Sensory-Aware Multimodal Fusion for Word Semantic Similarity Estimation

Georgios Paraskevopoulos, Giannis Karamanolakis, Elias Iosif, Aggelos Pikrakis† and Alexandros Potamianos
School of Electrical and Computer Engineering, National Technical University of Athens, Greece
†Department of Informatics, University of Piraeus, Greece
geopar@central.ntua.gr, giannis.karamanolakis@gmail.com, iosife@central.ntua.gr, pikrakis@unipi.gr, potam@central.ntua.gr

***Abstract*—Traditional semantic models are disembodied from the human perception and action. In this work, we attempt to address this problem by grounding semantic representations of words to the acoustic and visual modalities. Specifically we estimate multimodal word representations via the fusion of auditory and visual modalities with the text modality. We employ middle and late fusion of representations with modality weights assigned to each of the unimodal representations. We also propose a fusion method that assigns different weights to each word, based on how relevant that word is for the audio and visual modalities. The proposed methods are evaluated for the task of semantic similarity computation between words. To our knowledge, this is the first work that combines text, audio and visual features for the computation of multimodal semantic word representations. Multimodal models outperform the unimodal models, indicating the importance of multimodal fusion and perceptual grounding.**

I. Introduction

Semantic models represent the meaning of various language entities, such as words, phrases and sentences. For example, Distributional Semantic Models (DSMs) rely in the distributional hypothesis of meaning [1], according to which words that appear in similar contexts tend to have similar meaning. Based on the distributional hypothesis, DSMs encode patterns of word co-occurrence in text corpora. These models have been successfully applied for the estimation of lexical semantic similarity and used broadly in a variety of NLP tasks.

However, traditional semantic models have been criticized as "disembodied", since they rely solely on linguistic information without being grounded on human perception and action. This problem is often referred to as the symbol grounding problem [2]. In the past, experiments confirmed the connection between word semantics and human perception [3].

Multimodal semantic models have been employed to encounter the symbol grounding problem via the incorporation of multiple perceptual modalities into word representations [4], [5], [6]. Multimodal semantic models have also been applied for various multimodal tasks, including audio auto-tagging and music similarity estimation [7], music instrument clustering [8], image labeling and retrieval [9], etc. The term *multimodal fusion* is used to indicate the integration of information from multiple modalities.

In this work, we fuse text-, audio- and image-based models for the estimation of word semantic similarity. Two main fusion methods are employed here, namely middle and late fusion. In addition, a method is proposed for the estimation of fusion weights based on the sensorial scores of each word, i.e., the degree of relevance of each word with the auditory and visual modality. To the best of our knowledge, this is the first attempt to include both the auditory and visual modalities for the computation of joint, multimodal semantic representations of words.

II. Related Work

In this section, an overview of unimodal semantic models is provided, along with indicative references to research efforts dealing with the fusion of such models.

A. Text-based Semantic Models

DSMs aim to to represent the semantics of words as vectors in high-dimensional spaces. Such spaces enable the computation of similarity between words as the vector similarity of their representations. One of the most common ways of creating DSMs is through vector spaces [10], where the vectors are constructed by obtaining co-occurence counts from text corpora by considering the "context" of the target words. For example when the semantic space is constructed using the Latent Semantic Analysis [11] method, the context is defined as the paragraph (or document) where the target word occurs. Another example is the Hyperspace Analog to Language model (HAL) [12], where context is defined as the surrounding words of the target word within a predefined window size.

A more recent approach of creating DSMs is through the extraction of word embeddings, i.e., real-valued vector representations of words, from Neural Language Models (NLMs). One of the early NLM approaches is proposed in [13] relying on a probabilistic model that consists of a shallow feed-forward neural network. The network is trained on a large corpus to predict the next word given a sequence of words and is composed of three layers. Recent NLMs are based on the same principle, but the intermediate layer is substituted with an Long Short Term Memory network (LSTM) [14], [15]. An approach for constructing word embeddings that has gained in popularity is word2vec [16], which focuses on computational efficiency. To this end, word2vec forgoes with the expensive intermediate layers and proposes two training strategies, skip-gram where the model is trained to predict the context (surrounding words) of a target word and CBOW where the model is trained to predict a target word given a window of surrounding words. Another popular alternative for

computing word embeddings which focuses on computational efficiency is fastText [17].

