Zero-Shot Audio Classification using Image Embeddings

Duygu Dogan*, Huang Xie*, Toni Heittola*, Tuomas Virtanen*
*Unit of Computing Sciences, Tampere University, Finland

***Abstract*—Supervised learning methods can solve the given problem in the presence of a large set of labeled data. However, the acquisition of a dataset covering all the target classes typically requires manual labeling which is expensive and time-consuming. Zero-shot learning models are capable of classifying the unseen concepts by utilizing their semantic information. The present study introduces image embeddings as side information on zero-shot audio classification by using a nonlinear acoustic-semantic projection. We extract the semantic image representations from the Open Images dataset and evaluate the performance of the models on an audio subset of AudioSet using semantic information in different domains; image, audio, and textual. We demonstrate that the image embeddings can be used as semantic information to perform zero-shot audio classification. The experimental results show that the image and textual embeddings display similar performance both individually and together. We additionally calculate the semantic acoustic embeddings from the test samples to provide an upper limit to the performance. The results show that the classification performance is highly sensitive to the semantic relation between test and training classes and textual and image embeddings can reach up to the semantic acoustic embeddings when the seen and unseen classes are semantically similar.**

***Index Terms*—zero-shot learning, audio classification, semantic embeddings, image embeddings**

I. Introduction

Supervised learning methods have enjoyed widespread adoption in various fields (e.g., computer vision, audio analysis). These methods mainly require a large amount of annotated data to robustly learn the concepts in focus. Acquisition of such dataset typically requires manual labeling which is expensive and time-consuming. Additionally, these methods have the capability to learn only the concepts defined in the development stage of the models. In many practical scenarios, however, the application should have the ability to learn unseen concepts, e.g., when it is challenging to collect a sufficient amount of instances belonging to a target class that is rare or changes over time.

In order to overcome the aforementioned problems, zero-shot learning (also referred to as zero-data learning) was proposed in [1]. Zero-shot models aim to learn the concepts that have not been introduced during training by exploiting the semantic information about those classes. Based on the semantic information, both seen and unseen classes can be projected into the same representation space that is used to bridge between different domains. Consequently, the model acquires the ability to further mimic human cognition by transferring the knowledge between seen and unseen concepts.

Recently, zero-shot learning has drawn increased attention in computer vision tasks, e.g., in [2]–[6]. Due to the lack of labeled instances belonging to the unseen zero-shot classes, side information about the unseen classes is required. Commonly used side information for zero-shot object recognition is reviewed in [7]. The authors classified the side information into two sets: semantic attributes and beyond. Semantic attributes refer to the intrinsic characteristics or properties of visual classes. Semantic information beyond attributes includes representations that are learned from the textual descriptions of the classes.

Despite the increasing popularity of zero-shot learning in computer vision tasks, its applications and consequently the investigation of semantic embeddings in the audio content analysis have been fewer in comparison. Zero-shot audio classification based on class embeddings with a bilinear model was studied in [8], [9], and the work was extended with a non-linear model in [10]. The proposed approaches were based on the compatibility between the audio and the semantic embeddings of the corresponding concepts where class labels of the audio samples were used as semantic information by employing word embedding models in [8] and sentence embedding models in [9]. Zero-shot audio classification using image embeddings as semantic information has not been studied so far.

In this paper, we extend the previous work in [10] by introducing visual semantic information to the nonlinear acoustic-semantic projection model. Furthermore, we evaluate the performance of textual and visual embeddings both individually and together based on the classification accuracy. We compare our results with the model using acoustic embeddings as side information. In this work, we focus on the following questions: i) how does the performance of zero-shot audio classification change with different semantic spaces, ii) is the classification performance affected by the semantic relation between the training and test classes, and iii) if so, how does the performance change based on the hierarchical order of the semantically related classes.

The remainder of this paper is organized as follows. First, Section II introduces the selected zero-shot learning model for audio classification. Then, Section III describes the dataset and acoustic and semantic embeddings. Next, Section IV explains the evaluation setup and reports the experimental results. Finally, Section V concludes the paper with a discussion.

