Word2vec and beyond

presented by Eleni Triantafillou

March 1, 2016

The Big Picture

There is a long history of word representations

- Techniques from information retrieval: Latent Semantic Analysis (LSA)
- Self-Organizing Maps (SOM)
- Distributional count-based methods
- Neural Language Models

Important take-aways:

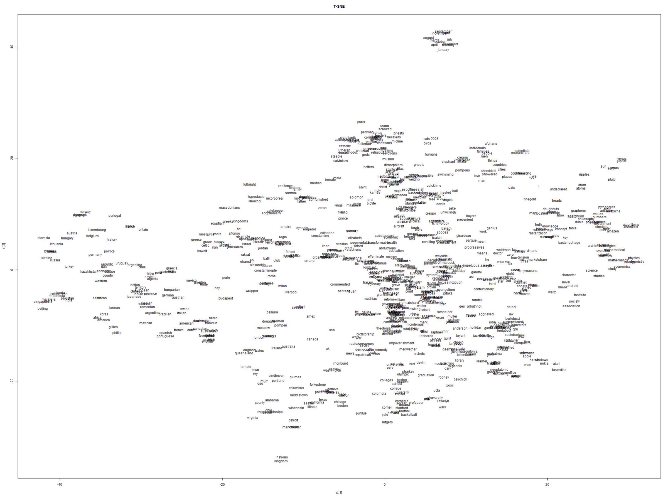
1. Don't need deep models to get good embeddings
2. Count-based models and neural net predictive models are not qualitatively different

source:
http://gavagai.se/blog/2015/09/30/a-brief-history-of-word-embeddings/

Continuous Word Representations

- Contrast with simple n-gram models (words as atomic units)
- Simple models have the potential to perform very well...
- ... if we had enough data
- Need more complicated models
- Continuous representations take better advantage of data by modelling the similarity between the words

Continuous Representations



source: http://www.codeproject.com/Tips/788739/Visualization-of-High-Dimensional-Data-using-t-SNE

Skip Gram

- Learn to predict surrounding words
- Use a large training corpus to maximize:

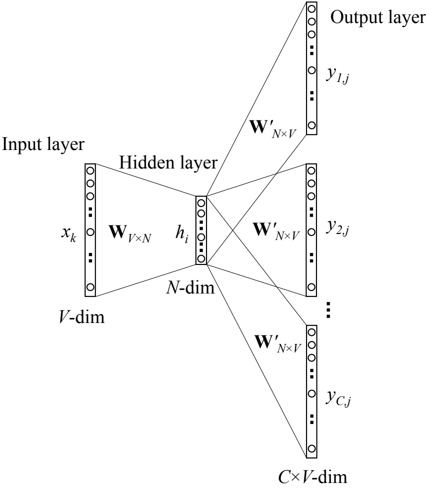
$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,\ j\neq 0}\log p(w_{t+j}|w_t)$$

where:

- T: training set size
- c: context size
- w_j: vector representation of the j_{th} word

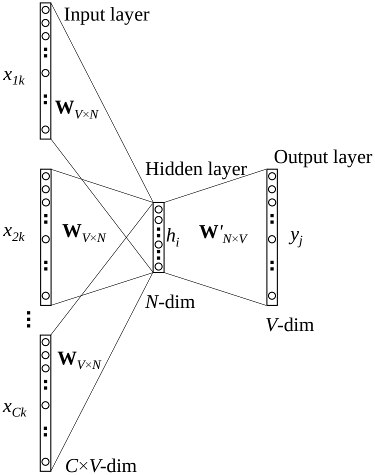
Skip Gram: Think of it as a Neural Network

Learn W and W' in order to maximize previous objective



source: "word2vec parameter learning explained." ([4])

CBOW



source: "word2vec parameter learning explained." ([4])

word2vec Experiments

- Evaluate how well syntactic/semantic word relationships are captured
- Understand effect of increasing training size / dimensionality
- Microsoft Research Sentence Completion Challenge

Semantic / Syntactic Word Relationships Task

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Semantic / Syntactic Word Relationships Results

Table 4: *Comparison of publicly available word vectors on the Semantic-Syntactic Word Relationship test set, and word vectors from our models. Full vocabularies are used.*

