UNSUPERVISED EXPECTATION LEARNING FOR MULTISENSORY BINDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Expectation learning is a continuous learning process which uses known multisensory bindings to modulate unisensory perception. When perceiving an event, we have an expectation on what we should see or hear which affects our unisensory perception. Expectation learning is known to enhance the unisensory perception of previously known multisensory events. In this work, we present a novel hybrid deep recurrent model based on audio/visual autoencoders, for unimodal stimulus representation and reconstruction, and a recurrent self-organizing network for multisensory binding of the representations. The model adapts concepts of expectation learning to enhance the unisensory representation based on the learned bindings. We demonstrate that the proposed model is capable of reconstructing signals from one modality by processing input of another modality for 43,500 Youtube videos in the animal subset of the AudioSet corpus. Our experiments also show that when using expectation learning, the proposed model presents state-of-the-art performance in representing and classifying unisensory stimuli.

1 INTRODUCTION

Multisensory binding is one of the most important processes that humans use to understand their environment. By using different sensory mechanisms, we are able to collect and process distinct information streams from the same experience which leads to a highly complex association learning. This mechanism allows us to improve the perception of individual stimuli (Frassinetti et al. (2002)), solve contextual, spatial and temporal conflicts (Diaconescu et al. (2011)), and progressively acquire and integrate novel information (Dorst & Cross (2001)).

An important process of multisensory binding is known as the expectation effect (Yanagisawa (2016)). When perceiving an event, we compare it to other events we have experienced before, and make certain assumptions based on our experience. For instance, when seeing a cat, we expect it to meow and not to bark. This effect modulates our multisensory association in terms of top-down expectation. When a cat barks to us, we assume that our perception is inconsistent. This means that either unisensory perception failed, the spatial or temporal attention was misleading, or we are learning a new species of a barking cat. For each of these scenarios, our brain adapts to the situation and we update our multisensory knowledge. This learning process referred to learning by expectation (Ashby & Vucovich (2016)), strongly suggests the role of unsupervised learning for multisensory binding, and leads to an adaptive mechanism for learning of novel concepts (Chen et al. (2017)).

Most existing computational models for multisensory learning apply explicit weighted connections while integrating the sensor information which are learned using early (Wei et al. (2010)) or late (Liu et al. (2016); de Boer et al. (2016)) fusion techniques. These weighted connections are usually tuned in a data-driven way whereby the data distribution affects the multisensory binding directly. This makes these models suitable to be used for real-world and complex problems but they are very sensitive to the training data used when performing the multisensory integration. More accurate models apply the neurophysiology findings on unisensory biasing for multisensory computational models (Pouget et al. (2002); Rowland et al. (2007); Kayser & Shams (2015)). Such models, although similar to the brain's neural behavior, are usually not feasible to be used on real-world data, as they are mostly applied to simple stimuli scenarios, and do not scale well. There exist other complex models that implement attention mechanisms based on multisensory information, but the most recent focus in this area is on data-driven fusion models (Hori et al. (2017); Barros et al. (2017);

Mortimer & Elliott (2017)). The introduction of expectation learning would give these models the ability to adapt better to novel situations and learn with its own errors in a online and continuous way.

Recent contributions make use of data-driven learning for multisensory representations Arandjelovic & Zisserman (2017); Senocak et al. (2018); Owens & Efros (2018); Kim et al. (2018). Such solutions make use of different transfer learning and attention mechanisms to improve unisensory recognition and localization. Although they present impressive result in this specific tasks, they still rely on strongly labelled data points or are not suitable for online learning given that they have a extensive training proceeding. In particular, the work by Arandjelovic and Zisserman (Arandjelović & Zisserman (2017)) introduces a data-driven model for multisensory binding with bottom-up modulation for spatial attention. Their model uses the network's activity to spatially identify which part of an image a certain sound is related to. Although the model is data-driven, the authors claim that it learns real-world biasing on a multisensory description for unisensory retrieval by using a large amount of real-world training data. Their results show that the model can also use a number of unisensory channels to compensate absent ones and identify congruent and incongruent stimuli. Overall, this approach has produced impressive results on spatial location and cross-sensory retrieval.

A similar approach was presented by Zhou et al. (Zhou et al. (2017)), but focused on audio generation. Their model relies on a sequence-to-sequence generator to associate audio events to visual information. The same generator is use to generate audio for newly presented video scenes. This requires an external teacher to identify congruent and incongruent stimuli which makes it impossible to be used in online learning scenarios. Also, all of them are deeply dependent on an end-to-end deep learning strategy. Furthermore, it cannot learn novel information without the need of extensive retraining the entire model again.

In this work, we propose a computational model based on expectation learning for multisensory binding. Our model implements a multichannel autoencoder to learn audio-visual representations. It also introduces the use of temporal binding to learn expectation by means of a recurrent Grow-When-Required (GWR) neural network. The capability to grow and shrink when necessary allows us to model a continuous learning scenario of multisensory binding. As the recurrent GWR network learns prototypes of multisensory binding, we use such prototypes to reconstruct unisensory stimuli. This allows us to reconstruct auditory information from visual stimuli and vice-versa.