II. Zero-Shot Learning for Audio Classification

In this section, we formalize the zero-shot audio classification setting and then describe the selected methodology.

The pipeline of the zero-shot audio classification model is illustrated in Figure 1. First, audio feature embeddings are extracted from audio clips using an acoustic embedding model. Based on the semantic information that is used in zero-shot learning, semantic embeddings are obtained with an image embedding model that takes image samples or a language embedding model that takes textual class labels as input. Next, the model learns an acoustic-semantic projection by measuring the compatibility between two spaces during training. In the test time, the model takes unknown audio samples and applies the learned acoustic-semantic projection between the acoustic embedding of the unknown audio recordings and semantic embeddings of the new sound classes. Finally, a classifier selects the label whose semantic representation has the highest compatibility with the unknown audio sample.

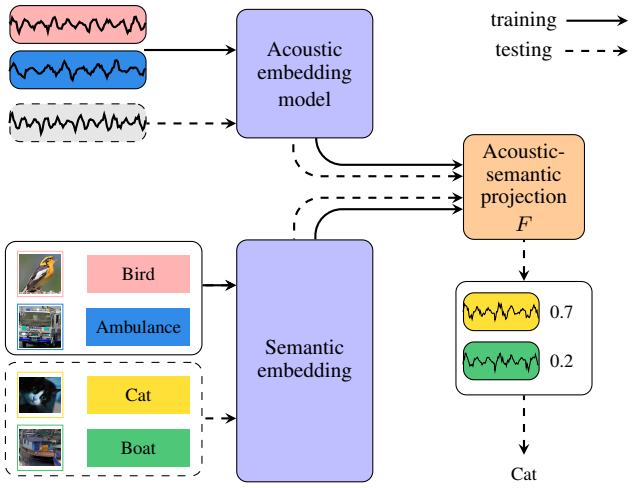

Fig. 1. Zero-shot audio classification using visual semantic information. Training is represented in solid line and testing in dashed line. Sound classes *bird* and *ambulance* are used during training and *cat* and *boat* are used in testing.

We denote the audio sample space by X, and the semantic space of the seen and unseen sound classes by Y and Z, respectively, where $Y \cap Z = \emptyset$. Given an audio instance $x \in X$, and a sound class $y \in Y$, the main goal of the zero-shot learning is to learn a classifier $f: X \rightarrow Y$, that is defined as

$$f(x) = \underset{y \in Y}{\operatorname{argmax}} F(x, y), \tag{1}$$

where F is the compatibility function between the representations of an audio instance and a sound class.

In the present work, we project semantic embeddings onto acoustic embeddings due to the higher dimensionality of the semantic embeddings compared with acoustic embeddings. By doing so, we aim to capture the semantic embeddings in a lower and more relevant representational space. We define the compatibility function F as the dot product between the acoustic embedding $\theta(x) \in R^{d_a}$ of an audio instance x and the projection of the semantic embedding $\phi(y) \in R^{d_s}$ of a sound class y as

$$F(x, y) = H(\phi(y))^T \theta(x), \tag{2}$$

where H is the acoustic-semantic projection.

We follow the nonlinear acoustic-semantic projection model proposed for zero-shot audio classification in [10]. In order to decrease the computational complexity, the model uses factorization of a bilinear model and introduces nonlinear activation to capture the nonlinear relationship between acoustic and semantic embeddings. Consequently, we define H as

$$H(\phi(y)) = V^T t(U^T \phi(y)), \tag{3}$$

where $U_{d_s \times r}$ and $V_{r \times d_a}$ are the learned projection matrices and t is a nonlinear activation function.

The underlying assumption of zero-shot learning is that the acoustic embedding of an audio instance is expected to be closer to the semantic embedding of the corresponding class in the acoustic-semantic space rather than those of other classes. Under this assumption, in test time, the model aims to assign an audio sample to an unseen class label that has the highest compatibility with.

