Learning to Ground Meaning in the Visual World

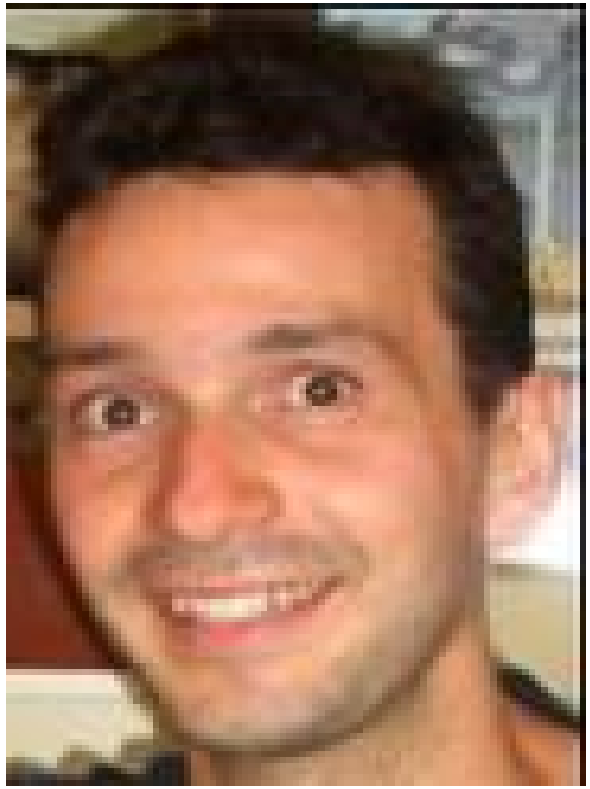
Mirella Lapata

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

UW MSR Summer Institute 2013

Joint work with

Carina Silberer



Vittorio Ferrari

Part I

Motivation

How Do We Learn the Meaning of Words?

How Do We Learn the Meaning of Words?

- Through language (e.g., parents talking to children)

How Do We Learn the Meaning of Words?

- Through language (e.g., parents talking to children)
- Through perception (e.g., visual, olfactory, haptic experience)

How Do We Learn the Meaning of Words?

- Through language (e.g., parents talking to children)
- Through perception (e.g., visual, olfactory, haptic experience)
- Through the physical world (e.g., through movement)

How Do We Learn the Meaning of Words?

- Through language (e.g., parents talking to children)
- Through perception (e.g., visual, olfactory, haptic experience)
- Through the physical world (e.g., through movement)
- Through social interaction (e.g., inferring speakers intentions)

Potter et al. (1986); Landau et al. (1998); Gernsbacher et al. (1990); Quinn et al. (1993); Schyns and Rodet (1997); Jones et al. (1991).

How Do We Describe the Meaning of words?

How Do We Describe the Meaning of words?

A Moose

How Do We Describe the Meaning of words?

A Moose

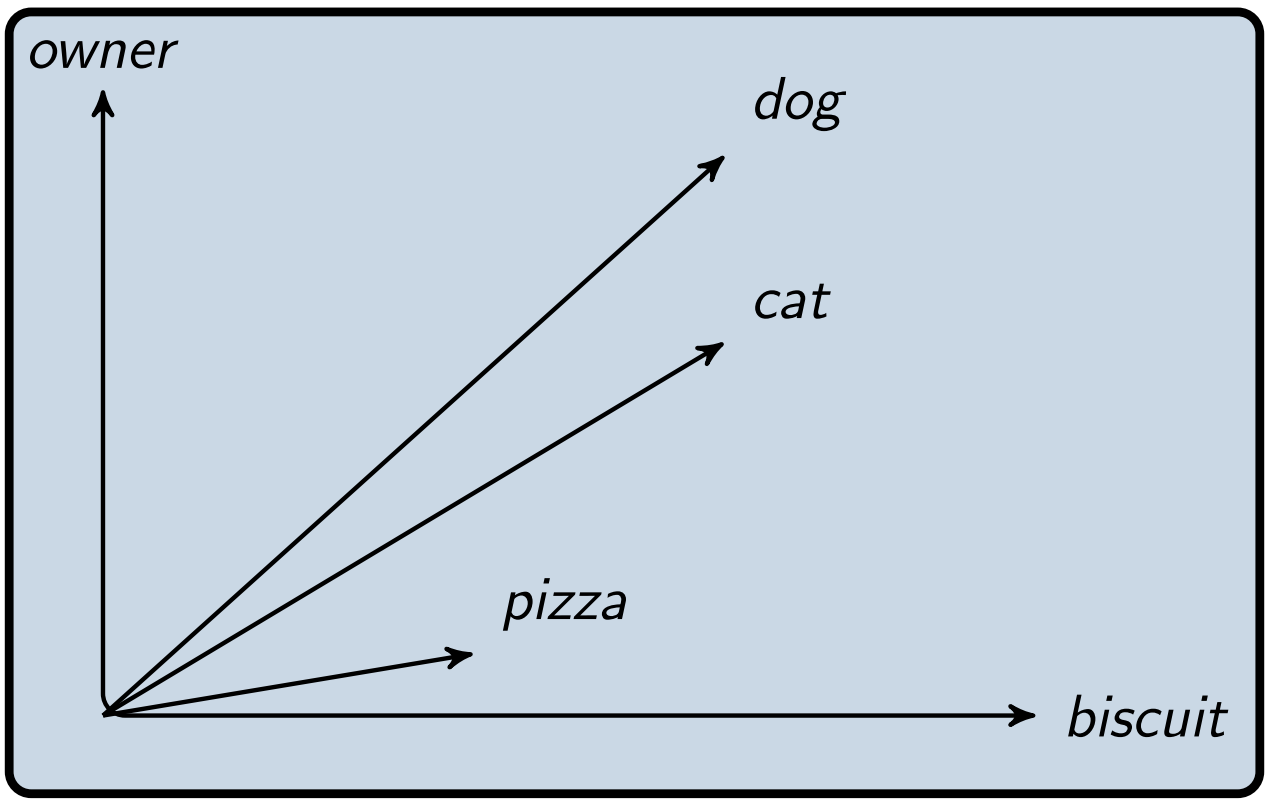
Feature	Freq	Classification
is_large	27	visual
has_antlers	23	visual
has_legs	14	visual
has_fur	7	visual
is_brown	10	visual
has_hooves	5	visual
eaten_as_meat	5	function
lives_in_woods	14	encyclopedic
an_animal	17	taxonomic
a_mammal	9	taxonomic

Feature norms from McRae et al. (2005).