Our hybrid approach allowed us to have a strong and robust encoding/decoding mechanism for real-world data based on our autoencoders and a temporal multisensory binding solution. The recurrent GWR performed particularly well on creating prototypical concepts which are important for our expectation learning mechanism, as we do not want to learn multisensory bindings of instances but of higher abstraction concepts. Our model is trained fully unsupervised, different from most of the current state-of-the-art solutions for multisensory binding. This allows us to learn concepts in an online manner, without the constraint of having a strong supervised training process.

Our model develops the temporal aspects of multisensory binding to continuously reconstruct the perceived stimuli. When perceiving the sound of a cat, the model can reconstruct the image of a cat, while when a dog enters a scene, the sound of the dog will be reconstructed. Also, by clustering the multisensory bindings on the GWR, we can create expected stimuli based on the learned concepts. This means that our model is able to reconstruct auditory and visual stimuli from prototypical concepts such as cats, dogs, and horses in the absence of sensory input. We hypothesize that our model can use the reconstruction of multimodal stimuli based on multisensory binding to boost the representation of unisensory stimuli.

To the best of our knowledge, there is no standard audio-visual binding association benchmark, and thus we evaluate our model using the animal sub-sets of the AudioSet corpus (Gemmeke et al. (2017)). This corpus contains human-labeled samples of Youtube videos based on the audio information. We use a total of 24 different animal classes and more than 44k samples to train the model. To proper measure the models capability to learn different abstraction levels of multisensory binding, we make use of overlapping animal classes. We measure our models behavior in two steps: first, we evaluate the performance of the model to recognize multimodal stimuli when one of the modalities is absent. Second, we analyze the learning behavior of the model and correlate it to multisensory imagery (Spence & Deroy (2013)) effect, where humans simplify absent stimuli into concepts to enhance perception.

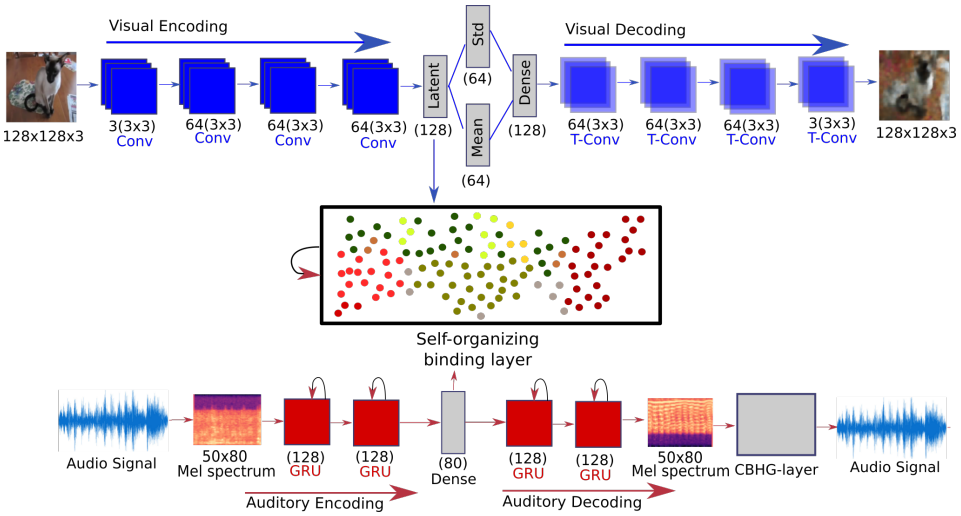

Figure 1: An overview of the proposed model with the audio/visual autoencoder structures and the self-organizing binding layer.

2 MULTISENSORY TEMPORAL BINDING

To reconstruct auditory and visual stimuli, we developed neural networks based on autoencoders for each of the unisensory channels. These networks encode high-dimensional data into a latent representation and reconstruct real-world audio-visual information. The binding between auditory and visual information is realized by means of a recurrent GWR network. The GWR is a self-organizing network that learns to create conceptual prototypes of a data distributions in an unsupervised, incremental learning manner. To address the temporal aspects of coincident binding, we extend the Gamma-GWR (Parisi & Wermter (2017)) which endows prototype neurons with a number of temporal contexts to learn the spatiotemporal structure of the data distribution. An overview of our model is illustrated in Figure 1

2.1 VISUAL CHANNEL

To process high-level information in the visual channel, we developed a variational autoencoder (Chen et al. (2016)). Which, instead of using a standard bottleneck layer, performs sampling over Gaussian unit reconstructions of individual nodes, which makes it robust to overfitting and which allows it to learn more general representations than a standard autoencoder.

To train the autoencoder, we implemented a composite loss function based on the reconstruction error and the Kullback-Leibler (KL) divergence between the encoded representation and a sample from the Gaussian distribution. This composite loss function is important to learn representations of general concepts, represented by Gaussian distributions, instead of reconstructing input images from memorized parameters.