III. Audio, Visual, and Textual Embeddings

In this section, we describe the dataset and acoustic and semantic embeddings.

A. Dataset

To our knowledge, there is no joint audio and image dataset where mutual classes are defined. Therefore, we conducted our experiments based on separate datasets; the audio dataset, AudioSet [11] and image dataset, Open Images [12]. AudioSet is an unbalanced general audio dataset that contains 527 sound classes with over two million audio clips in total. The samples in AudioSet contain clip-level textual labels that consist of one or several words and a longer sentence description for every sound class. In this work, we use only the class labels rather than the sentence descriptions. Open Images is a dataset of around nine million images with image-level labels, object bounding boxes, object segmentation masks, and visual relationships. It contains a total of 16 M bounding boxes for 600 object classes on 1.9 M images. In our experiments, we use the images whose bounding boxes are defined.

In order to semantically exploit every audio instance that belongs to one class by their visual representations, we defined mutual classes that occur in both AudioSet and Open Images. Both datasets provide the textual class labels with their Knowledge Graph Machine IDs (MID) [13] which are unique identifier codes for each entity in all languages. We define the mutual classes between AudioSet and Open Images by using the MID of the classes instead of string matching. Therefore, we prevent mistakenly capturing homonym word labels as well as losing information about the mutual classes that are defined with synonym words. Among 527 sound classes and

600 object classes, there are 79 mutual classes defined with the same MID in AudioSet and Open Images.

AudioSet Ontology further provides a tree-like hierarchical structure for individual classes, where each audio sample is annotated with a more specific description when traversed to the deeper nodes. Using the hierarchical structure, we defined the child classes of each class until its deepest node. Among 79 mutual classes, we did not include the classes that have more than two children in the mutual class set (i.e. the class labels of Animal, Musical Instrument, Vehicle). In this way, we prevent the model from attempting to classify an audio instance either as "Animal" or "Cat". For the classes that have only one or two child classes in the mutual class set, we only kept the parent class in the mutual class set (i.e. Bird, Insect, Car, Aircraft, Tools) as the parent classes provide with more samples without significantly reducing the number of classes. We used the remaining 69 classes in our experiments. The least populated classes have 10 and 4 samples in AudioSet and Open Images, respectively. In order to provide a more balanced set, we randomly sampled the over-represented audio and image classes until each class has at most 300 and 1000 samples, respectively.

B. Acoustic Embeddings

Following [8]–[10], we used a pre-trained VGGish [14] to obtain acoustic embeddings. First, a 10-second audio clip was split into ten 1-second audio segments. Next, log mel spectrograms with the size of 96x64 were extracted for each segment and passed into VGGish that produced 128-dimensional audio embeddings. Then, the clip-level audio embeddings were obtained by averaging the embeddings extracted from the audio segments within the clip they belong to. It should be noted that the VGGish should be trained from scratch to reduce the bias arising from the seen classes used to train the model.

C. Semantic Embeddings

We compare class embeddings from three different domains: textual, audio, and image embeddings. We utilized pre-trained Word2Vec [15], ResNet101 [16], and VGGish [14] models for textual, image, and audio embeddings, respectively. Additionally, we concatenated the image and textual embeddings to capture the semantic information from both domains and denoted them as hybrid embeddings.

Word2Vec is a two-layered fully-connected neural network. The model learns the word embeddings by either predicting the current word by looking at its surroundings (bag-of-words) or the surrounding words from the current word (skip-gram). The model produces a 300-dimensional vector to represent the input word. We extracted the word embeddings based on the textual class labels of the sound classes defined in AudioSet. For the sound classes that are defined with multiple words, we averaged the extracted word embeddings.

ResNet101 is a 101-layered convolutional neural network that is trained on 1.28 million images composed of 1000 classes in the ImageNet database. The model takes images with the size of 224x224. First, we extracted the objects that belong to the mutual classes from the defined bounding boxes and resized them into 224x224. Then, we obtained a 1000-dimensional vector from the classification layer of the ResNet101 model for each image instance. We averaged the image embeddings over the samples within each class to represent the sound classes. It should be noted that the ResNet101 model with the current setup achieves an accuracy of 84% on the supervised image classification task.

