affects the learning of unisensory descriptors, the expectation learning gives the model the capability to learn top-down crossmodal concepts, and use this information to fine-tune unisensory pathways.

The unity assumption modulation was shown to be capable, after a prior learning with coincident stimuli, to identify incongruent stimuli based on two factors: crossmodal correspondence and semantic congruence. These factors are known to be part of the unity assumption in humans [9], and in our model they were able to modulate the behavior of the unity assumption. These factors contribute in different complementary ways to the unity assumption: the semantic congruence gives a general idea of unity assumption by estimating if the two stimuli are conceptually correlated, and the crossmodal correspondence modulates this assumption by identifying if the stimuli are congruent on a description level.

Our expectation learning experiments show that our model can learn crossmodal descriptors using unisensory information. Also, we show that the model is able to correlate conceptual information from different modalities: the expected reconstructions contain enough information that even a human can identify the animal by looking at it. The model itself can be improved for reconstructing audio information and we believe that this could be solved by training the network with more auditory information.

Our model does not take into consideration the temporal aspects of crossmodal learning, and developing the model further into this direction could make it adaptable to focus on one particular stimulus over time, filtering out background noise for example. The use of recurrent connections on the self-organizing layer could help to achieve this.

Acknowledgements

This work was partially supported by German Research Foundation (DFG) under project CML (TRR 169) and the NSFC (61621136008) and the China Scholarship Council.

References

[1] P. E. Patton and T. J. Anastasio, "Modeling cross-modal enhancement and modality-specific suppression in multisensory neurons," *Neural Computation*, vol. 15, no. 4, pp. 783–810, 2003.

[2] F. Frassinetti, N. Bolognini, and E. Làdavas, "Enhancement of visual perception by crossmodal visuo-auditory interaction," *Experimental Brain Research*, vol. 147, no. 3, pp. 332–343, 2002.

[3] A. O. Diaconescu, C. Alain, and A. R. McIntosh, "The co-occurrence of multisensory facilitation and cross-modal conflict in the human brain," *Journal of Neurophysiology*, vol. 106, no. 6, pp. 2896–2909, 2011.

[4] K. Dorst and N. Cross, "Creativity in the design process: co-evolution of problem–solution," *Design Studies*, vol. 22, no. 5, pp. 425–437, 2001.

[5] S. Wei, Y. Zhao, Z. Zhu, and N. Liu, "Multimodal fusion for video search reranking," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 8, pp. 1191–1199, 2010.

[6] J.-C. Liu, C.-Y. Chiang, and S. Chen, "Image-based plant recognition by fusion of multimodal information," in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2016 10th International Conference on*. IEEE, 2016, pp. 5–11.

[7] M. H. de Boer, K. Schutte, H. Zhang, Y.-J. Lu, C.-W. Ngo, and W. Kraaij, "Blind late fusion in multimedia event retrieval," *International Journal of Multimedia Information Retrieval*, vol. 5, no. 4, pp. 203–217, 2016.

[8] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[9] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4405–4425, 2017.

[10] C. Kayser and L. Shams, "Multisensory causal inference in the brain," *PLoS biology*, vol. 13, no. 2, pp. 100 – 125, 2015.

[11] R. B. Welch and D. H. Warren, "Immediate perceptual response to intersensory discrepancy." *Psychological bulletin*, vol. 88, no. 3, p. 638, 1980.

[12] C. Spence, "Audiovisual multisensory integration," *Acoustical science and technology*, vol. 28, no. 2, pp. 61–70, 2007.

[13] B. R. Orvis, J. D. Cunningham, and H. H. Kelley, "A closer examination of causal inference: The roles of consensus, distinctiveness, and consistency information." *Journal of personality and social psychology*, vol. 32, no. 4, p. 605, 1975.

[14] A. Vatakis and C. Spence, "Crossmodal binding: Evaluating the unity assumption using audiovisual speech stimuli," *Attention, Perception, & Psychophysics*, vol. 69, no. 5, pp. 744–756, 2007.

[15] C. V. Parise and C. Spence, "When birds of a feather flock together: synesthetic correspondences modulate audiovisual integration in non-synesthetes," *PLoS One*, vol. 4, no. 5, pp. 56– 68, 2009.

[16] O. Doehrmann and M. J. Naumer, "Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration," *Brain research*, vol. 1242, pp. 136–150, 2008.

[17] H. Yanagisawa, "Expectation effect theory and its modeling," in *Emotional Engineering*, vol. 4.

[18] R. J. Tomlin, S. V. Stevenage, and S. Hammond, "Putting the pieces together: Revealing face–voice integration through the facial overshadowing effect," *Visual Cognition*, pp. 1–15, 2016.

[19] F. G. Ashby and L. E. Vucovich, "The role of feedback contingency in perceptual category learning." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 42, no. 11, p. 1731, 2016.

[20] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE ICASSP*, 2017.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[22] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

[23] M. Schutz and M. Kubovy, "The effect of tone envelope on sensory integration: support for the'unity assumption'," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3412–3425, 2008.

[24] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas, "Recognizing expressions from face and body gesture by temporal normalized motion and appear-

ance features," *Image and Vision Computing*, vol. 31, no. 2, pp. 175 – 185, 2013, affect Analysis In Continuous Input.

[25] Y. Chiba, A. Ito, and T. Shinozaki, "Voice conversion from arbitrary speakers based on deep neural networks with adversarial learning," in *Advances in Intelligent Information Hiding and Multimedia Signal Processing: Proceedings of the Thirteenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, August, 12-15, 2017, Matsue, Shimane, Japan*, vol. 82. Springer, 2017, p. 97.