Our model receives as input a 128x128 color image. The input data is processed by our encoding architecture which is composed of a series of four convolution layers, with a stride of 2x2, and kernel sizes of dimension 3x3. The first convolution layer has three channels and the subsequent three layers have 64 filters. The latent representation starts with a fully connected layer with 128 units. We compute the standard deviation and mean of this layer's output, generate a Gaussian distribution from it and sample an input for another fully connected hidden layer with 64 units, which is our final latent representation. The decoding layer has the same structure as our encoding layer but in the opposite direction and applying transpose convolutions.

2.2 AUDITORY CHANNEL

For the auditory channel, we implement a recurrent autoencoder based on Gated Recurrent Units (GRU) (Cho et al. (2014)). Recurrent units allow us to reconstruct audio with better quality than using non-recurrent layers since auditory signals are sequential, and each audio frame depends highly on previous contextual information.

As input and output of the auditory autoencoder, we compute a Mel spectrum which we generate from the raw waveform. To reconstruct the audio from the output Mel spectrum, we develop a Convolutional Bottleneck CBHG-layer (Lee et al. (2017)) which consists of a 1-D convolutional bank, a highway network and a bi-directional GRU layer. This network receives as input the Mel spectrum, and outputs a linear frequency spectrum which is then transformed into waveform using the Griffin Lim algorithm. This approach of transforming Mel coefficients to waveform achieved better audio synthesis quality than performing Griffin Lim on the Mel spectrum directly (Wang et al. (2017)), which is important for our expectation learning.

We performed hyperparameter optimization for the autoencoder and found that an audio spectrum window length of 50ms, a window shift of 12.5 ms with 80 Mel coefficients and 1000 linear frequencies yield best results. We also found that 80 units for the dense bottleneck layer and two GRU layers with 128 units each for both the encoder and decoder network are sufficient for achieving a high audio quality. An additional number of Mel coefficients, GRU layers, and neural units did not significantly improve the reconstruction quality. The number of bottleneck units is important for the multisensory binding as it determines the number of connections between the binding layer and the audio encoder and decoder.

2.3 SELF-ORGANIZING TEMPORAL BINDING

To learn coincident bindings between audio and visual stimuli, we make use of an unsupervised binding layer. This layer is implemented as a recurrent GWR network which is an unsupervised model for learning spatiotemporal prototype representations. The layer receives as input the latent representation of our visual and auditory channels which occur coincidentally. To synchronize the two data streams, we make sure that the videos have a temporal resolution of 20 frames per second. As a result, each frame of our video is associated with 12.5 ms of auditory information. In contrast to traditional self-organizing models with winner-take-all dynamics for the processing of spatial patterns, the Gamma-GWR (Parisi & Wermter (2017)) computes the winner neuron taking into account the activity of the network for the current input and a temporal context. Each neuron of the map consists of a weight vector $\mathbf{w}_j$ and a number K of context descriptors $\mathbf{c}_j^k$ (with $\mathbf{w}_j, \mathbf{c}_j^k \in \mathbb{R}^n$). As a result, recurrent neurons in the map will encode prototype sequence-selective snapshots of the input. Given a set of N neurons, the best-matching unit (BMU), $\mathbf{w}_b$, with respect to the input $\mathbf{x}(t) \in \mathbb{R}^n$ is computed as:

$$b = \arg\min_{j \in N} \left(\alpha_0 \|\mathbf{x}(t) - \mathbf{w}_j\|^2 + \sum_{k=1}^{K} \alpha_k \|\mathbf{C}_k(t) - \mathbf{c}_{j,k}\|^2 \right), \tag{1}$$

$$\mathbf{C}_k(t) = \beta \cdot \mathbf{w}_{I-1} + (1 - \beta) \cdot \mathbf{c}_{I-1,k-1}, \tag{2}$$

where α_i and $\beta \in (0; 1)$ are constant values that modulate the influence of the current input with respect to previous neural activity, $\mathbf{w}_{I-1}$ is the weight of the winner neuron at $t-1$, and $\mathbf{C}_k \in \mathbb{R}^n$ is the global context of the network ($\mathbf{C}_k(t_0) = 0$).

New connections are created between the BMU and the second-BMU of an input. When a BMU is computed, all the neurons the BMU is connected to are referred to as its topological neighbors. Given the current input $\mathbf{x}(t)$, the activity of the network $a(t)$ is defined by the activity of the BMU as $a(t) = \exp(d_b)$ where d_b is defined by Eq. 6. Each neuron is equipped with a habituation counter $h_i \in [0, 1]$ expressing how frequently it has fired based on a simplified model of how the efficacy of a habituating synapse reduces over time. In the Gamma-GWR, the habituation rule is given by $\Delta h_i = \tau_i \cdot \kappa \cdot (1 - h_i) - \tau_i$, where κ and τ_i are constants that control the decreasing behavior of the habituation counter Marsland et al. (2002) To establish whether a neuron is habituated, its habituation counter h_i must be smaller than a given habituation threshold h_T. The network is initialized with two neurons and, at each learning iteration, it inserts a new neuron whenever the activity of the network $a(t)$ of a