By the nature of zero-shot learning, the model aims to exploit semantic information assuming that there is not enough labeled data in test time. Therefore, in the task of zero-shot audio classification, it is not possible to use the audio data as semantic information. However, in this study, similar to the visual embeddings, we utilize the acoustic embeddings as semantic information by averaging the acoustic embeddings of audio instances that are used in the classification over the classes. As these semantic audio embeddings are calculated from the samples of test data, we use semantic audio embeddings with the purpose of measuring the known upper limit of the zero-shot learning performance that can be reached by using the semantic information provided with the test set itself.

IV. Results and Analysis

In this section, we first describe the evaluation setup used to examine the effect of different semantic spaces and next present the experimental results of the zero-shot learning model.

A. Evaluation

In our experiments, we evaluate zero-shot audio classification based on two different strategies, random-based class partitions, and category-based class partitions. On random-based class partition strategy, we randomly split the extracted dataset into three disjoint subsets for training, validation, and testing, which all at the end have 23 sound classes and a similar number of audio samples and image samples. We ran the experiments five times with different splits. Table I displays the selected subsets on one trial with random-based class partitions.

TABLE I
SELECTED SUBSETS FOR ONE TRIAL WITH THE RANDOM-BASED CLASS PARTITIONS

Subset	Class Labels
Training	violin, bus, cat, door, horse, goose, scissors, ratchet, sheep, alarm clock, goat, tap, snake, chicken, car, trombone, harp, chainsaw, tool, frog, cupboard, coin, microwave oven
Validation	camera, ambulance, motorcycle, aircraft, horn, duck, guitar, drum, cattle, mechanical fan, skateboard, chime, television, bike, piano, harpsichord, sink, bird, bathtub, sewing machine, telephone, mouse, drawer
Test	hand, boat, dog, computer keyboard, truck, insect, flute, train, clock, pig, maracas, banjo, tableware, printer, trumpet, accordion, saxophone, cello, harmonica, blender, organ, hair dryer, toothbrush

On category-based class partition strategy, we conduct zero-shot learning within and across categories with the aim of

investigating whether the hierarchical order of the semantically related classes affects the behavior of the classifier. We define a category as a class whose several child classes occur in the mutual classes, i.e. Animal, Musical Instrument, and Vehicle. Table II shows the classes that belong to each of the categories. First, we conducted zero-shot learning within each category by randomly splitting the classes into training, validation, and test sets. In each split, the test sets of each category contain two different classes, and training and validation sets of the categories include 8, 6, and 4 different classes for Musical Instrument (MI), Animal (A), and Vehicle (V), respectively. Second, we evaluated the zero-shot learning across categories by combining two different categories and randomly splitting the classes where each subset has the same number of classes from both categories. When performed across categories, we included same number of classes from each category in all the subsets by limiting the number of classes based on the smallest category, i.e. Vehicle. We randomly selected four classes from each represented category for training and validation sets in each trial. In the test set, we combined one random class from each category. We ran the experiments five times with different splits for each strategy and category.

TABLE II
CLASS LABELS FOR DIFFERENT CATEGORIES

Category	Class Labels
Animal	cat, horse, goose, sheep, goat, snake, chicken, frog, duck, cattle, bird, mouse, dog, pig
Musical Instrument	trombone, harp, guitar, cello, drum, organ, chime, piano, harpsichord, flute, maracas, banjo, trumpet, accordion, saxophone, harmonica, horn, violin
Vehicle	skateboard, bus, car, motorcycle, boat, aircraft, bike, train, truck, ambulance

In all the setups, we implemented (3) with two fully-connected layers where r equals to d_s and t is the tanh activation function. We trained all the models using a batch size of 32, and SGD optimizer with the learning rate of 10^{-2} for 200 epochs.