Model	Vector Dimensionality	Training words	Accuracy [%]		
			Semantic	Syntactic	Total
Collobert-Weston NNLM	50	660M	9.3	12.3	11.0
Turian NNLM	50	37M	1.4	2.6	2.1
Turian NNLM	200	37M	1.4	2.2	1.8
Mnih NNLM	50	37M	1.8	9.1	5.8
Mnih NNLM	100	37M	3.3	13.2	8.8
Mikolov RNNLM	80	320M	4.9	18.4	12.7
Mikolov RNNLM	640	320M	8.6	36.5	24.6
Huang NNLM	50	990M	13.3	11.6	12.3
Our NNLM	20	6B	12.9	26.4	20.3
Our NNLM	50	6B	27.9	55.8	43.2
Our NNLM	100	6B	34.2	**64.5**	50.8
CBOW	300	783M	15.5	53.1	36.1
Skip-gram	300	783M	**50.0**	55.9	**53.3**

Learned Relationships

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Microsoft Research Sentence Completion

Table 7: *Comparison and combination of models on the Microsoft Sentence Completion Challenge.*

Architecture	Accuracy [%]
4-gram [32]	39
Average LSA similarity [32]	49
Log-bilinear model [24]	54.8
RNNLMs [19]	55.4
Skip-gram	48.0
Skip-gram + RNNLMs	**58.9**

Linguistic Regularities

- "king" - "man" + "woman" = "queen"!
- Demo
- Check out gensim (python library for topic modelling): https://radimrehurek.com/gensim/models/word2vec.html

Multimodal Word Embeddings: Motivation

Are these two objects similar?

Multimodal Word Embeddings: Motivation

And these?

Multimodal Word Embeddings: Motivation

What do you think should be the case?

sim( , ) $<$ sim( , ) ?

or

sim( , ) $>$ sim( , ) ?

When do we need image features?

It's surely task-specific. In many cases can benefit from visual features!

- Text-based Image Retrieval
- Visual Paraphrasing
- Common Sense Assertion Classification
- They are better-suited for zero shot learning (learn mapping between text and images)

Two Multimodal Word Embeddings approaches...

1. Combining Language and Vision with a Multimodal Skip-gram Model (Lazaridou et al, 2013)
2. Visual Word2Vec (vis-w2v): Learning Visually Grounded Word Embeddings Using Abstract Scenes (Kottur et al, 2015)

Two Multimodal Word Embeddings approaches...

1. **Combining Language and Vision with a Multimodal Skip-gram Model (Lazaridou et al, 2013)**
2. Visual Word2Vec (vis-w2v): Learning Visually Grounded Word Embeddings Using Abstract Scenes (Kottur et al, 2015)

Multimodal Skip-Gram

- **The main idea**: Use visual features for the (very) small subset of the training data for which images are available.
- Visual vectors are obtained by CNN and are fixed during training!
- Recall, Skip-Gram objective:

$$L_{ling}(w_t) = \sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log(p(w_{t+j} | w_t))$$

- New Multimodal Skip-Gram objective:

$$L = \frac{1}{T} \sum_{t=1}^{T} (L_{ling}(w_t) + L_{vision}(w_t)),$$

where

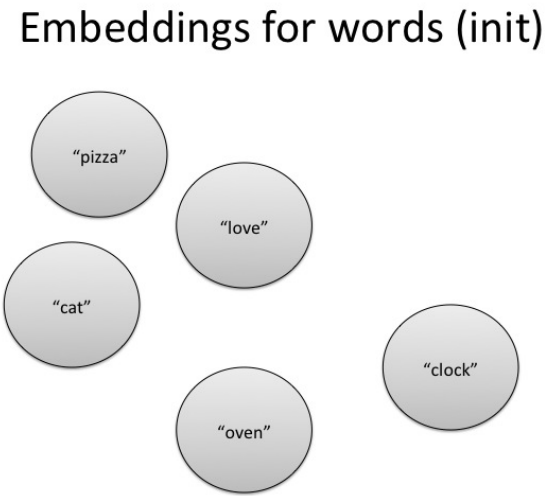
- $L_{vision}(w_t) = 0$ if w_t does not have an entry in ImageNet, and otherwise
- $L_{vision}(w_t) =$
$$- \sum_{w' \sim P(w)} \max(0, \gamma - cos(u_{w_t}, v_{w_t}) + cos(u_{w_t}, v_{w'}))$$