B. Audio-based Semantic Models

Audio-based semantic models enable the computation of word semantic representations by taking into account the association of tags with audio clips in clip collections. In [18] an audio-based DSM was constructed using the Bag-of-Audio-Words (BoAW) technique. The model was extended in [8] by combining linguistic and auditory features. Typically, ADSMs are constructed via a three step procedure. First, acoustic features (e.g., Mel-scale Frequency Cepstral Coefficients - MFCCs) are extracted from segmented audio clips. Then, clip vector representations are computed using the Bag-of-Audio-Words (BoAW) method. Specifically, the acoustic feature vectors extracted from the segments of a clip, c, are quantized to the nearest clusters (called audio-words) and clip c is represented as a bag (or histogram) of audio-words. The final step is the computation of tag representations. Each tag derives a BoAW representation by averaging the representations of the clips annotated with this tag. The aforementioned steps, are detailed in [19].

C. Image-based Semantic Models

As in the case of audio-based semantic models, semantically tagged images (i.e., images associated with lexical descriptions of the depicted content) can be exploited for the construction of semantic models. In [20], text-based and visual-based multimodal models using Bag of Visual Words (BoVW) representations [21] are extracted from a collection of news articles along with article images. The BoVW representations are derived via the extraction of (scale-invariant feature transform) SIFT [22] features from images and the vector quantization using k-means clustering. In [23], [24], traditional DSMs and visual DSMs (VDSMs) are constructed and two methods are proposed for fusion. According to the first method, the vectors of the respective spaces are concatenated into a joint multimodal space. Using the second method, the unimodal representations are used individually to produce similarity scores between pairs of words and the individual scores are then averaged (using weights) for the computation of the final score. A method to extract embeddings for images with the use of a Deep Convolutional Neural Network (CNN) is proposed in [25], where the CNN architecture proposed in [26] was adopted and trained on ImageNet [27]. In [25], is demonstrated that CNN embeddings outperform their BoVW counterpart on semantic similarity tasks.

III. Multimodal Fusion

In this section, three types of fusion schemes are presented. The first scheme (middle fusion, see Section III-A) deals with the creation of multimodal semantic representation for the words of interest. Such representations are exploited for computing word semantic similarity. A late fusion scheme is presented in Section III-B, where modality-specific word similarity scores are combined. The information exploited in middle and late fusion schemes (semantic representations and similarity scores) are linearly combined via weights that do not depend on the words (i and j) for which the similarity is computed. In Section III-C, the sensory content ("audio-ness" and "visual-ness" scores) of i and j is taken into account for weighting their respective semantic representations (middle fusion) and similarity scores (late fusion).

A. Middle Fusion

Middle fusion works in two steps. First, given a word i, unimodal representations derived from text-, audio- and image-based models (denoted as r_T^i, r_A^i, r_V^i respectively) are fused. Thus, a joint multimodal representation r_{MM}^i is computed for i:

$$r_{MM}^i = u_T r_T^i \quad || \quad u_A r_A^i \quad || \quad u_V r_V^i, \tag{1}$$

where $||$ denotes the vector concatenation operator and $u_T, u_A, u_V \in [0, 1]$ correspond to the fusion weights for each modality and sum up to one. The estimation of semantic similarity between two words, i and j, is estimated in the multimodal space as a similarity metric between the corresponding multimodal representations:

$$S_{MF}(i, j) = s(r_{MM}^i, r_{MM}^j), \tag{2}$$

where $s(x, y)$ denotes a similarity metric between vectors x and y. Here, cosine similarity is used as a similarity metric.