habituated neuron is smaller than a given threshold a_T, i.e., a new neuron r is created if $a(t) < a_T$ and $h_b < h_T$. The training of the neurons is carried out by adapting the BMU b and its topological neurons n according to:

$$\Delta \mathbf{w}_i = \epsilon_i \cdot h_i \cdot (\mathbf{x}(t) - \mathbf{w}_i), \tag{3}$$

$$\Delta \mathbf{c}_{k,i} = \epsilon_i \cdot h_i \cdot (\mathbf{C}_k(t) - \mathbf{c}_{k,i}), \tag{4}$$

where ϵ_i is a constant learning rate. The learning process of the Gamma-GWR is unsupervised and driven by bottom-up sensory observations, thereby either allocating new neurons or adapting existing ones in response to novel input. In this way, fine-grained multisensory representations can be acquired and fine-tuned through experience.

As an extension of the Gamma-GWR, we implement temporal synapses for the purpose of predicting future frames from an onset frame. We implement temporal connections as sequence-selective synaptic links that are incremented between those two neurons that are consecutively activated. When the two neurons i and j are activated at time $t-1$ and t respectively, their synaptic link $P_{(i,j)}$ is strengthened. Thus, at each learning iteration, we set $\Delta P_{(I-1,b)} = 1$, where $I-1$ and b are respectively the indexes of the BMUs at time $t-1$ and t. As a result, for each neuron $i \in N$, we can retrieve the next neuron v of a prototype sequence by selecting

$$v = \arg\max_{j \in N \setminus i} P_{(i,j)}. \tag{5}$$

This approach results in the learning of trajectories of neural activations that can be reconstructed in the absence of sensory input.

3 Expectation Learning

As the self-organizing layer is updated in an unsupervised Hebbian manner, it learns to associate audio-visual stimuli online. That means that the binding process is entirely data-driven, without the necessity of supervision. More specifically, after finding the BMU related to a unimodal perceived stimulus, the associated absent stimuli will be reconstructed based on the prototypical concept that this neuron learned. This is possible because each neuron in the self-organizing layer processes the union of the auditory and visual encodings at training time, where both signals are provided.

This reconstruction and expectation learning capability is the basis for our following proposal of a re-training mechanism for the self-organizing layer. We first pre-train our self-organizing binding to generate prototype neurons with strong audio-visual encodings. This allows the model to learn a prior association between auditory and visual concepts. After the network has learned these associations, we use unknown data to fine-tune the bindings with the expectation learning.

We first encode a visual or auditory stimulus (s), and compute the BMU (b_{av}) using only the associated auditory or visual weights:

$$b_{av} = \arg\min_{j \in N} \left(\alpha_0 \|\mathbf{s}(t) - \widetilde{\mathbf{w}}_j^s\|^2 + \sum_{k=1}^{K} \alpha_k \|\widetilde{\mathbf{C}}_k^s(t) - \widetilde{\mathbf{c}}_{j,k}\|^2 \right), \tag{6}$$

where $\widetilde{\mathbf{w}}_{j,s}$ represents the audio or visual representation encoded on the neuron's weights. In this case, the global context of the network at any time step ($\widetilde{C}_{s,k}t$) is represented by the stimulus encoding, the same happens with the BMU context ($\widetilde{\mathbf{c}}_{j,k}$). We then use the auditory and vision parts of the multisensory representation stored on b_{av} to reconstruct the auditory (a') and visual (v') information using the specific channel decoding D_v for vision and D_a for audio:

$$\begin{aligned} a' &= D_a(b_a), \\ s' &= D_v(b_v). \end{aligned} \tag{7}$$

By doing this for both perceived auditory and visual signals, we create two extra pairs of multisensory stimuli by combining the perceived auditory and visual ones with the reconstructed auditory and visual. We bind the encoded information of the reconstructed audio-visual information with the original perceived stimuli and re-train the self-organizing layer with the new pairs. By pairing

the perceived and the reconstructed stimuli representations, we enforce the self-organizing layer to learn concepts and not instances of the animals. That means that animals which sound similar will be paired together. This mechanism helps us enforce the connections of coincident stimuli without needing a large amount of data. In the other hand, incongruence will make the model pair different audio-visual stimuli which will create prototype neurons which will be forgotten quickly by the self-organizing layer as they do not happen often.