Multimodal Skip-Gram: An example

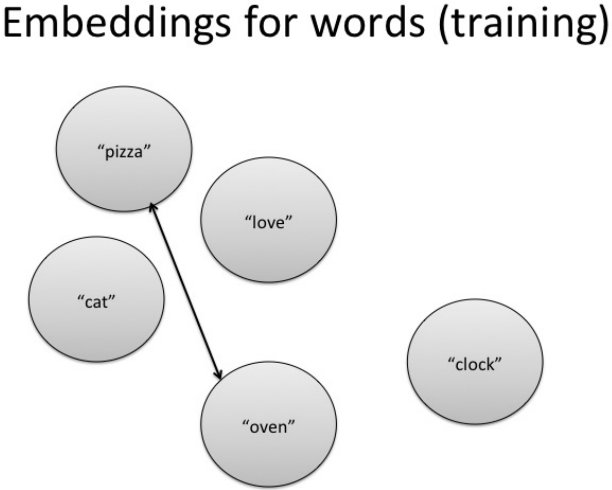
Training Set

Words	Image Available?
pizza	yes
cat	yes
clock	yes
love	no
oven	no

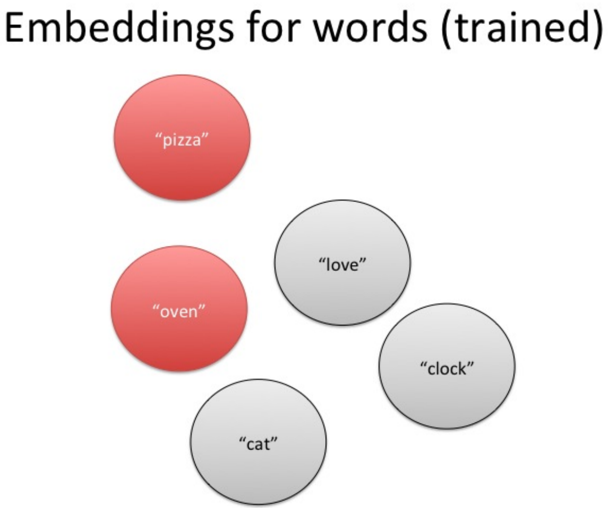
Multimodal Skip-Gram: An example



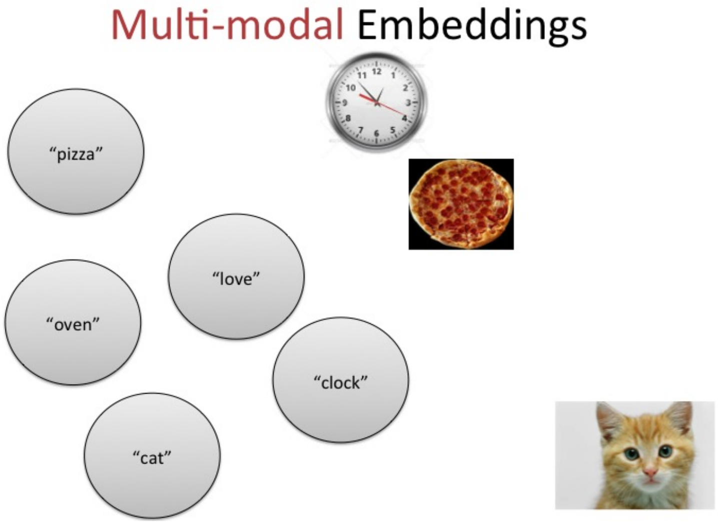
Multimodal Skip-Gram: An example



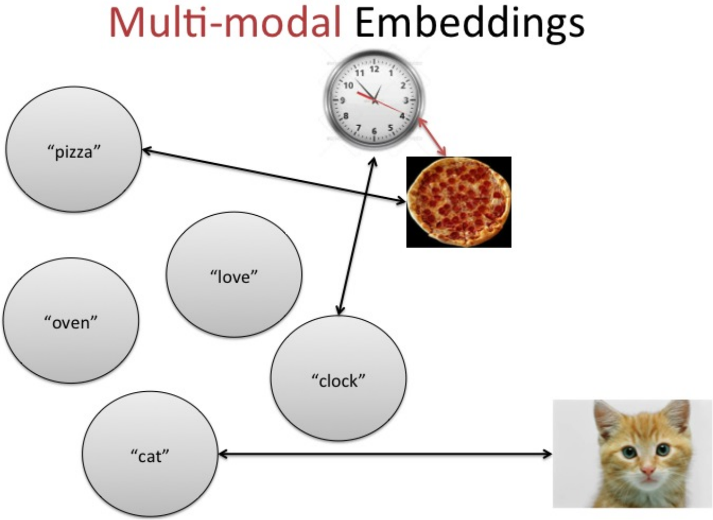
Multimodal Skip-Gram: An example



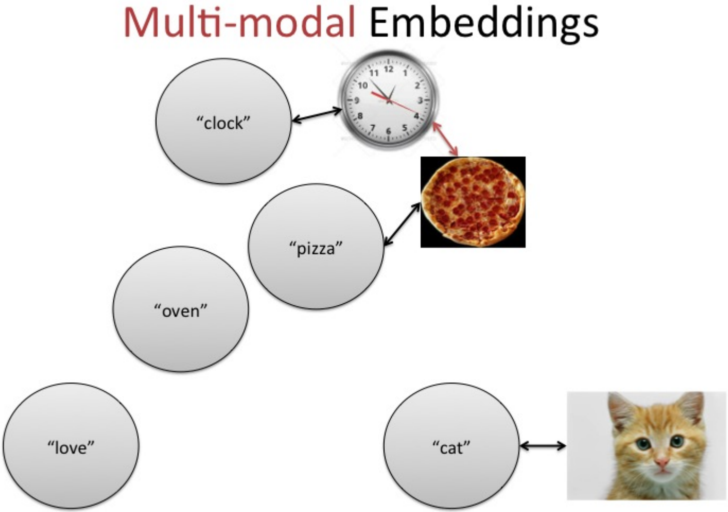
Multimodal Skip-Gram: An example



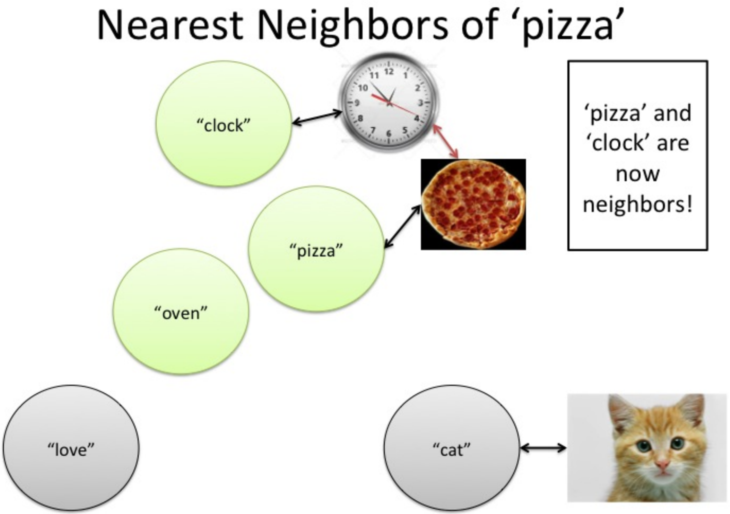
Multimodal Skip-Gram: An example



Multimodal Skip-Gram: An example



Multimodal Skip-Gram: An example



Multimodal Skip-Gram: Comparing to Human Judgements

Model	*MEN*		*Simlex-999*		*SemSim*		*VisSim*	
	100%	42%	100%	29%	100%	85%	100%	85%
Kiela and Bottou	-	0.74	-	0.33	-	0.60	-	0.50
Bruni et al.	-	0.77	-	0.44	-	0.69	-	0.56
Silberer and Lapata	-	-	-	-	0.70	-	0.64	-
CNN features	-	0.62	-	0.54	-	0.55	-	0.56
Skip-gram	0.70	0.68	0.33	0.29	0.62	0.62	0.48	0.48
Concatenation	-	0.74	-	0.46	-	0.68	-	0.60
SVD	0.61	0.74	0.28	0.46	0.65	0.68	0.58	0.60
MMSkip-gram-A	0.75	0.74	0.37	0.50	0.72	0.72	0.63	0.63
MMSkip-gram-B	0.74	0.76	0.40	0.53	0.66	0.68	0.60	0.60

MEN: general relatedness ("pickles", "hamburgers"), **Simplex-999**: taxonomic similarity ("pickles", "food"), **SemSim**: Semantic similarity ("pickles", "onions"), **VisSim**: Visual Similarity ("pen", "screwdriver")

Multimodal Skip-Gram: Examples of Nearest Neighbors

Only "donut" and "owl" trained with direct visual information.

