Multimodal Semantic Word Representations Grounded in the Human Perception

Giannis Karamanolakis

Department of Electrical and Computer Engineering
National Technical University of Athens (NTUA)

12 June, 2017

Goal of this thesis

Distributional Semantic Models (DSMs)

- Information Retrieval & Natural Language Processing
- Modeling **word semantics**

Our Goal:

- Extend DSMs to **Multimodal DSMs**
  - **Audio-based DSM (ADSM)**: Acoustic properties of words
- Fuse Audio-DSM with Text-DSM and Visual-DSM
- Evaluate Multimodal DSMs on **Word Semantic Similarity**
- Apply Audio-DSM for Music Information Retrieval tasks
  - **Audio Auto-Tagging**
  - **Music Similarity**

Prior work:

- E. Bruni and M. Baroni (2014, 2016): VDSM and Fusion with DSM
- A. Lazaridou (2015, 2016): VDSM and Fusion with DSM
- A. Lopopolo and E. Miltenburg (2015): First approach of ADSM
- D. Kiela and S. Klark (2016, 2017): Extended ADSM and Fusion with DSM

Outline

Distributional Semantic Models (DSMs)

- **Vector representations** of word semantics

Distributional Hypothesis

"**Words** that appear in **similar contexts** tend to have **similar meanings**"

- Counting **co-occurences** between **target words** and their **contexts**

	wheel	transport	passenger	tournament	London	goal	match
automobile	1	1	1	0	0	0	0
car	1	2	1	0	1	0	0
soccer	0	0	0	1	1	1	1
football	0	0	1	1	1	2	1

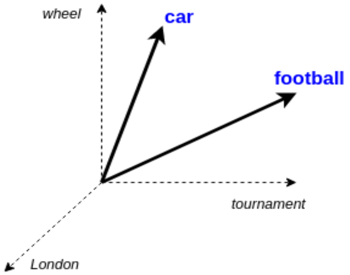
Table: The word-context matrix.

Introduction - Distributional Semantic Models (DSMs)

	wheel	transport	passenger	tournament	London	goal	match
automobile	1	1	1	0	0	0	0
car	1	2	1	0	1	0	0
soccer	0	0	0	1	1	1	1
football	0	0	1	1	1	2	1

Table: The word-context matrix.

- **Weighting**: TF-IDF, Pointwise Mutual Information (PMI)
- **Dimensionality Reduction**: PCA, Truncated SVD
- **Word Semantic Similarity**: Vector Similarity (e.g. cosine similarity)



Motivation: The Grounding Problem

- **Human Perception of words**
  - banana

  - guitar

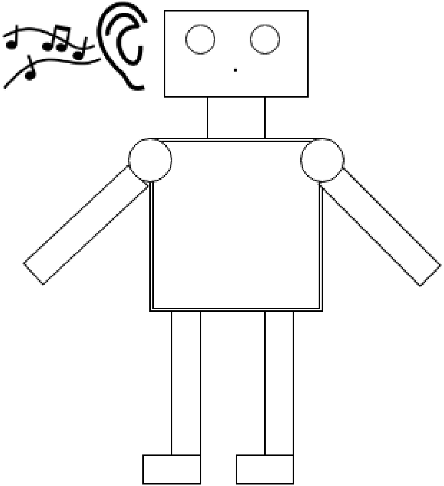
Motivation: The Grounding Problem

- **Human Perception of words**
  - banana

  - guitar

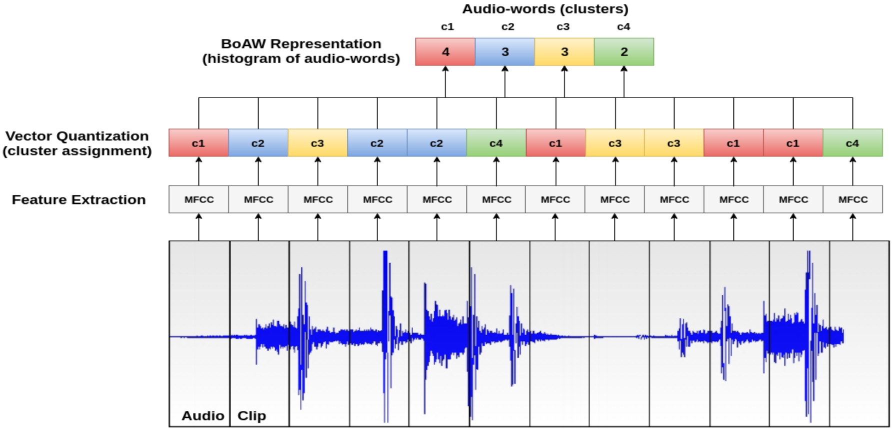
- **Grounding Problem**
  - DSMs rely solely on **text**
  - Acoustic/Visual properties of words?
  - DSMs are “disembodied” from the human **perception** and **action**