4 EXPERIMENTAL SETUP

Our goal is to evaluate the performance of the model to reconstruct audio/visual stimuli based on unimodal perception. Also, we intend to evaluate the conceptual relations learned by the network. Thus, although there exists several datasets with multimodal information, the animal subset of the AudioSet corpus (Gemmeke et al. (2017)) presents a unique advantage for our evaluation. It contains natural scenarios with different levels of conceptual binding - from images of cats and meowing sounds association to high pitch barking with small dogs. Each video has a duration of 10s and it is possible that, e.g., there is both a cat and a dog present in the video. As there are no standard published results on this specific task for the AudioSet corpus, we run a series of baseline recognition experiments that serves as main comparison to measure our model's performance. To obtain a precise measure of the contribution of the expectation learning, we decided to cluster some overlapping classes and use 16 single labels, one per video: Cats ("Cat" + "Meow"+ "Purr"), Dogs ("Bark"+"Dog"+"Howl"), Pigs ("Oink" + "Pig"), Cows ("Moo"+"Cattle, bovinae"), Owls ("Owl" + "Coo"), Birds, Goats, Bee ("Bee, was, etc.."), Chickens ("Chicken, rooster"), Ducks ("Duck"), Pidgeons ("Pidgeon, dove"), Crows ("Crow"), Horses ("Horse"), Frogs ("Frogs"), Flies ("Fly, housefly"), Lions ("Roaring cats (lions, tigers)"). We used the unbalanced training subset consisting of approximately 43.500 videos to train our model and evaluated it with the test subset consisting of approximately 20.000 videos. The labels of this dataset were crowdsourced based on the video descriptions.

In our first baseline experiment, we train classic deep learning architectures for individual stimuli recognition. We make use of the Inception V3 network (Ioffe & Szegedy (2015)) for the visual stimuli, and the SoundNet (Aytar et al. (2016)) for the auditory stimuli. Our second baseline experiment makes use of the learned representations of the individual autoencoers. We implement a supervised classifier for each channel. The classifier is composed of a dense layer with 128 units and an output softmax layer. We optimized our autoencoders and classifier to maximize the recognition accuracy on the training subset using a tree-structured Parzen estimator (TPE) (Bergstra et al. (2011)) and use the optimal parameters through all of our experiments. This is an important step, as we need the encoding/decoding processes of our model to work robustly with real-world data.

To evaluate our expectation model in a proper manner, we first trained our network using half of the training subset in order to guarantee that the model learned prior meaningful audio-visual bindings. Then we use the other half of the training subset to update the networks bindings using the expectation learning. We measure the performance of the model to generate meaningful information. For that, we present one audio or visual stimulus of a video and generate the absent stimulus. We then e feed the autoencoder with the generated stimulus and calculate the accuracy of the generated stimulus regarding the true label of the video. To measure the contribution of the expectation learning we also computed the recognition accuracy for each class of the evaluation subset when training only with half of the training subset. This is important to explicit the impact of the expectation learning.

For all of our experiments We trained the models 10 times and report the mean accuracy and standard deviation for each modality. We used the same 10% of the training subset as a validation set for each experiment. We used the accuracy convergence on the validation subset to stop the training of each model. This guarantees that each model is trained for an optimized number of epochs to maximize the performance.

5 RESULTS

Our final results are shown on Table 1. The contribution of the expectation is highlighted by a comparison with the results of training the autoencoder with half the training data. Training the model with half of the data, to create the strong binding associations, was enough to obtain a baseline

Table 1: Mean accuracy, in percentage, and standard deviation of our experiments: the baselines results, training the network with half and all the samples of the training set and reconstructing the absent modality using the expectation mechanism.

Model	Audio	Vision
Inception V3 (Ioffe & Szegedy (2015))	-	90.4 (1.3)
SoundNet (Aytar et al. (2016))	68.5 (2.4)	-
Autoencoder (Half)	58.5 (3.1)	69.0 (3.9)
Without Expectation	66.4 (2.4)	86.8 (3.2)
With Expectation Learning	70.8 (3.2)	89.8 (1.9)

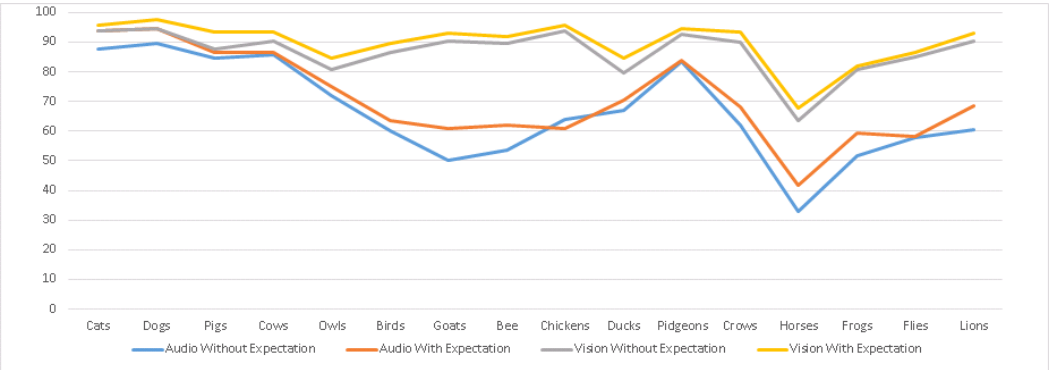

Figure 2: Mean accuracy per class, in percentage, of the reconstructed absent stimuli. We compare audio and visual reconstruction with the results when training the network with all the samples of the training set.

performance. Training the model with all the data points improved 8% of recognition accuracy for audio and more than 17% for audio. While training with the expectation learning, we obtained an improvement of more than 12% for audio recognition and more than 20% for vision.