How Do We Model the Meaning of Words?

Words are represented through their relations to other words.

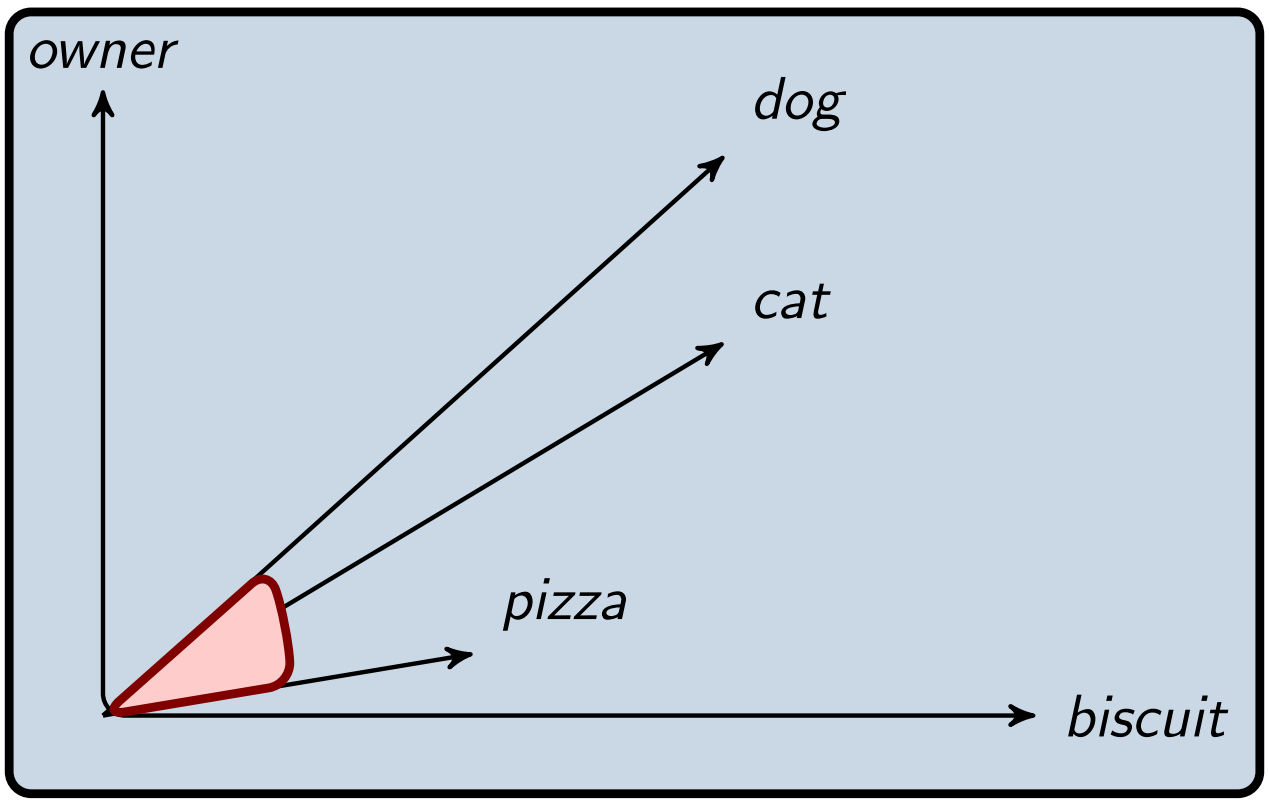
Latent Semantic Analysis (LSA; Landauer and Dumais, 1997)
Hyperspace Analogue to Language (HAL; Lund and Burges, 1996)
Latent Dirichlet Allocation (LDA; Griffiths et al., 2007)
Neural Language Model (NLM; Collobert and Weston 2008)
Hierarchical Distributed Language Model (HLBL; Mnih and Hinton, 2008)



How Do We Model the Meaning of Words?

Words are represented through their relations to other words.