Our Goal



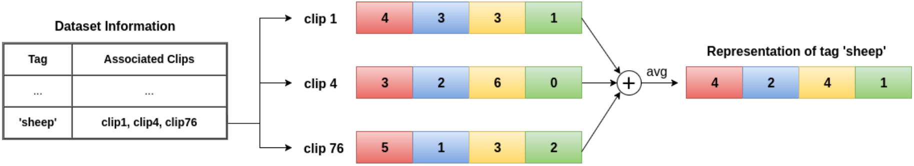
Audio-based DSM (ADSM)

- Extract **acoustic features** from **audio clips**
- Audio Clip Representations: Bag-of-Audio-Words (BoAW) approach (extension of the traditional Bag-of-Words method)



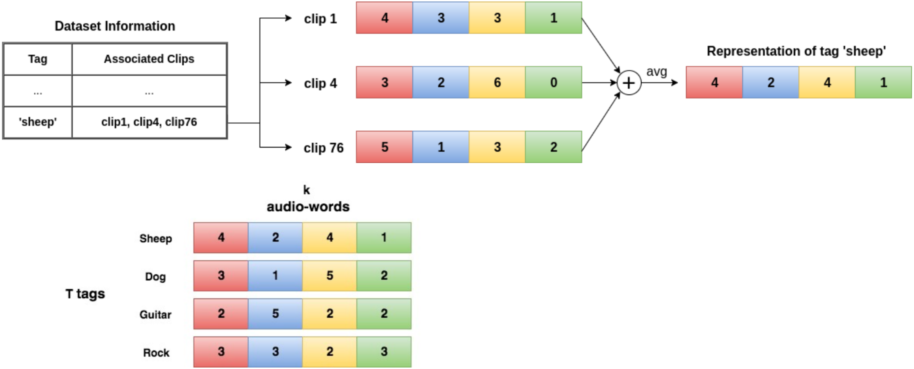
Audio-based DSM (ADSM) - Tag Representations

- **Word Representations via the ADSM**:
  - Metadata: **tags** describe clip **content**
  - Tag Representations: **averaging** the clip representations



Audio-based DSM (ADSM) - Tag Representations

- **Word Representations via the ADSM**:
  - Metadata: **tags** describe clip **content**
  - Tag Representations: **averaging** the clip representations



Audio-DSM - Summary

ADSM Computation Steps (Baseline):

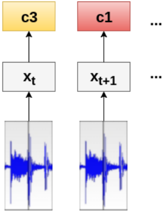
1. **Acoustic Feature Extraction**
2. **Clustering** (k-means)
3. **Vector Quantization (BoAW)** for clip encodings
4. **Average** clip encodings for tag encodings
5. **Weighting** (PMI)
6. **Dimensionality Reduction** (SVD)

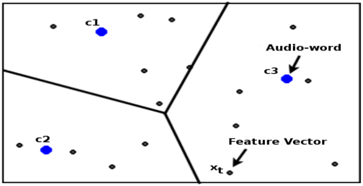
ADSM Extensions:

- Soft Cluster Assignment (Soft Encoding)
- Weighted Fusion of Feature Spaces

ADSM Extension: Soft Cluster Assignment

- Before: **Hard Cluster Assignment (Hard Encoding)**

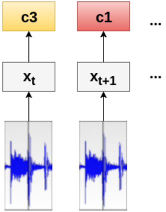


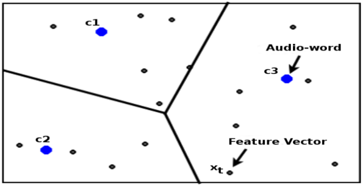


$$x_t \rightarrow e_t = (0, ..., 1, 0, ..., 0). \tag{1}$$

ADSM Extension: Soft Cluster Assignment

- Before: **Hard Cluster Assignment (Hard Encoding)**





$$x_t \rightarrow e_t = (0, ..., 1, 0, ..., 0). \tag{1}$$

- After: **Soft Cluster Assignment (Soft Encoding)**

$$x_t \rightarrow e_t^{'} = (w_1, w_2, ..., w_k), \tag{2}$$

where $\sum_{i=1}^{k} w_i = 1$

Soft Cluster Assignment: Calculation of weights

- **Calculation of weights**:
  - t-th acoustic vector: $x_t \in \mathrm{I\!R}^{\mathrm{d}}$
  - i-th acoustic word: $c_i \sim N(\mu_i, \Sigma_i), \quad \mu_i \in \mathrm{I\!R}^{\mathrm{d}}, \Sigma_{\mathrm{i}} \in \mathrm{I\!R}^{\mathrm{d \times d}}$

$$w_i = \frac{p(c_i|x_t)}{\sum_{j=1}^{k} p(c_j|x_t)}, \tag{3}$$

  - Using Bayes Rule and assuming Σ_i is diagonal:

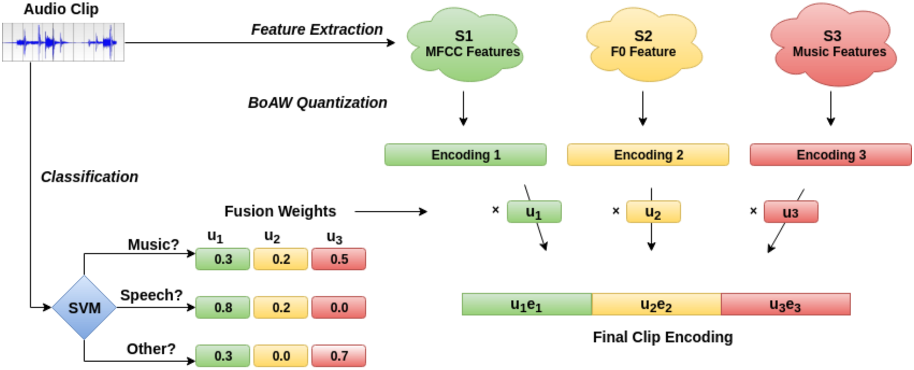
$$p(c_i|x_t) = \frac{p(x_t|c_i)p(c_i)}{p(x_t)} = \frac{p(c_i)e^{-\frac{1}{2}h_{ti}^2}}{(2\pi)^{d/2}|\Sigma_i|^{1/2}p(x_t)}, \tag{4}$$

  - h_{ti}: Mahalanobis distance between x_t and c_i,
  - $p(c_i)$: a-priori probability of cluster c_i,
  - $p(.)$: probabilities computed via ML estimation.

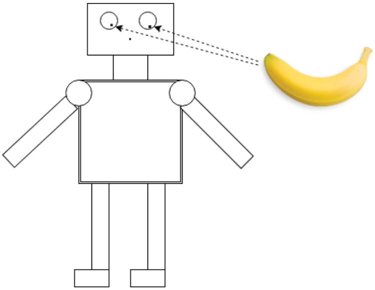
Finally:

$$w_i = \frac{p(c_i)|\Sigma_i|^{-1/2}e^{-h_{ti}^2}}{\sum_{j=1}^{k} p(c_j)|\Sigma_j|^{-1/2}e^{-h_{tj}^2}}, \tag{5}$$

ADSM Extension 2: Fusion of Feature Spaces



Visual Properties?