The performance of the network follows the general behavior of other models to recognize vision stimuli better than auditory stimuli. This effect is demonstrated by the results of the SoundNet and Inception-V3 models. This is probably due the dataset presenting challenging audio stimuli with much background noise. When compared to SoundNet and Inception-V3, our expectation model presents better audiotry recognition, and competitive vision recognition performance. When compared to training the autoencoder with all the datapoints the expectation learning improved the recognition of auditory stimuli by more than 4% while improved the visual stimuli in 3%. This behavior is a side effect of the expectation model which enforces the multimodal binding. The auditory stimuli is more affected, as it presents much more noisy information. The network then relies more on the visual stimuli and creates neurons with strong visual encoding. This effect is represented by creating neurons with similar visual encoding associated to different auditory encoding. When training with expectation learning, the network created an average of 5400 neurons, while training without the expectation it created 4000 neurons.

The latent representation from the auditory and visual channels encode different characteristics of the stimulus, which are then bound by our self-organizing layer. The expectation learning enforced the generation of robust bindings, especially for distinct animals. The network ended up creating specific neurons for cats and dogs, and shared neurons for chickens and ducks, for example. This explains the improvement on the recognition of the reconstructed stimuli of easily separable animals, as illustrated in Figure 2.

6 Discussion

As the self-organizing layer is updated in an unsupervised manner, it learns to associate audio-visual stimuli online. That means that our model learning mechanism is entirely data-driven without

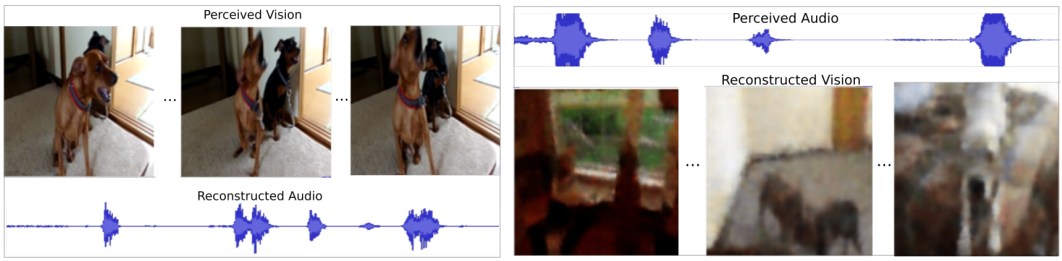

Figure 3: Example of the reconstruction output. The left image displays the audio reconstruction when the visual stimulus is perceived. The right image displays the vision reconstruction when the audio stimulus is perceived.

a strong teaching supervisor. More specifically, after finding the BMU related to a single perceived stimulus, the associated absent stimuli can be reconstructed based on the concept that this neuron learned, and not on the original high-abstraction data. When sending an image of a dog to the network, it will reconstruct the sound of barking, but not the exact sound that this specific dog would make.

This effect is related to the concept of multisensory imagery (Spence & Deroy (2013)) where humans tend to simplify and cluster the absent stimuli when asked to reconstruct it. That means that every time you see a different small yellow bird, you will expect it to sound very similar to the ones you have seen before. This is an important effect that helps our model to reconstruct concepts instead of specific instances.

As there is no quantitative metric to measure this effect, thus we proceed with a simple overlapping analysis which gives us an indication of how well the model is binding and clustering audio-visual information. We first train the model with expectation learning and then we classify every single neuron using both audio and visual classifiers which generate two labels for each neuron: one for auditory and one for visual information. The total overlap between visual and auditory labels for each prototype neuron in our self-organizing layer is 93%, suggesting that our prototype neurons are very concise when storing audio-visual information.

The training of the binding layers causes the prototype neurons to be associated to high-abstraction concepts and to reconstruct absent modality-specific information which is the basis for our expectation learning mechanism. For example, when humans see a barking dog, they are continuously reinforcing the association between dogs and barking sounds. We also create associations between small dogs and high-pitch barking, an example of an effect called multisensory correspondence Spence & Driver (2000).

This effect was also observed in some examples where the variety of animals was higher, such as the dogs. We illustrate one of these examples in Figure 3. It is possible to see that when reconstructing the visual information based on an auditory stimulus of different dogs barking, a high-pitch dog barking sound generated images related to a small dog. It is also possible to see that, when the sound of more than one dog appears barking, the network generated an image of more than one dog.