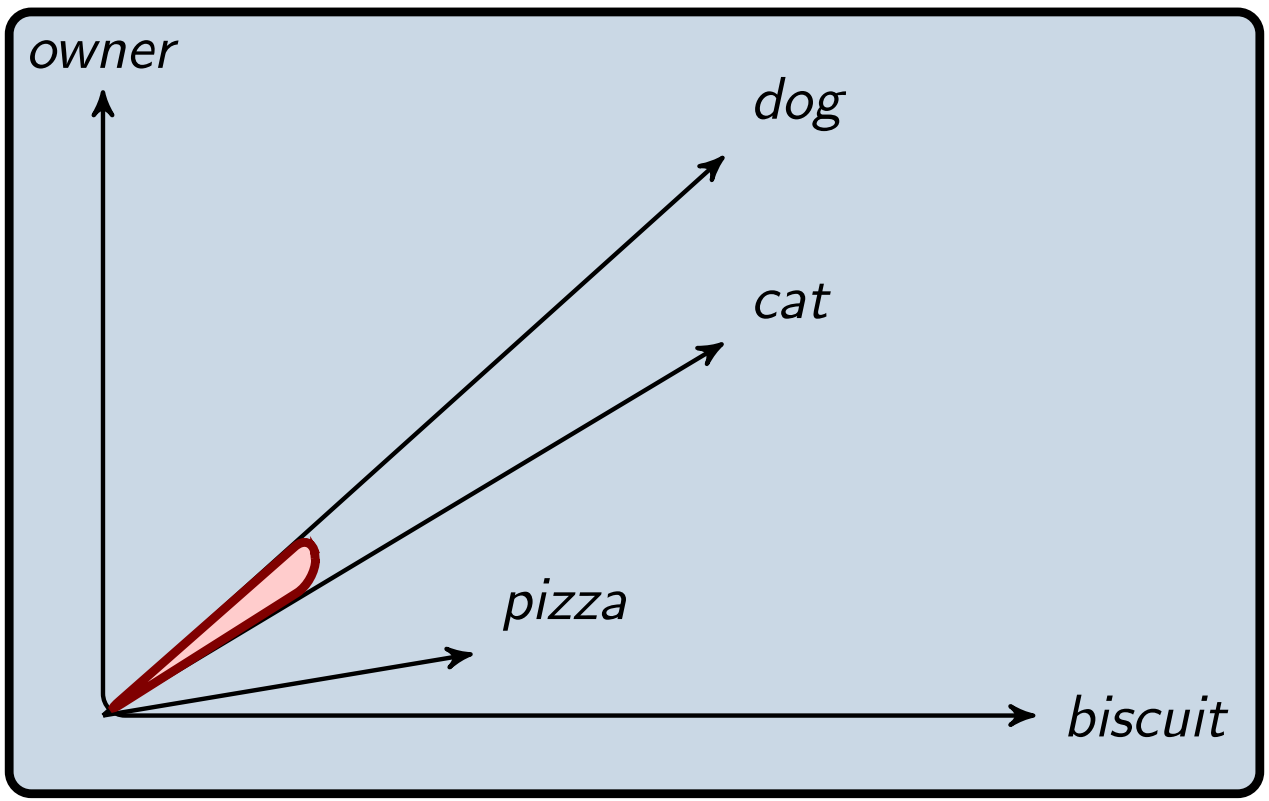
Latent Semantic Analysis (LSA; Landauer and Dumais, 1997)
Hyperspace Analogue to Language (HAL; Lund and Burges, 1996)
Latent Dirichlet Allocation (LDA; Griffiths et al., 2007)
Neural Language Model (NLM; Collobert and Weston 2008)
Hierarchical Distributed Language Model (HLBL; Mnih and Hinton, 2008)



How Do We Model the Meaning of Words?

Words are represented through their relations to other words.

Latent Semantic Analysis (LSA; Landauer and Dumais, 1997)
Hyperspace Analogue to Language (HAL; Lund and Burges, 1996)
Latent Dirichlet Allocation (LDA; Griffiths et al., 2007)
Neural Language Model (NLM; Collobert and Weston 2008)
Hierarchical Distributed Language Model (HLBL; Mnih and Hinton, 2008)



How Do We Model the Meaning of Words?

- Through language (e.g., parents talking to children)
- ~~Through perception (e.g., visual, olfactory, haptic experience)~~
- ~~Through the physical world (e.g., through movement)~~
- ~~Through social interaction (e.g., inferring speakers intentions)~~

How Do We Model the Meaning of Words?

- Through language (e.g., parents talking to children)
- ~~Through perception (e.g., visual, olfactory, haptic experience)~~
- ~~Through the physical world (e.g., through movement)~~
- ~~Through social interaction (e.g., inferring speakers intentions)~~

Learning meaning by listening to the radio (Elman, 1990).

Grounding Meaning in Perception

Q_1: How should we represent perceptual information?

Grounding Meaning in Perception

Q_1: How should we represent perceptual information?

Via images (Feng and Lapata, 2010; Bruni et al., 2011), image labels (Bruni et al., 2012), or feature norms (Andrews et al., 2009; Johns & Jones, 2012; Silberer and Lapata, 2012).

Grounding Meaning in Perception

Q_1: How should we represent perceptual information?

Q_2: How do we integrate perceptual and textual information?

Grounding Meaning in Perception

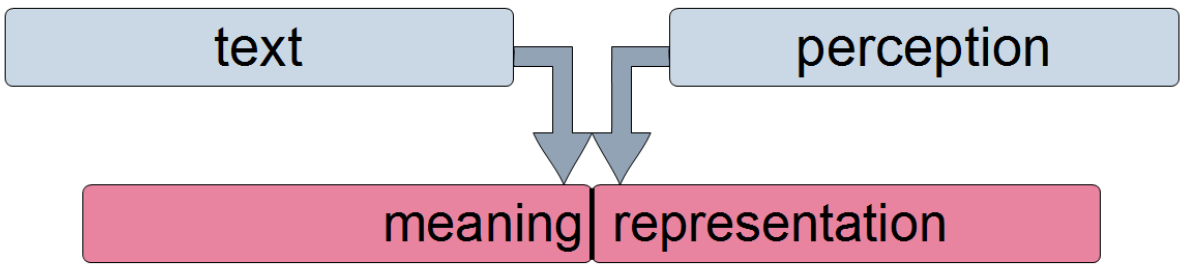
Q_1: How should we represent perceptual information?

Q_2: How do we integrate perceptual and textual information?

Grounding Meaning in Perception

Q_1: How should we represent perceptual information?

Q_2: How do we integrate perceptual and textual information?