Target	SKIP-GRAM	MMSKIP-GRAM-A	MMSKIP-GRAM-B
donut	fridge, diner, candy	pizza, sushi, sandwich	pizza, sushi, sandwich
owl	pheasant, woodpecker, squirrel	eagle, woodpecker, falcon	eagle, falcon, hawk
mural	sculpture, painting, portrait	painting, portrait, sculpture	painting, portrait, sculpture
tobacco	coffee, cigarette, corn	cigarette, cigar, corn	cigarette, cigar, smoking
depth	size, bottom, meter	sea, underwater, level	sea, size, underwater
chaos	anarchy, despair, demon	demon, anarchy, destruction	demon, anarchy, shadow

Multimodal Skip-Gram: Zero-shot image labelling and image retrieval

	P@1	P@2	P@10	P@20	P@50
SKIP-GRAM	1.5	2.6	14.2	23.5	36.1
MMSKIP-GRAM-A	2.1	3.7	16.7	24.6	37.6
MMSKIP-GRAM-B	2.2	5.1	20.2	28.5	43.5

Table 3: Percentage precision@k results in the zero-shot image labeling task.

	P@1	P@2	P@10	P@20	P@50
SKIP-GRAM	1.9	3.3	11.5	18.5	30.4
MMSKIP-GRAM-A	1.9	3.2	13.9	20.2	33.6
MMSKIP-GRAM-B	1.9	3.8	13.2	22.5	38.3

Table 4: Percentage precision@k results in the zero-shot image retrieval task.

Multimodal Skip-Gram: Survey to evaluate on Abstract Words

Metric: Proportion (percentage) of words for which number votes in favour of "neighbour" image significantly above chance.
Unseen: Discard words for which visual info was accessible during training.

	global	*|words|*	*unseen*	*|words|*
all	48%	198	30%	127
concrete	73%	99	53%	30
abstract	23%	99	23%	97

Multimodal Skip-Gram: Survey to evaluate on Abstract Words

Left: subject preferred the nearest neighbour to the random image

freedom

theory

wrong

god

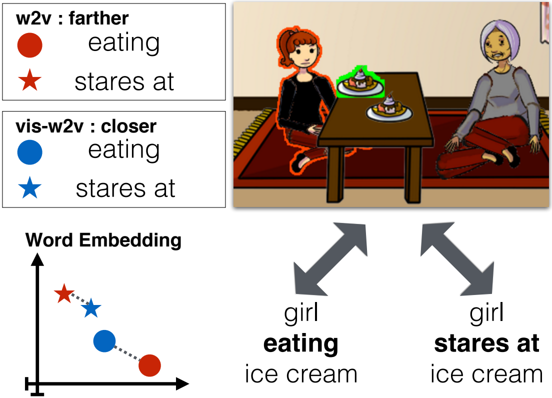
together

place

Two Multimodal Word Embeddings approaches...

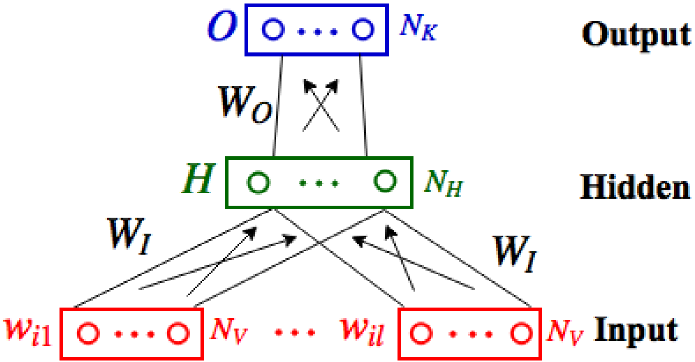
1. Combining Language and Vision with a Multimodal Skip-gram Model (Lazaridou et al, 2013)
2. **Visual Word2Vec (vis-w2v): Learning Visually Grounded Word Embeddings Using Abstract Scenes (Kottur et al, 2015)**

Visual Word2Vec (vis-w2v): Motivation



Visual Word2Vec (vis-w2v): Approach

- Multimodal train set: tuples of (description, abstract scene)
- **Finetune** word2vec to add visual features obtained by abstract scenes (clipart)
- Obtain surrogate (visual) classes by clustering those features
- W_I: initialized from word2vec
- N_K: number of clusters of abstract scene features