B. Late Fusion

Using late fusion, the semantic similarity between two words, i and j, is estimated for each modality separately, resulting in the following three similarity scores $s_T(i, j)$, $s_A(i, j)$, $s_V(i, j)$, corresponding to the text, audio and visual space respectively:

$$\begin{aligned} s_T(i, j) &= s(r_T^i, r_T^j), \\ s_A(i, j) &= s(r_A^i, r_A^j), \\ s_V(i, j) &= s(r_V^i, r_V^j). \end{aligned} \tag{3}$$

Then, the final similarity score is computed as the weighted combination of the modality-specific similarity scores:

$$S_{LF}(i, j) = \lambda_T s_T(i, j) + \lambda_A s_A(i, j) + \lambda_V s_V(i, j), \tag{4}$$

where λ_T, λ_A and λ_V are the fusion weights for the three modalities and $\lambda_T + \lambda_A + \lambda_V = 1$.

C. Sensory Aware Multimodal Models (SaMM)

The motivation behind this method is that some words are inherently more relevant to specific perceptual modalities and less relevant to other modalities. For example, the word 'red' is primarily relevant to the visual modality while 'guitar' is primarily associated with the acoustic modality. The weights that described in Sections III-A and III-B for the fusion of the three modalities are common for each word and independent from its sensorial properties. Therefore, we propose a new method that takes into account the sensorial properties of words and assigns different weights to each word based on their sensorial properties. To achieve this we use the

Sensicon lexicon [28] which contains 22684 English words and associates each word with 5 numerical scores. The scores correspond to the relevance of the word to each of the 5 senses, namely vision, hearing, taste, smell and touch. In this work, values corresponding to visual and audio scores are used. Some examples scores from Sensicon are presented in Table I. We observe that the word "red" is primarily associated with the visual sense, "dog" has a balanced score between hearing and vision, while "guitar" is primarily associated to hearing.

Word	Visual Score	Audio Score
red	0.83	0.31
dog	0.36	0.44
guitar	0.23	0.74

TABLE I: Examples from the Sensicon lexicon

Next, we describe two sensory-aware fusion weight computation strategies, one for middle fusion and one for late fusion.

Middle Fusion: the text modality weight u_T is a word-independent parameter

$$u_T = a, \quad 0 \leq a \leq 1. \tag{5}$$

The audio weight for the word i is computed using the following equation:

$$u_A^i = \sqrt{\frac{\beta_A^i}{\beta_V^i}} u_V^i, \tag{6}$$

where β_A^i and β_V^i are the "hearing" and "vision" score respectively, retrieved from Sensicon. The motivation behind (6) is that if a word is e.g., more relevant to the hearing sense than the vision, the audio modality should contribute more than the visual. The square root function smooths the ratios of the Sensicon scores. We compute u_V^i from the constraint $u_V^i + u_T^i + u_A^i = 1$. In the case of middle fusion, these weights are used for the weighting of the multimodal vectors in (1).

Late Fusion: Here, we follow a diffrent approach from middle fusion, since we are weighting similarities for word pairs instead of individual words, by modifying (6) to

$$\lambda_A^{(i,j)} = \sqrt{\frac{\beta_A^i + \beta_A^j}{\beta_V^i + \beta_V^j}} \lambda_V^{(i,j)}. \tag{7}$$

λ_T and $\lambda_V^{(i,j)}$ are computed the same way as u_T and u_V^i.

IV. Experimental Settings

Text-based Model: For the text model we use the freely available fastText pretrained embeddings[1] for the English language. The model consists of 300-dimensional vectors for 2519371 English words trained on the entire Wikipedia corpus. The vectors were constructed using a subword model [29], which is an extension of the skip-gram model [16] that takes into account the morphological properties of words and character-level information. Because of this, the subword model is able to create better representations for rare words and even produce vectors for words that do not exist in the training corpus.