Visual DSM (VDSM) - Bag of Visual Words

- Extract **visual features** from **images**
- **Image Representations**: BoVW

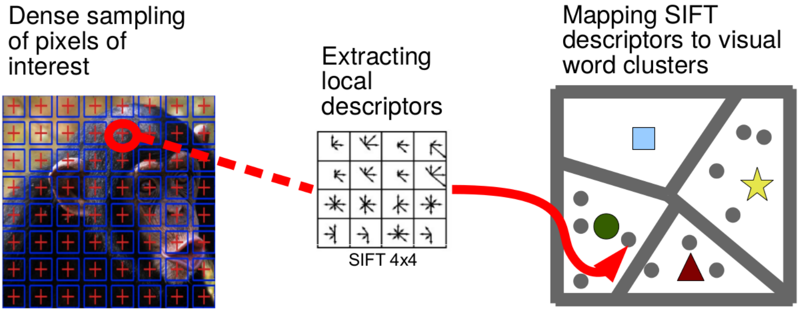


Figure: Bag of Visual Words approach. Source: Multimodal Distributional Semantics (Bruni et al. 2014)

- **Tag Representations**:

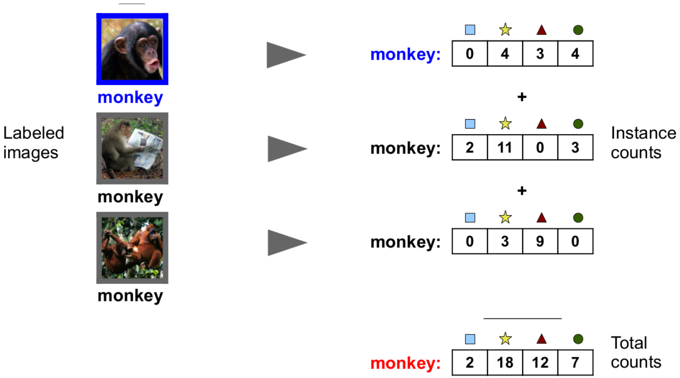
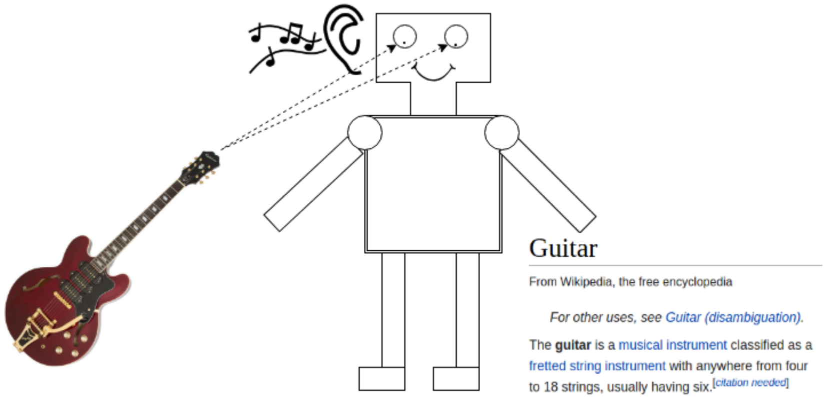


Figure: VDSM. Source: Multimodal Distributional Semantics (Bruni et al. 2014)

Multimodal Fusion?



Multimodal Fusion

Our work:

- Fuse **DSM**, **ADSM** and **VDSM**
- Estimate Word Semantic Similarity

Fusion Strategies:

- Early (Feature Level) Fusion
  1. **Fuse** (e.g. concatenate) the unimodal representations x_i, y_i, z_i
  2. Compute **cosine similarity** in the multimodal space

$$sim(fuse(x_1, y_1, z_1), fuse(x_2, y_2, z_2)) \tag{6}$$

- Late (Scoring Level) Fusion
  1. Compute **cosine similarity** separately for every modality
  2. **Fuse** (e.g. average) the similarity scores

$$fuse(sim(x_1, x_2), sim(y_1, y_2), sim(z_1, z_2)) \tag{7}$$

Outline

- **Task**: Estimation of Word Semantic Similarity
- **Groundtruth Data**: MEN (3000 pairs), SimLex-999 (999 pairs)

automobile	car	**0.50**
birds	mammals	**0.29**
airplane	market	**0.11**
...	...	...

- **Task**: Estimation of Word Semantic Similarity
- **Groundtruth Data**: MEN (3000 pairs), SimLex-999 (999 pairs)

automobile	car	**0.50**
birds	mammals	**0.29**
airplane	market	**0.11**
...	...	...

- **Evaluation procedure**
  - $\forall(w_1, w_2)$: **predict similarity scores**: $sim(w_1, w_2) = cos(\vec{r_1}, \vec{r_2})$
  - **Evaluation metric**: Spearman correlation coefficient

		GT	PRED
automobile	car	**0.50**	0.35
birds	mammals	**0.29**	0.42
airplane	market	**0.11**	0.28
...	...	...	..