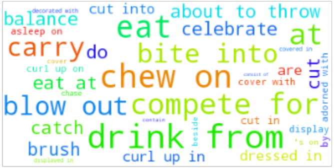
Clustering abstract scenes

Interestingly, "prepare to cut", "hold", "give" are clustered together with "stare at" etc. It would be hard to infer these semantic relationships from text alone.

lay next to

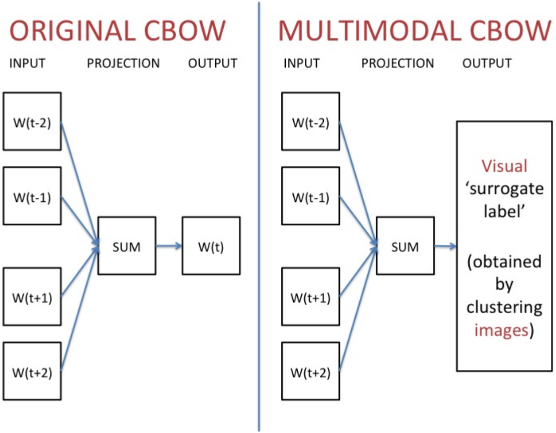
stand near

enjoy



stare at

Visual Word2Vec (vis-w2v): Relationship to CBOW (word2vec)



Surrogate labels play the role of *visual context*.

Visual Word2Vec (vis-w2v): Visual Paraphrasing Results

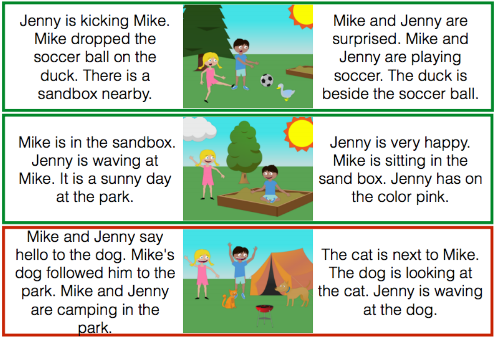


Figure 5: The visual paraphrasing task is to identify if two textual descriptions are paraphrases of each other. Shown above are three positive instances, *i.e.*, the descriptions (left, right) actually talk about the same scene (center). Green boxes show two cases where `vis-w2v` correctly predicts and `w2v` does not, while red box shows the case where both `vis-w2v` and `w2v` predict incorrectly. Note that the red instance is tough as the textual descriptions do not intuitively seem to be talking about the same scene, even for a human reader.

Visual Word2Vec (vis-w2v): Visual Paraphrasing Results

Approach	Visual Paraphrasing AP (%)
w2v-wiki	94.1
w2v-wiki	94.4
w2v-coco	94.6
vis-w2v-wiki	95.1
vis-w2v-coco	**95.3**

Table: Performance on visual paraphrasing task

Visual Word2Vec (vis-w2v): Common Sense Assertion Classification Results

Given a tuple (Primary Object, Relation, Secondary Object), decide if it is plausible or not.

Approach	common sense AP (%)
w2v-coco	72.2
w2v-wiki	68.1
w2v-coco + vision	73.6
vis-w2v-coco (shared)	74.5
vis-w2v-coco (shared) + vision	74.2
vis-w2v-coco (separate)	**74.8**
vis-w2v-coco (separate) + vision	**75.2**
vis-w2v-wiki (shared)	72.2
vis-w2v-wiki (separate)	74.2

Table: Performance on the common sense task

Thank you!

[-0.0665592 -0.0431451 ... -0.05182673 -0.07418852 -0.04472357 0.02315103 -0.04419742 -0.01104935]

[0.08773034 0.00566679 ... 0.03735885 -0.04323553 0.02130294 -0.09108844 -0.05708769 0.04659363]

Bibliography

- Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
- Kottur, Satwik, et al. "Visual Word2Vec (vis-w2v): Learning Visually Grounded Word Embeddings Using Abstract Scenes." arXiv preprint arXiv:1511.07067 (2015).
- Lazaridou, Angeliki, Nghia The Pham, and Marco Baroni. "Combining language and vision with a multimodal skip-gram model." arXiv preprint arXiv:1501.02598 (2015).
- Rong, Xin. "word2vec parameter learning explained." arXiv preprint arXiv:1411.2738 (2014).
- Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.