Grounding Meaning in Perception

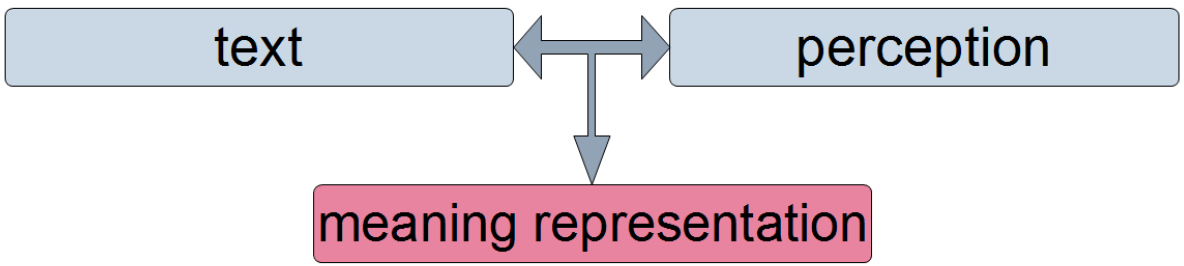
Q_1: How should we represent perceptual information?

Q_2: How do we integrate perceptual and textual information?

Grounding Meaning in Perception

Q_1: How should we represent perceptual information?

Q_2: How do we integrate perceptual and textual information?



Grounding Meaning in Perception

Q_1: How should we represent perceptual information?

Q_2: How do we integrate perceptual and textual information?

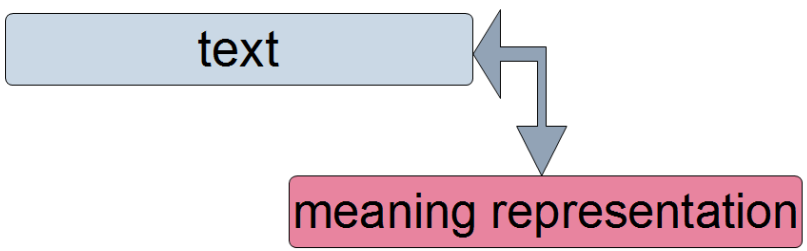
Q_3: Are two modalities better than one?

Grounding Meaning in Perception

Q_1: How should we represent perceptual information?

Q_2: How do we integrate perceptual and textual information?

Q_3: Are two modalities better than one?

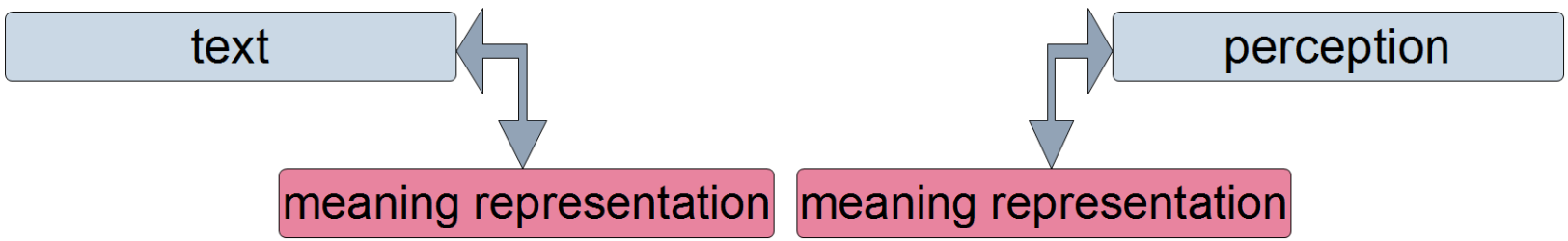


Grounding Meaning in Perception

Q_1: How should we represent perceptual information?

Q_2: How do we integrate perceptual and textual information?

Q_3: Are two modalities better than one?



Grounding Meaning in Perception

Q_1: How should we represent perceptual information?

Q_2: How do we integrate perceptual and textual information?

Q_3: Are two modalities better than one?

Q_4: How do we deal with unseen information?

Grounding Meaning in Perception

Q_1: How should we represent perceptual information?

Q_2: How do we integrate perceptual and textual information?

Q_3: Are two modalities better than one?

Q_4: How do we deal with unseen information?

Q_5: How do we measure success?

This Talk

1. We represent images using words
2. We use visual attributes as an approximation of perceptual information
3. We use autoencoders for integrating perceptual and linguistic representations

Part II

Perception via Visual Attributes

Ferrari and Zisserman, 2007; Lampert et al., 2009; Farhadi et al., 2009; Berg et al., 2010; Patterson and Hays, 2012

Visual Attribute Dataset

- Create dataset of images and attributes, train attribute classifiers, make predictions (Farhadi et al., 2009)
- 688K images from ImageNet (Deng et al., 2009) covering all concepts of McRae et al. ($\approx$500 basic nouns)
- Taxonomy of $>$400 visual attributes
- Start with McRae et al.'s visual features, modify and extend so that they are consistent and exhaustive
- Per-concept rather than per-image annotations
- 9,688 features based on color, texture, visual words, and edges

Visual Attribute Dataset (Example 1)

(Attribute listed in McRae et al.'s feature norms)

behavior	eats, walks, climbs, swims, runs
diet	drinks_water, eats_anything
shape_size	is_tall, is_large
anatomy	has_a_head, has_a_mouth, has_a_snout, has_jaws, has_teeth, has_a_tongue, has_a_nose, has_eyes, has_ears, has_a_neck, has_4_legs, has_feet, has_paws, has_claws, has_fur, has_a_tail
color_patterns	is_black, is_brown, is_white

Visual Attribute Dataset (Example 2)

(Attribute listed in McRae et al.'s feature norms)

behavior	rolls
parts	has_step_through_frame, has_fork, has_2_wheels, has_a_chain, has_pedals has_gears, has_handlebar, has_seat has_a_bell, has_brakes, has_spokes
texture_material	made_of_metal
color_patterns	is_black, is_red, is_grey, is_silver, different_colors