Word Semantic Similarity Estimation via the ADSM

- **Experimental Dataset for ADSM**

Number of clips	**4474**	**Number of unique tags**	**940**
Min duration	0.1s	Avg tags per clip	8
Max duration	120s	Avg clips per tag	40
Avg duration	16.6s	Total number of tags	37203

Table: **Audio clips** & **tags** from the online search engine **FreeSound**.

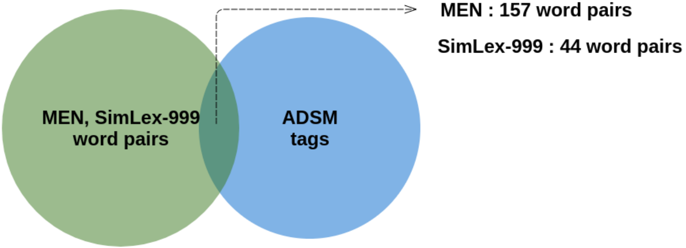
Word Semantic Similarity Estimation via the ADSM

- **Experimental Dataset for ADSM**

Number of clips	**4474**	**Number of unique tags**	**940**
Min duration	0.1s	Avg tags per clip	8
Max duration	120s	Avg clips per tag	40
Avg duration	16.6s	Total number of tags	37203

Table: **Audio clips** & **tags** from the online search engine **FreeSound**.

- **Evaluation Procedure for ADSM**



Word Semantic Similarity: Evaluating the ADSM

- Adding two **text models** as **evaluation datasets**[1]:
  - **CDSM**: state-of-the-art DSM (Iosif & Potamianos, LREC 2016)
  - **word2vec**: the word2vec model (Mikolov et al. 2013 a,b,c)

Dataset	MEN	SimLex-999	CDSM	word2vec
Word Pairs	157	44	1084	785

[1] Both CDSM and word2vec are used as evaluation datasets, because they have state-of-the-art performance and provide estimations for unlimited word pairs

Evaluating the ADSM

- **ADSM parameters**

Parameter	Description	Default Value
k	# audio words	300
SVD	SVD dimensions	- (no SVD)

- **Baseline ADSM** vs **Literature Results**[2]

Method	k	SVD	MEN	SimLex-999	CDSM	word2vec
[1]	100	60	0.402	0.233	n/a	n/a
[2]	300	-	0.325	0.161	n/a	n/a
Baseline	100	60	0.382	**0.302**	0.321	0.294
Baseline	300	-	**0.416**	0.235	**0.333**	**0.332**

Table: ADSM Accuracy, i.e., Spearmann Correlation.

[2] First row: A. Lopopolo and E. Miltenburg (2015)
Second row: D. Kiela and S.Clark (2016)

ADSM Evaluation: Fusion of Feature Spaces

- **Feature Spaces**: S_1: MFCCs, S_2: F0 feature, S_3: Music features
- **Classification (music, speech, other)**: SVM classifier (linear kernel)

Class	u_1	u_2	u_3
Music	0.3	0.2	0.5
Speech	0.8	0.2	0.0
Other	0.3	0.0	0.7

Feat. Space	k	SVD	MEN	SimLex-999	CDSM	word2vec
S_1	300	-	0.416	0.235	0.333	0.332
S_2			0.308	0.313	0.269	0.248
S_3			0.418	0.205	0.278	0.315
S_{123}			**0.468**	**0.387**	**0.388**	**0.382**
S_1		90	0.436	0.209	0.283	0.320
S_2			0.302	0.34	0.275	0.26
S_3			0.422	0.252	0.343	0.337
S_{123}			**0.480**	**0.374**	**0.402**	**0.401**
S_1	400	-	0.457	0.24	0.298	0.309
S_2			0.304	0.334	0.283	0.259
S_3			0.423	0.300	0.384	0.343
S_{123}			**0.462**	**0.437**	**0.404**	**0.379**
S_1		90	0.427	0.317	0.375	0.331
S_2			0.314	0.351	0.278	0.254
S_3			0.46	0.225	0.293	0.302
S_{123}			**0.477**	**0.407**	**0.416**	**0.407**

Table: ADSM Accuracy, i.e., Spearmann Correlation.