Audio-based Model: The ADSM was built on 11192 audio clips and 2467 unique tags downloaded from the online search engine Freesound [30] using the Freesound API. The audio clips were converted to WAV format and resampled to 22.05kHz. For each clip, a feature vector is extracted from windows of 250 ms with a step of 100 ms. The feature vectors consist of 13 MFCCs (concatenated with spectral energy), and their 1st and 2nd order derivatives, yielding a vector of 39 coefficients. These features are clustered to $k = 300$ clusters using mini-batch k-means [31]. The ADSM was built using the BoAW method described in Section II-B, resulting in 2467 word representations of length 300.

Image-based Model: For the visual model (VDSM) we considered using both BoVW and CNN image representations of the images in the ESP game dataset [32]. The ESP game dataset contains 100000 images labeled with 29845 unique tags by human annotators. The annotation of the images is done through a game with a purpose, where 2 randomly matched annotators, who cannot communicate with each other, are presented with the same image and must agree on an appropriate tag that describes the image. The ESP game dataset images illustrate complex scenes with multiple and frequently off-center objects with a variety of tags, which leads to a dataset with more noisy images than e.g., ImageNet [23] but also to a dataset with increased word coverage which can capture more complex interpretations of images. The image embeddings were extracted from the 7th layer of AlexNet using the Caffe deep learning framework [33] and the MMFeat framework [34].

Multimodal Fusion: We experimented both with middle and late fusion for the multimodal model construction as described in Sections III-A, III-B. The optimal set of fusion weights (u_T, u_A, u_V) and $(\lambda_T, \lambda_A, \lambda_V)$ were computed using exhaustive search and applying the methods described in Section III-C.

V. Evaluation Results

The Multimodal Semantic Models (MMSMs) are evaluated for the task of word semantic similarity computation. We use the MEN [23] and SimLex-999 [35] datasets as the ground truth. Both datasets are provided in the form of lists of word pairs, where each pair is associated with a similarity score. This score was computed by averaging the similarities that provided by human annotators. In order to provide equal comparisons between different semantic models, all models are evaluated on words for which text-, audio- and image-based representations are available. This process reduces the number of pairs from 3000 to 2243 for MEN and from 999 to 244 for Simlex-999. Regarding the proposed (automatic) models, the similarity score between two words is estimated as the cosine similarity between the corresponding vector representations. The Spearman correlation coefficient between

[1] https://github.com/facebookresearch/fastText

Text	Audio	Visual	(u_T,u_A,u_V)	**MEN**	(u_T,u_A,u_V)	**SimLex-999**
✓			(1.0, 0.0, 0.0)	0.768	(1.0, 0.0, 0.0)	0.378
	✓		(0.0, 1.0, 0.0)	0.428	(0.0, 1.0, 0.0)	0.296
		✓	(0.0, 0.0, 1.0)	0.530	(0.0, 0.0, 1.0)	0.119
✓	✓		(0.7, 0.3, 0.0)	0.785	(0.6, 0.4, 0.0)	**0.420**
✓		✓	(0.6, 0.0, 0.4)	0.782	(1.0, 0.0, 0.0)	0.378
	✓	✓	(0.0, 0.6, 0.4)	0.608	(0.0, 1.0, 0.0)	0.296
✓	✓	✓	(0.5, 0.2, 0.3)	**0.795**	(0.6, 0.4, 0.0)	**0.420**
✓	✓	✓	SaMM ($a = 0.5$)	**0.793**	SaMM ($a = 0.6$)	0.401

TABLE II: Middle fusion: correlation coef. for MEN and SimLex-999 datasets.

Text	Audio	Visual	(λ_T,λ_A,λ_V)	**MEN**	(λ_T,λ_A,λ_V)	**SimLex-999**
✓			(1.0, 0.0, 0.0)	0.768	(1.0, 0.0, 0.0)	0.378
	✓		(0.0, 1.0, 0.0)	0.428	(0.0, 1.0, 0.0)	0.296
		✓	(0.0, 0.0, 1.0)	0.530	(0.0, 0.0, 1.0)	0.119
✓	✓		(0.8, 0.2, 0.0)	0.786	(0.8, 0.2, 0.0)	**0.421**
✓		✓	(0.7, 0.0, 0.3)	0.782	(1.0, 0.0, 0.0)	0.378
	✓	✓	(0.0, 0.7, 0.3)	0.609	(0.0, 1.0, 0.0)	0.296
✓	✓	✓	(0.6, 0.2, 0.2)	**0.797**	(0.8, 0.2, 0.0)	**0.421**
✓	✓	✓	SaMM ($a = 0.6$)	**0.796**	SaMM ($a = 0.7$)	0.402

TABLE III: Late fusion: correlation coef. for MEN and SimLex-999 datasets.

the human and the automatically computed similarity scores was used as evaluation metric.