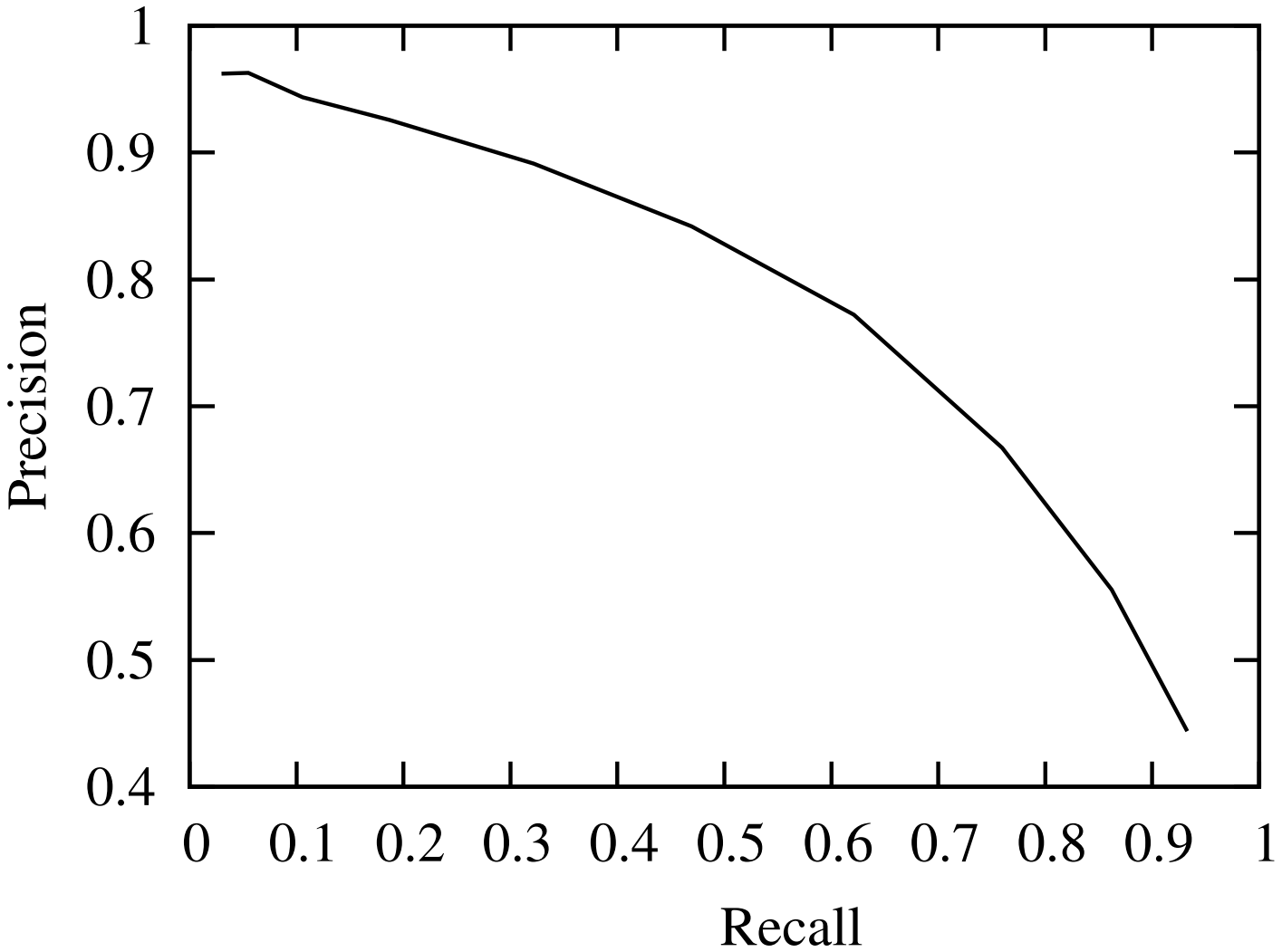
Predicted Visual Attributes (Examples)

has_2_pieces, has_a_pointed_end, has_a_strap, has_a_thumb, has_buckles, has_heels, has_shoe_laces, has_soles, is_black, is_brown, is_white, made_of_leather, made_of_rubber

climbs, climbs_trees, crawls, hops, jumps, eats, eats_nuts, is_small, has_a_bushy_tail, has_4_legs, has_a_head, has_a_neck, has_a_nose, has_a_snout, has_a_tail, has_claws, has_eyes, has_feet, has_toes

diff_colours, has_2_legs, has_2_wheels, has_windshield, has_floorboard, has_stand, has_tank, has_mudguard, has_seat, has_exhaust_pipe, has_frame, has_handlebar, has_lights, has_mirror, has_step-through_frame, is_black, is_blue, is_red, is_white, made_of_aluminum, made_of_steel

Attribute Classifier Performance: ROC Curve



Prediction scores for each attribute are thresholded (from 0 to 0.9).