Word Semantic Similarity: Multimodal Fusion

- **Multimodal Fusion**

Model	Dimensions	Train Data	Train Features
ADSM	300	FreeSound clips	MFCCs
DSM (CDSM)	300	English documents	-
VDSM	300	ESP-Game images	SIFT (HSV Space)

- **Evaluation**: keep the intersection of DSM, ADSM and VDSM tags (1613 unique tags)
- Addition of three evaluation datasets
  - **AMEN**: the **auditory relevant** subset of MEN (e.g. guitar-rock)
  - **TMEN**: the **text relevant** subset of MEN (complementary to AMEN)[3]
  - **ASLex**: the **auditory relevant** subset of SimLex-999

Dataset	MEN	AMEN	TMEN	SimLex-999	ASLex
Word Pairs	1533	141	135	207	100

[3] To provide equal comparisons, we random sample TMEN to obtain equal number of words as in AMEN. The final score is computed as the average score of 10 random samples.

Word Semantic Similarity: Multimodal Fusion

- Early Fusion:
  1. **Concatenation** of ADSM, DSM, VDSM representations[4]
  2. **Dimensionality reduction** to 300 dimensions using PCA
  3. Final score: **cosine similarity** between **multimodal** representations

Model	MEN	AMEN	TMEN	SimLex-999	ASLex
ADSM	0.433	0.554	0.532	0.352	0.292
DSM	0.774	0.762	**0.812**	0.427	0.398
VDSM	0.233	0.435	0.181	0.248	0.269
ADSM&DSM	**0.783**	0.815	0.759	0.475	0.424
ADSM&VDSM	0.470	0.632	0.438	0.401	0.348
DSM&VDSM	0.762	0.814	0.772	0.481	**0.497**
ADSM&DSM&VDSM	0.776	**0.827**	0.798	**0.502**	0.476

Table: Early Fusion - Spearmann Correlation

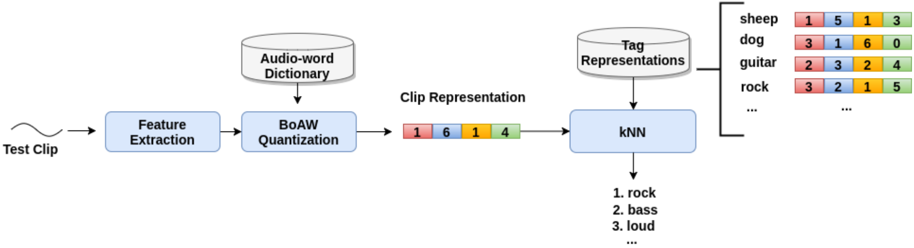
[4] Before concatenation, L2 normalization is performed

Outline

ADSM Applications

Audio auto-tagging

- **Task**: Audio auto-tagging, i.e., predict multiple labels from audio
- **Applications**: Indexing & Querying Music Collections
- **Auto-tagging using ADSM**:



- **Experimentation Dataset**: MagnaTagATune
  - 25,863 audio clips (mostly music) of 30s duration
  - 188 unique tags
- **Acoustic Features for ADSM**
  - **EchoNest**:
    - 12 chromagram features
    - 12 timbre (MFCC-like) features
  - **MFCCdd**:
    - 13 MFCCs, first and second order derivatives

ADSM Application: Auto-tagging

Examples

Clip id	Groundtruth Tags	Predicted Tags
3843	**indian**, **sitar**	**sitar**, **indian**, eastern, india, oriental
13526	bass, **drums**, drum, **funky**, **reggae**	**funky**, beat, **drums**, **reggae**, funk
15380	**classical**, **solo**, **cello**, **violin**, strings	**cello**, viola, **violin**, **solo**, **classical**
19920	-	orchestra, violins, flutes, fiddle, violin
21725	choir, **choral**, **men**, man	monks, chant, chanting, **men**, **choral**
29231	**acoustic**, **guitar**	classical guitar, **guitar**, **acoustic**, lute, spanish
43390	**rock**, loud, **pop**, vocals, **male vocals**	**male vocals**, **pop**, male vocal, male singer, **rock**
48010	silence	low, soft, no singing, quiet, wind
57081	**piano**	piano solo, **piano**, classic, solo, classical

Table: Examples of auto-tagging outputs for MagnaTagATune clips. Number of predicted tags: $N = 5$.