7 Conclusion

Multisensory binding is a crucial aspect of how humans understand the world. Consequently, the development of computational systems able to adapt this aspect into information processing is important to many research fields. An extensive number of models has been proposed that incorporate different aspects of multisensory binding. However, none of them made use of expectation learning which is a mechanism that helps humans learn from their mistakes in a continuous manner. In this paper, we introduce a model that adapts different concepts of expectation learning. Our model makes use of autoencoders to be able to reconstruct audio-visual stimuli and use a recurrent self-organizing network to learn multisensory bindings. We show that our expectation learning mechanism provides a robust solution for reconstructing absent, modality-specific stimuli and discussed the effects of such mechanism on the learning behavior of our model. Our experiments demonstrate that the pro-

posed model and expectation learning mechanisms can improve unisensory recognition. We would like to extend our model to deal with spatial expectation, which would give us a complete view on learning based on contextual, temporal, and spatial incongruence. We also observed the development of continuous and incremental learning in our self-organizing binding layer. We would like to explore these characteristics with further experiments on life-long and transfer learning.

ACKNOWLEDGMENTS

Given the double-blind review nature of ICLR, we will only update the Acknowledgments session after the reviewing process.

REFERENCES

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 609–617. IEEE, 2017.

Relja Arandjelović and Andrew Zisserman. Objects that sound. *arXiv preprint arXiv:1712.06651*, 2017.

F Gregory Ashby and Lauren E Vucovich. The role of feedback contingency in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(11):1731, 2016.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pp. 892–900, 2016.

Pablo Barros, German I Parisi, Cornelius Weber, and Stefan Wermter. Emotion-modulated attention improves expression recognition: A deep learning model. *Neurocomputing*, 253:104–114, 2017.

James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*, pp. 2546–2554, 2011.

Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. A survey of depth and inertial sensor fusion for human action recognition. *Multimedia Tools and Applications*, 76(3):4405–4425, 2017.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014. ISSN 09205691. doi: 10.3115/v1/D14-1179. URL `http://arxiv.org/abs/1406.1078`.

Maaike HT de Boer, Klamer Schutte, Hao Zhang, Yi-Jie Lu, Chong-Wah Ngo, and Wessel Kraaij. Blind late fusion in multimedia event retrieval. *International Journal of Multimedia Information Retrieval*, 5(4):203–217, 2016.

Andreea Oliviana Diaconescu, Claude Alain, and Anthony Randal McIntosh. The co-occurrence of multisensory facilitation and cross-modal conflict in the human brain. *Journal of Neurophysiology*, 106(6):2896–2909, 2011.

Kees Dorst and Nigel Cross. Creativity in the design process: co-evolution of problem–solution. *Design Studies*, 22(5):425–437, 2001.

Francesca Frassinetti, Nadia Bolognini, and Elisabetta Làdavas. Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research*, 147(3):332–343, 2002.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE ICASSP*, 2017.

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 4203–4212. IEEE, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Christoph Kayser and Ladan Shams. Multisensory causal inference in the brain. *PLoS biology*, 13 (2):100 – 125, 2015.

Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. On learning associations of faces and voices. *arXiv preprint arXiv:1805.05553*, 2018.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017. URL `http://aclweb.org/anthology/Q17-1026`.

Jen-Chang Liu, Chieh-Yu Chiang, and Sam 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*, pp. 5–11. IEEE, 2016.

Stephen Marsland, Jonathan Shapiro, and Ulrich Nehmzow. A self-organising network that grows when required. *Neural Networks*, 15(8):1041–1058, 2002.

Bruce JP Mortimer and Linda R Elliott. Information transfer within human robot teams: Multimodal attention management in human-robot interaction. In *Cognitive and Computational Aspects of Situation Management (CogSIMA), 2017 IEEE Conference on*, pp. 1–3. IEEE, 2017.

Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. *arXiv preprint arXiv:1804.03641*, 2018.

German I Parisi and Stefan Wermter. Lifelong Learning of Action Representations with Deep Neural Self-Organization. In *AAAI Spring Symposium*, pp. 608 – 612, 2017.

Alexandre Pouget, Sophie Deneve, and Jean-René Duhamel. A computational perspective on the neural basis of multisensory spatial representations. *Nature Reviews Neuroscience*, 3(9):741, 2002.

Benjamin A Rowland, Terrence R Stanford, and Barry E Stein. A model of the neural mechanisms underlying multisensory integration in the superior colliculus. *Perception*, 36(10):1431–1443, 2007.

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4358–4366, 2018.

Charles Spence and Ophelia Deroy. Crossmodal mental imagery. In *Multisensory imagery*, pp. 157–183. Springer, 2013.

Charles Spence and Jon Driver. Attracting attention to the illusory location of a sound: reflexive crossmodal orienting and ventriloquism. *Neuroreport*, 11(9):2057–2061, 2000.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards End-to-End Speech Synthesis. Technical report, Google, Inc., 2017. URL `http://arxiv.org/abs/1703.10135`.

Shikui Wei, Yao Zhao, Zhenfeng Zhu, and Nan Liu. Multimodal fusion for video search reranking. *IEEE Transactions on Knowledge and Data Engineering*, 22(8):1191–1199, 2010.

Hideyoshi Yanagisawa. Expectation effect theory and its modeling. 4.

Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. *arXiv preprint arXiv:1712.01393*, 2017.

8 APPENDIX

8.1 TRAINING PARAMETERS

8.1.1 VISION CHANNEL

Table 2 exhibits all the important parameters used to train our visual channel. We used ADAM optimizer with an adaptive learning rate.