Attribute Classifier Performance: Nearest Neighbors

BOAT	ROOSTER	SHIRT	SPINACH
ship	chicken	blouse	lettuce
sailboat	turkey	robe	parsley
yacht	owl	cape	peas
submarine	pheasant	vest	celery
canoe	peacock	dress	broccoli
whale	stork	coat	cabbage
airplane	pigeon	jacket	cucumber
jet	woodpecker	skirt	rhubarb
helicopter	dove	camisole	zucchini
tank	raven	nightgown	asparagus

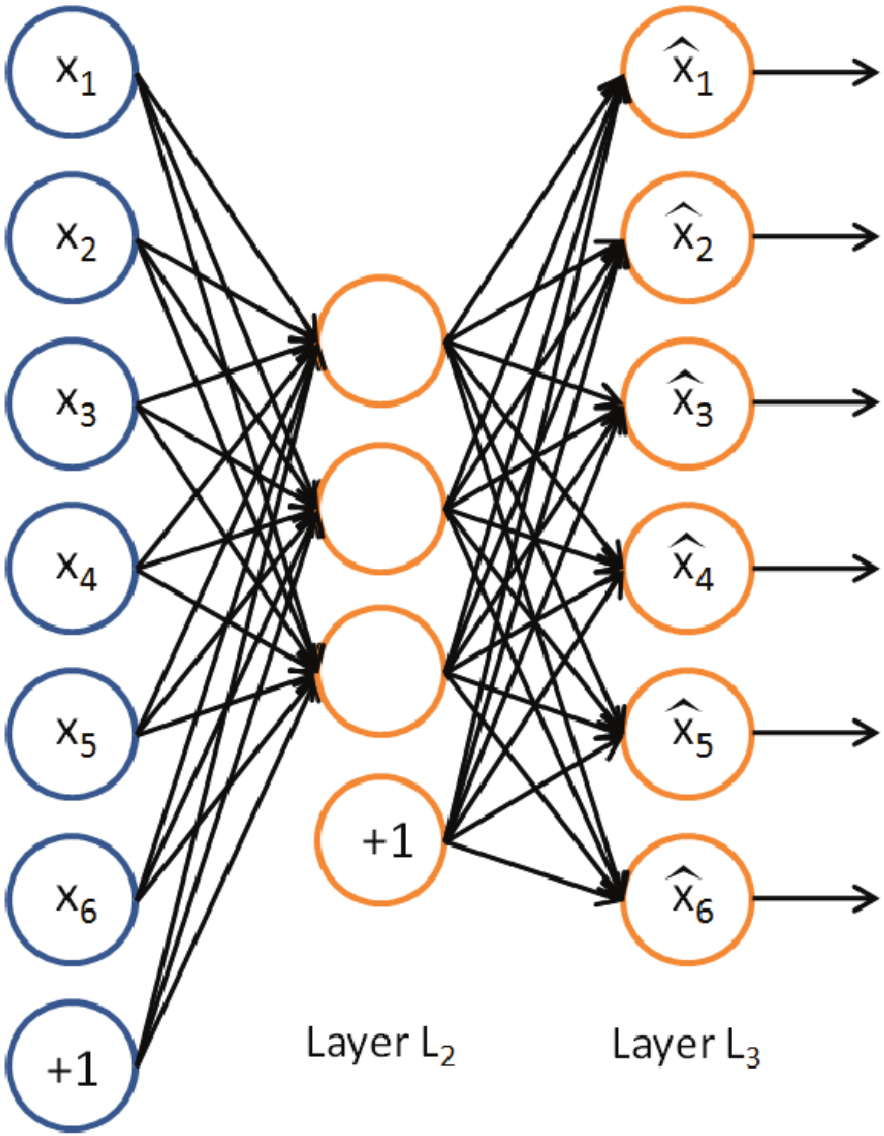
Part III

Autoencoders

Hinton and Salakhutdinov, 2006; Socher and Fei-Fei, 2010; Ngiam et al. (2011); Strivastava and Salakhudinov (2012)

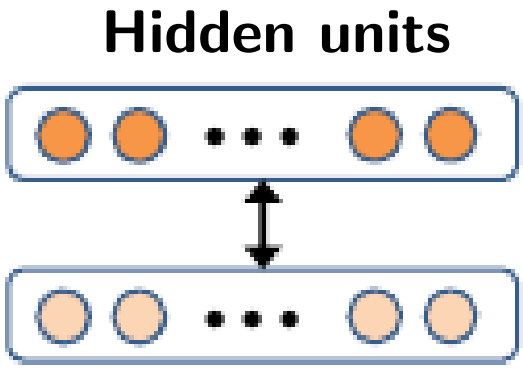
Model Definition

An autoencoder is a neural network that applies backpropagation setting the target values to be equal to the inputs.

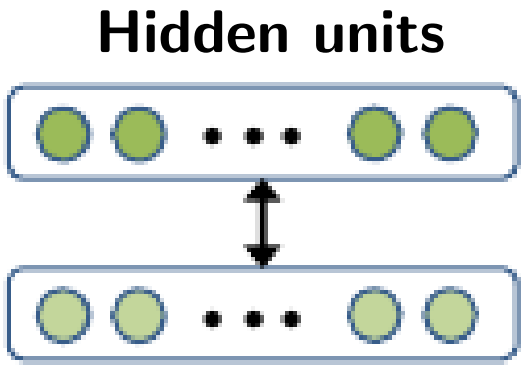


- Learns function $h_{W,b}(x) \approx x$
- Limit number of hidden units
- Learns compressed representation of input
- Can also impose sparsity constraints
- Can be stacked to form highly non-linear representations