B. Random-based zero-shot learning

Random-based zero-shot audio classification results are reported in Table III. The table shows the average and standard deviation of the accuracies with five trials where each trial has different randomization for class partitioning. In each run, training, validation, and test sets included 23 different classes. Overall, each model performs better than the random guess of 0.04. As the averaged audio embeddings represent the class-level audio instances that are used during classification, the model using audio embeddings determined the known upper limit for the performance. Between image and textual embeddings, we did not observe a significant advantage of choosing one over another in the current setup. Moreover, the model using the hybrid embeddings (image + textual) did not provide any improvement on the model performance.

TABLE III
ZERO-SHOT AUDIO CLASSIFICATION WITH RANDOM-BASED CLASS PARTITIONS WHERE RANDOM GUESS IS 0.04

Class embeddings	Accuracy [0–1] (mean ± SD)
Audio	0.32 ± 0.02
Image	**0.17 ± 0.02**
Text	0.15 ± 0.04
Hybrid	0.16 ± 0.04

Additionally, we conducted experiments by randomly selecting audio and image embeddings among each class in each run to represent the class to which the selected instance belongs. The accuracy results of using audio and image embeddings with five runs were 0.16 ± 0.14 and 0.12 ± 0.03, respectively. Comparing the performance between averaging the embeddings over classes and randomly selecting, these results indicate that the averaging eliminates the undesirable effect of the outliers, and classes can be represented more accurately by taking advantage of the representational variety.

C. Category-based zero-shot learning

Zero-shot audio classification results within categories are reported in Table IV. In all cases, the models perform better than the random guess of 0.50. The average of the accuracy scores shows that both individual and combined image and textual embeddings can easily reach up to the performance of audio embeddings when the seen and unseen concepts are semantically similar. Yet, the standard deviation of the accuracy scores indicates that the performance of the model is sensitive to the variation of the classes within the category. For instance, the accuracy of the model using hybrid embeddings is 0.75 when the test classes are selected as "boat" and "ambulance" whereas the accuracy is reported as 0.54 when the test set consists of "car" and "truck". In the hierarchical structure of the AudioSet, the class of "boat, water vehicle" is defined as the immediate subcategory of Vehicle whereas "car", "truck", and "ambulance" are defined under the "motor vehicle (road)". For further improvement, the intermediate categories should be explored.

TABLE IV
TOP-1 ACCURACIES (MEAN AND SD) OF ZERO-SHOT AUDIO CLASSIFICATION WITHIN CATEGORIES WHERE RANDOM GUESS IS 0.50

Class Embedding	Categories		
	Animal	Musical Instrument	Vehicle
Audio	0.61 ± 0.05	0.66 ± 0.09	0.61 ± 0.05
Image	0.53 ± 0.07	0.57 ± 0.19	**0.62 ± 0.13**
Text	0.60 ± 0.06	0.60 ± 0.09	0.61 ± 0.10
Hybrid	**0.63 ± 0.08**	**0.62 ± 0.12**	0.59 ± 0.12

Zero-shot audio classification results across categories are reported in Table V. As can be seen from the results, the model performs significantly better than the random guess of 0.50 in each setup. Compared with the results of within-category experiments, we observe that the model can easily

TABLE V
TOP-1 ACCURACIES (MEAN AND SD) OF ZERO-SHOT AUDIO CLASSIFICATION ACROSS CATEGORIES WHERE RANDOM GUESS IS 0.50

Class Embedding	**Categories**		
	A + MI	A + V	MI + V
Audio	0.92 ± 0.02	0.80 ± 0.18	0.94 ± 0.04
Image	0.89 ± 0.03	0.77 ± 0.07	0.87 ± 0.15
Text	**0.90 ± 0.05**	0.84 ± 0.08	0.95 ± 0.04
Hybrid	0.88 ± 0.08	**0.85 ± 0.06**	**0.96 ± 0.04**

learn different categories. This indicates that the classifier can transfer the knowledge from one class to another in the same category.