Table 2: Training parameters of the Vision channel

Parameter	Value
Epochs	200
Batch size	32
Optimizer	ADAM (Adaptive Learning rate)
Initial learning rate	0.05
ADAM beta1	0.9
ADAM beta2	0.999

8.1.2 AUDITORY CHANNEL

Table 3 exhibits all the important parameters used to train our auditory channel. We follow the same training procedure as the vision channel, and also use ADAM optimizer with an adaptive learning rate.

Table 3: Training parameters of the Auditory channel

Parameter	Value
Epochs	250
Batch size	32
Optimizer	ADAM
Initial learning rate	ADAM (Adaptive Learning rate)
ADAM beta1	0.9
ADAM beta2	0.999

8.1.3 SELF-ORGANIZING BINDING LAYER

Table 4 exhibits all the important parameters used to train our gamma Growing-When-Required (GWR) network. We use a small insertion threshold, which helps the network to maintain a limited number of neurons, reinforcing the generation of high-abstract clusters.

Table 4: Training parameters of the Auditory channel

Parameter	Value
Epochs	50
Insertion threshold	0.01
Context size	4
Initial Gamma Weights	0.64391426, 0.23688282, 0.08714432, 0.0320586
β_b	0.5
ϵ_b	0.2
ϵ_n	0.003

8.2 DETAILED EXPERIMENTAL RESULTS

8.2.1 BASELINE EXPERIMENTS

Table 6 details the mean accuracy of our baseline experiments (SoundNet and Inception V3) for each of the classes of the evaluation subset.

Table 5: Detailed results in accuracy (in %) and standard deviation for our baseline experiments.

Animal	SoundNet	Inception V3
Cats	90,2 (3.2)	94,8 (2.4)
Dogs	92,5 (4.1)	96,7 (2.5)
Pigs	80,7 (3.7)	95,6 (3.4)
Cows	83,8 (3.5)	94,8 (1.7)
Owls	71,8 (1.4)	87,8 (1.0)
Birds	62,7 (2.2)	90,6 (3.6)
Goats	60,2 (3.9)	95,8 (2.1)
Bee	63,1 (1.1)	91,2 (4.7)
Chickens	59,8 (3.0)	96,8 (2.3)
Ducks	68,7 (4.1)	85,1 (1.7)
Pidgeons	76,8 (2.6)	92,5 (3.1)
Crows	67,9 (1.8)	91,3 (2.7)
Horses	43,6 (3.7)	69,8 (4.1)
Frogs	57,8 (1.4)	79,8 (2.5)
Flies	53,1 (1.3)	89,8 (1.9)
Lions	63,5 (3.4)	94,5 (2.5)

8.3 Expectation learning Experiments

Table 5 exhibits the detailed accuracy per class when evaluating our expectation learning model with the evaluation subset. Here we detail the results based on reconstructed audio, when the audio is absent on the perceived stimuli, and on reconstructed vision, when the vision is absent. We detail the experiments with and without the expectation.

Table 6: Detailed accuracy (in %) for our baseline experiments.

Animal Class	Audio		Vision	
-	Without Expectation	With Expectation	Without Expectation	With Expectation
Cats	87,6 (3.2)	93,8 (2.1)	93,8 (1.9)	95,6 (2.1)
Dogs	89,5 (3.6)	94,4 (2.9)	94,6 (2.2)	97,5 (1.8)
Pigs	84,6 (3.2)	86,5 (3.7)	87,5 (1.4)	93,4 (1.7)
Cows	85,9 (4.1)	86,7 (2.7)	90,4 (1.6)	93,4 (2.8)
Owls	71,8 (3.7)	74,9 (2.9)	80,7 (1.8)	84,7 (1.9)
Birds	60,1 (2.6)	63,7 (1.9)	86,7 (4.7)	89,7 (3.7)
Goats	50,2 (1.6)	60,7 (3.7)	90,4 (2.8)	93,2 (1.9)
Bee	53,7 (2.7)	62,1 (3.9)	89,5 (2.7)	91,7 (3.1)
Chickens	63,8 (1.9)	60,7 (2.1)	93,8 (1.7)	95,7 (1.9)
Ducks	66,9 (1.9)	70,5 (2.8)	79,5 (1.6)	84,6 (2.9)
Pidgeons	83,6 (4.7)	83,8 (2.6)	92,6 (2.7)	94,7 (2.9)
Crows	62,1 (1.9)	68,3 (2.2)	90,1 (2.0)	93,4 (2.8)
Horses	32,8 (2.6)	41,6 (3.9)	63,7 (3.1)	67,8 (1.8)
Frogs	51,8 (3.7)	59,4 (2.7)	80,6 (2.7)	82,1 (3.4)
Flies	57,8 (3.0)	58,3 (2.5)	84,9 (1.6)	86,7 (2.6)
Lions	60,3 (2.9)	68,5 (2.6)	90,4 (2.4)	93,2 (3.8)