Model Variants

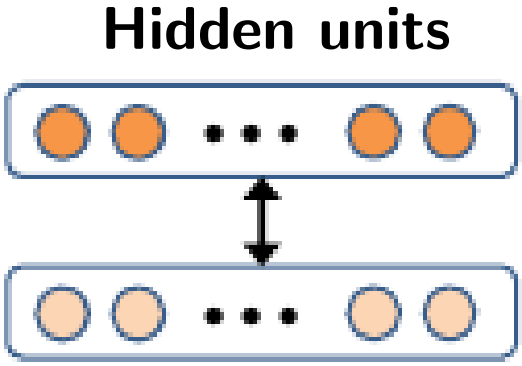


(a) Textual AE

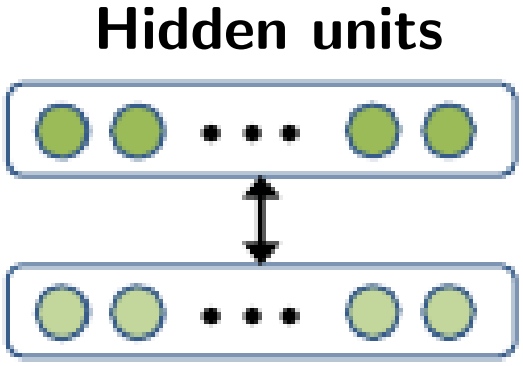


(b) Perceptual AE

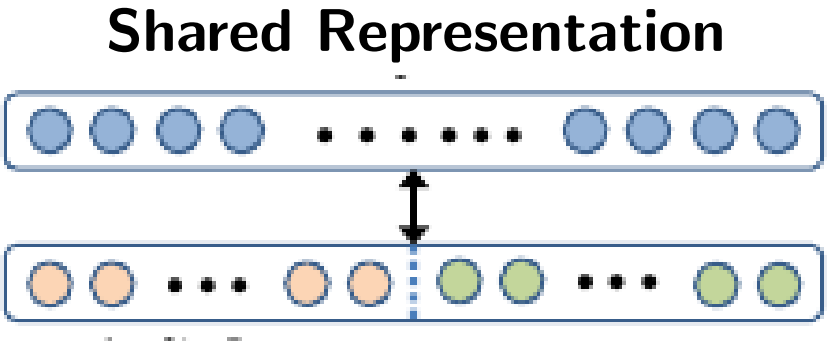
Model Variants



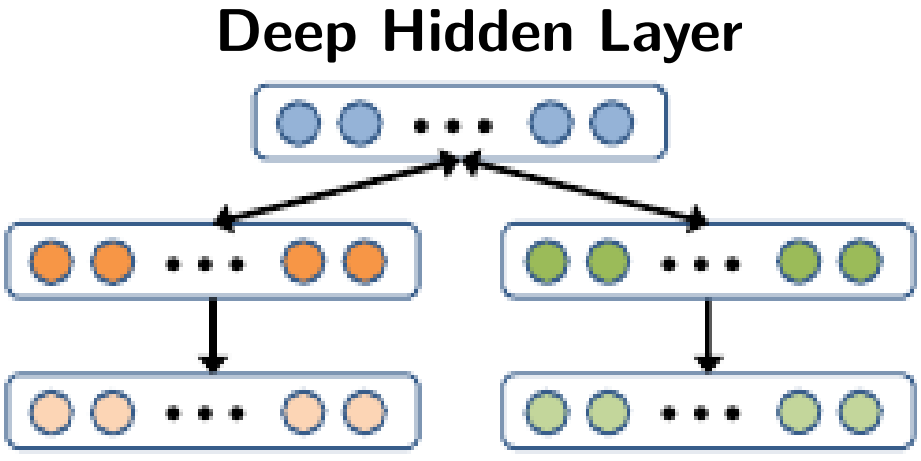
(a) Textual AE



(b) Perceptual AE



(c) Shallow Bimodal AE



(d) Stacked Bimodal AE

Experimental Setup

Textual Input: Strudel (Baroni et al., 2010) on English Wikipedia
eggplant–cook-v, *eggplant*–vegetable-n, *eggplant*–plant-n

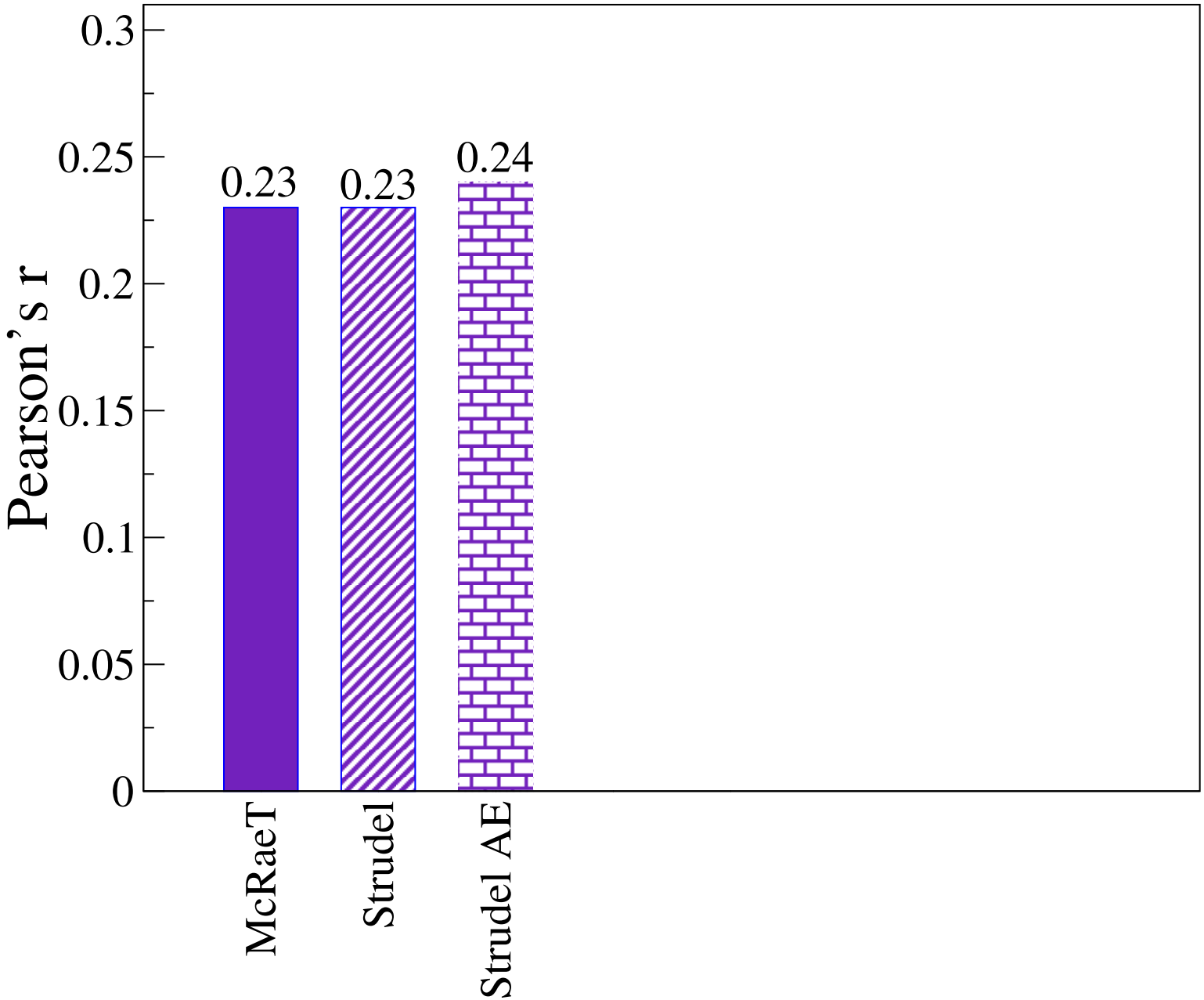
Perceptual Input: each image corresponds to 400-dimensional vector, each component is prediction score of single attribute classifier.