The results[2] for all MMSMs[3] using middle fusion are presented in Table II. The three leftmost columns indicate whether the corresponding model (Text, Audio, Visual) is used (✓). The first three rows illustrate the performance of each individual (i.e., unimodal) model, while the next three rows deal with the performance of each combination of two unimodal models. The last two rows contain the 3-modality MMSM and the 3-modality SaMM respectively. Similarly to middle fusion, we report the late fusion evaluation results for the optimal parameters of all MMSMs in Table III.

The best correlation achieved for MEN using middle fusion is 0.795 via the fusion of the three models outperforming any other combination as well as the three unimodal models. Using middle fusion for SimLex-999, it is observed that the weights assigned to the visual modality are zero for all model combinations. The highest correlation (0.420) is achieved via the fusion of the text-based with the audio-based model. Regarding late fusion (see Table III), the best correlation achieved for MEN is 0.797 via the fusion of the three models, while for SimLex-999 the best correlation is 0.421 via the fusion of the text and audio modalities.

In this paragraph, a brief overview of the performance of state-of-the-art MMSMs is provided for the word similarity task. Those results are not directly comparable to the results presented in this work because the text model and the subsets of MEN and Simlex-999 used for evaluation are different. [4] In [23], a BoVW-based visual DSM was combined with a text DSM and a correlation of 0.78 was reported for the MEN dataset. In [18] a BoAW-based audio DSM was fused with a text DSM yielding 0.689 and 0.493 correlation for MEN and Simlex-999, respectively. In [5], a CNN-based visual DSM was combined with a text DSM and a correlation of 0.727 and 0.38 was reported for MEN and Simlex-999, respectively. Also in [5] a BoAW model was fused with a text DSM achieving 0.697 correlation for the MEN dataset.

Unlike the aforementioned approaches, the present work constitutes the first research effort where all three modalities are fused for estimating word semantic similarity.

VI. Conclusions

In this work, we created MMSMs with the goal to ground semantic representations on the audio and visual modalities. Also, we proposed a fusion method that assigns different fusion weights to the text, audio and visual spaces based on the relevance of each word to the auditory and visual modalities. All fusion methods were evaluated for the task of semantic similarity computation between words. To the best of our knowledge, this is the first work that exploits text-based, audio-based and visual-based models for this task. It was shown that the multimodal model derived from the fusion of the three unimodal models outperforms each unimodal model (relative improvement 11.3% compared to the best unimodal model)

[2]All the reported Spearmans coefficients are statistically significant with respect to a random model at 95% level according to paired-sample t-test.

[3]We observed that CNN vectors outperformed BoVW in all experimental configurations. So, the experimental results are reported only for the case of CNN vectors.

[4]The lack of a shared evaluation dataset for multimodal semantic models is a common issue. However, we think useful to mention the performance of related works.

and every other combination of unimodal models. Also the SaMM achieved comparable performance with that of the optimal MMSMs, and exceeded the unimodal and most of the bimodal MMSMs, both using middle fusion and late fusion. In the case of MEN, we see that the optimal weight distribution favors a balanced audio and visual modality contribution. This means that the audio and visual modalities provide complementary information and both enhance the text modality. In the case of Simlex we observe that incorporating the visual modality has detrimental effects. This may be caused by the choice of ESP game for the visual model construction.