ADSM Application: Auto-tagging

Visualization of tag representations using t-SNE

ADSM Application: Auto-tagging

Visualization of tag representations using t-SNE

ADSM Application: Auto-tagging

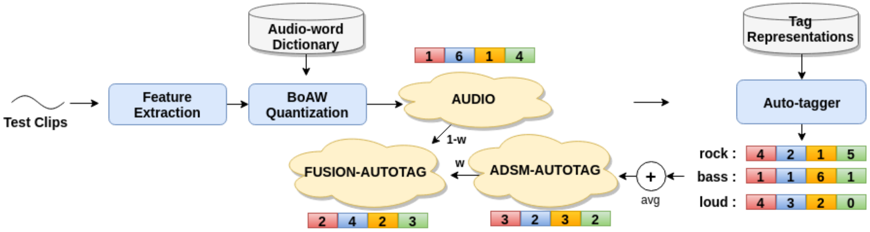
Visualization of tag representations using t-SNE

ADSM Application: Auto-tagging

Visualization of tag representations using t-SNE

Outline

- **Music Similarity**: The core of Music Recommendation and Music Information Retrieval
- **ADSM for Music Similarity**: Combine tags and audio



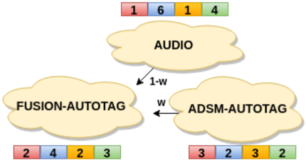
ADSM Application: Music Similarity

- Music Similarity Estimation is **subjective**
- **Relative similarity**:
  - Given songs (a, b, c): "Which is the most irrelevant (odd) song?"
  - c irrelevant $\Rightarrow$ $sim(c, a) < sim(b, a)$ and $sim(c, b) < sim(a, b)$
  - similarity constraints = distance constraints
- **Groundtruth similarity data**
  - Collected from a "odd one out" game
  - 860 Triplets of songs (a, b, c), where c is the "odd" song
- **Evaluation Metric**: % constraints satisfied by the algorithm
- **Experimentation dataset**: MagnaTagATune

ADSM Application: Music Similarity

Literature Method	EchoNest Features
Euclidean	0.598
RITML	0.711
SVM	**0.712**
MLR	0.689

Table: Literature methods.



Proposed Method	EchoNest		MFCCdd	
	k=300	svd=10	k=300	svd=10
AUDIO	0.613	0.644	0.636	0.646
ADSM-REALTAG	0.705	0.719	**0.717**	**0.720**
FUSION-REALTAG	**0.720**	**0.731**	0.681	0.684
ADSM-AUTOTAG	0.705	0.705	0.693	0.696
FUSION-AUTOTAG	0.705	0.709	0.662	0.672

Table: Proposed methods.

Conclusions

Multimodal DSMs - Grounding to the auditory and visual modalities

- **Multimodal Fusion for Word Semantic Similarity**: Higher correlation with human ratings compared to unimodal representations. First attempt to fuse text, visual and acoustic features for multimodal word representations.
- **Dimensionality Reduction** can give significant improvements for Word Semantic Similarity and Music Similarity
- **Fusion of Feature Spaces** outperforms the baseline ADSM
- **Soft Encoding** did not outperform Hard Encoding
- **ADSM for Auto-tagging**: Satisfactory performance. Can be used to tag unknown clips or enrich provided annotations
- **ADSM for Music Similarity**: Combine audio and tags (groundtruth or predicted) for getting better estimations in an unsupervised way

Future Work

- **ADSM Pipeline**
  - Audio-word Dictionary: Replace k-means with a **dictionary learning** algorithm (e.g. k-SVD)
  - Perform **audio segmentation** before building the ADSM
  - **Hard Encoding** vs **Soft Encoding**. Test assertion: clustering samples $\uparrow \Rightarrow$ Hard Encoding=Soft Encoding
- **Multimodal Fusion**: Early, Middle and Late Fusion
  - Early Fusion: Learn **simultaneously** text, image and audio representations (e.g. multimodal skip-gram)
  - Use the **auditory/visual relevance** of words to **weight the contribution** of ADSM and VDSMs to the multimodal representation
- **Zero-shot learning via cross-modal mapping**: e.g. use multimodal Deep Boltzmann Machines
- Apply for **Audio/Video tasks** (e.g. Multimedia Event Detection)
- **Build deep neural auditory embeddings**
  - end2end learning: CNN (or CNN-LSTM) on log mel spectrogram
- **BoAW** approach ignores temporal order $\rightarrow$ train LSTMs

Thank You!