V. DISCUSSION AND CONCLUSION

In this paper, we introduced image embeddings as semantic information for zero-shot audio classification. The zero-shot model was evaluated by comparing the performance of the models using visual, audio, and textual embeddings as semantic information. We used VGGish, ResNet101, and word2vec models to extract audio, image, and textual embeddings, respectively. We performed the audio classification task by extracting the mutually observed classes between audio and image datasets.

The experiments showed that using visual semantic embeddings both individually and combined with textual embeddings can achieve better accuracy than a random guess on the zero-shot audio classification task. Yet, the difference between the results of using averaged and randomly selected semantic embeddings suggested that the performance of the model can be improved in the presence of a larger dataset where the averaging eliminates outliers more accurately. Additionally, the experimental results showed that the image, textual, and hybrid embeddings can further reach up to the performance of acoustic semantic embeddings when the seen and unseen classes are semantically similar. However, the varying behavior on the different splits indicates that the performance of the model is sensitive to the class splits when performed within the same category. The model reached up to the highest accuracy when performed across categories.

Even though the visual embeddings might increase the representational power for each class by introducing rich variety to the model with a large number of image instances, they lack the ability to express more abstract classes. On the other hand, textual embeddings can capture the abstract classes in the presence of the nonambiguous textual class description. Yet, they provide each class only with a singular representation. Cognitively motivated approaches suggest that human semantic knowledge relies on perceptual information rather than only linguistic information [17]. In order to allow the model to capture complementary information from different modalities, future work should explore more advanced hybrid models. Furthermore, for a fully unbiased evaluation, the embedding models should be trained from scratch where the zero-shot classes are excluded from the training data.

ACKNOWLEDGMENT

This research has been supported by the Academy of Finland, project no. 338502. The authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources used towards this research.

REFERENCES

[1] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data Learning of New Tasks," in Proceedings of the National Conference on Artificial Intelligence, vol. 2, pp. 646-651, 2008.

[2] M. Palatucci, D. Pomerleau, G. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS), 2009, pp. 1410–1418.

[3] M. Rohrbach, M. Stark, and B.Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in CVPR, 2011.

[4] X. Yu and Y. Aloimonos, "Attribute-based transfer learning for object categorization with zero or one training example," in ECCV, 2010.

[5] X. Xu, Y. Yang, D. Zhang, H. T. Shen, and J. Song, "Matrix trifactorization with manifold regularizations for zero-shot learning," in CVPR, 2017.

[6] Z. Ding, M. Shao, and Y. Fu, "Low-rank embedded ensemble semantic dictionary for zero-shot learning," in CVPR, 2017.

[7] Y. Fu, T. Xiang, Y. G. Jiang, X. Xue, L. Sigal, and S. Gong, "Recent Advances in Zero-Shot Recognition: Toward Data-Efficient Understanding of Visual Content," in IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 112-125, 2018.

[8] H. Xie and T. Virtanen, "Zero-Shot Audio Classification Based on Class Label Embeddings," in Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustic. (WASPAA), 2019, pp. 264-267.

[9] H. Xie and T. Virtanen, "Zero-Shot Audio Classification Via Semantic Embeddings," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1233-1242, 2021.

[10] H. Xie, O. Räsänen and T. Virtanen, "Zero-Shot Audio Classification with Factored Linear and Nonlinear Acoustic-Semantic Projections," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 326-330.

[11] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrance, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 776-780.

[12] I. Krasin, T. Duerig, N. Alldrin, A. Veit, S. Abu-El-Haija, S. Belongie, D. Cai, Z. Feng, V. Ferrari, and V. Gomes, "OpenImages: A public dataset for large-scale multi-label and multi-class image classification", 2016.

[13] A. Singhal, "Introducing the knowledge graph: things, not strings," Official Google blog, https://blog.google/products/search/introducing-knowledge-graph-things-not/ (accessed March 7, 2022).

[14] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 131-135.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in Proceedings of International Conference on Learning Representation (ICLR), 2013.

[16] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[17] M. Louwerse, "Symbol Interdependency in Symbolic and Embodied Cognition", in Topics in Cognitive Science, vol. 3, pp. 273-302, 2011