In the future we plan to investigate more methods for the construction of SaMMs and use machine learning techniques to automate the parameter tuning process. We also plan to experiment with more image datasets (e.g., ImageNet) and investigate their effectiveness especially on Simlex-999. Finally, we plan to apply multimodal semantic models and more fusion methods for various multimodal tasks, such as zero-shot learning via cross-modal mappings.

ACKNOWLEDGMENT

This work has been partially supported by the BabyRobot project supported by EU H2020 (grant # 687831).

REFERENCES

[1] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[2] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[3] F. Pulvermüller, "Brain mechanisms linking language and action," *Nature Reviews Neuroscience*, vol. 6, no. 7, pp. 576–582, 2005.

[4] E. Bruni, G. B. Tran, and M. Baroni, "Distributional semantics from text and images," in *Proc. of the workshop on geometrical models of natural language semantics*, 2011, pp. 22–32.

[5] D. Kiela, "Deep embodiment: grounding semantics in perceptual modalities," University of Cambridge, Computer Laboratory, Tech. Rep., 2017.

[6] E. Iosif and A. Potamianos, "Crossmodal network-based distributional semantic models," *10th Language Resources and Evaluation Conference*, 2016.

[7] G. Karamanolakis, E. Iosif, A. Zlatintsi, A. Pikrakis, and A. Potamianos, "Audio-based distributional semantic models for music auto-tagging and similarity measurement," *arXiv preprint arXiv:1612.08391*, 2016.

[8] D. Kiela and S. Clark, "Multi- and cross-modal semantics beyond vision: Grounding in auditory perception," in *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2461–2470.

[9] A. Lazaridou, N. T. Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," *arXiv preprint arXiv:1501.02598*, 2015.

[10] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of artificial intelligence research*, vol. 37, pp. 141–188, 2010.

[11] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104, no. 2, p. 211, 1997.

[12] K. Lund and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence," *Behavior Research Methods*, vol. 28, no. 2, pp. 203–208, 1996.

[13] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[14] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models." in *AAAI*, 2016, pp. 2741–2749.

[15] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," *arXiv preprint arXiv:1602.02410*, 2016.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[17] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[18] A. Lopopolo and E. van Miltenburg, "Sound-based distributional models," in *Proc. of the 11th International Conference on Computational Semantics*, 2015, pp. 70–75.

[19] G. Karamanolakis, E. Iosif, A. Zlatintsi, A. Pikrakis, and A. Potamianos, "Audio-based distributional representations of meaning using a fusion of feature encodings," *Interspeech 2016*, pp. 3658–3662, 2016.

[20] Y. Feng and M. Lapata, "Visual information in semantic representation," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 91–99.

[21] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. of the International Workshop on Workshop on Multimedia Information Retrieval*, ser. MIR '07. New York, NY, USA: ACM, 2007, pp. 197–206.

[22] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. of the Seventh IEEE International Conference on Computer Vision*. IEEE, 1999.

[23] E. Bruni, N. K. Tran, and M. Baroni, "Multimodal distributional semantics," *J. Artif. Int. Res.*, vol. 49, no. 1, pp. 1–47, Jan. 2014.

[24] E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran, "Distributional semantics in technicolor," in *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 136–145.

[25] D. Kiela and L. Bottou, "Learning image embeddings using convolutional neural networks for improved multi-modal semantics," in *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[28] S. S. Tekiroglu, G. Özbal, and C. Strapparava, "Sensicon: An automatically constructed sensorial lexicon," in *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014.

[29] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *CoRR*, vol. abs/1607.04606, 2016.

[30] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proc. of the 21st ACM international conference on multimedia*. ACM, 2013, pp. 411–412.

[31] D. Sculley, "Web-scale k-means clustering," in *Proc. of the 19th international conference on World wide web*. ACM, 2010, pp. 1177–1178.

[32] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '04. New York, NY, USA: ACM, 2004, pp. 319–326.

[33] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[34] D. Kiela, "MMFeat: A toolkit for extracting multi-modal features," in *Proc. of ACL-2016 System Demonstrations*. Association for Computational Linguistics, 2016.

[35] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *Computational Linguistics*, 2015.