Backup slides

Word Semantic Similarity: Multimodal Fusion

Late Fusion

- Late Fusion:
  1. Computation of **cosine similarity** separately for ADSM, DSM, VDSM
  2. Final score: Fusion (average) of similarity scores

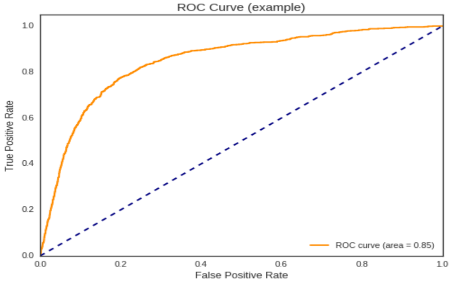
Model	**MEN**	**AMEN**	**TMEN**	**SimLex-999**	**ASLex**
ADSM	0.433	0.554	0.532	0.352	0.292
DSM	**0.774**	0.762	**0.812**	0.427	0.398
VDSM	0.233	0.435	0.181	0.248	0.269
ADSM&DSM	0.741	0.719	0.718	0.406	0.317
ADSM&VDSM	0.474	0.635	0.428	0.405	0.340
DSM&VDSM	0.762	**0.814**	0.737	**0.478**	**0.492**
ADSM&DSM&VDSM	0.459	0.639	0.308	0.403	0.345

Table: Late Fusion - Spearmann Correlation

ADSM Application: Auto-tagging

Evaluation

- Auto-tagging as **Multi-label classification** (tags = labels)
- **Evaluation metric**: AUC-ROC (Area Under ROC curve)



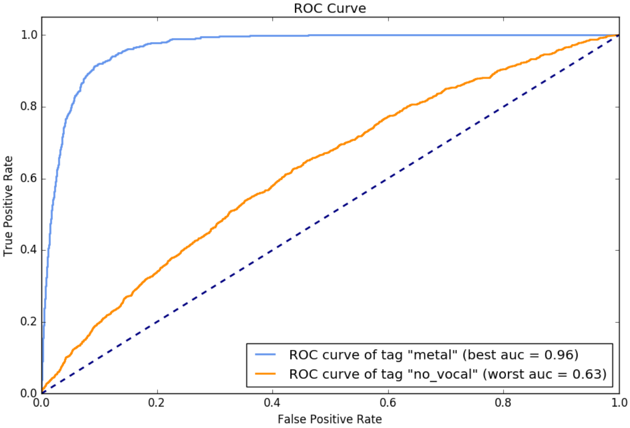
- **ADSM Evaluation Procedure**
  1. Compute **AUC** for each tag
  2. Final Score: **Avg AUC** over the tags

ADSM Application: Auto-tagging

Evaluation

k	EchoNest	MFCCdd
300	0.809	0.806

Table: Avg AUC



ADSM Application: Auto-tagging

Evaluation

tag	AUC	tag	AUC	tag	AUC	tag	AUC
metal	**0.965**	loud	0.868	male	0.795	female vocal	0.755
choral	**0.946**	techno	0.862	male vocal	0.786	vocal	0.747
choir	**0.942**	country	0.847	guitar	0.786	female voice	0.747
opera	**0.941**	piano	0.838	electronic	0.786	synth	0.738
rock	**0.927**	classical	0.828	male voice	0.783	weird	0.737
harp	**0.906**	pop	0.827	ambient	0.775	vocals	0.733
harpsichord	**0.900**	solo	0.826	soft	0.773	voice	0.728
cello	0.897	classic	0.823	fast	0.768	slow	0.727
dance	0.891	quiet	0.823	indian	0.767	**no voice**	**0.642**
beats	0.881	sitar	0.821	singing	0.766	**no vocals**	**0.629**
beat	0.877	drums	0.818	woman	0.757	**no vocal**	**0.626**
flute	0.875	man	0.808	female	0.757		
violin	0.870	strings	0.799	new age	0.755		

Table: 50 MagnaTagATune tags sorted by AUC