Word association task: given cue, write down words that come to mind (Nelson et al., 1998); 63,619 normed cue-associate pairs

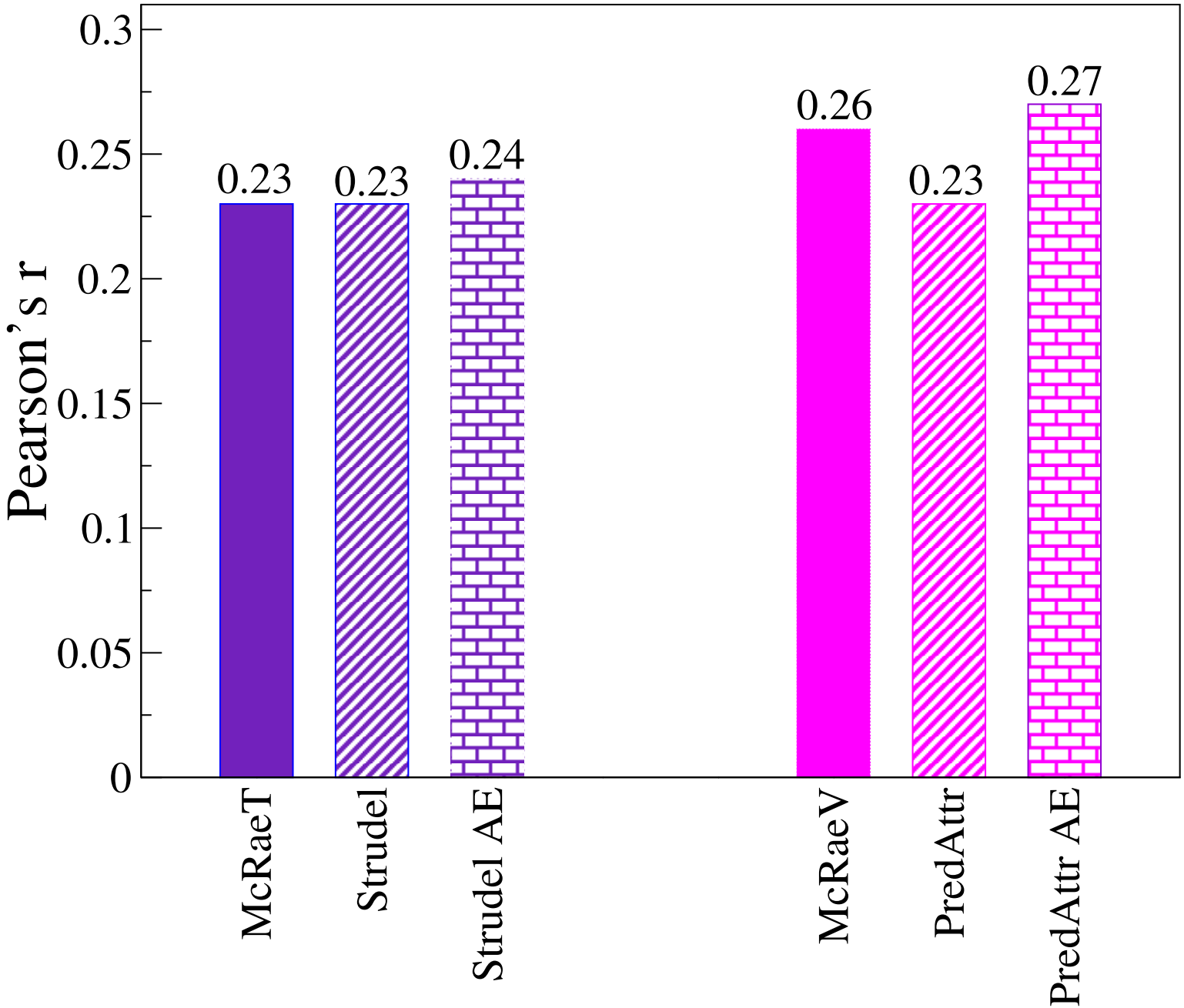
SPOTS	RICE	MOOSE
dog	chinese	animal
dirty	wedding	antler
dirt	food	deer
stripes	white	horn
dark	china	big

Evaluation: measure word-word cosine similarity; correlate with human association judgments (Pearson's r).

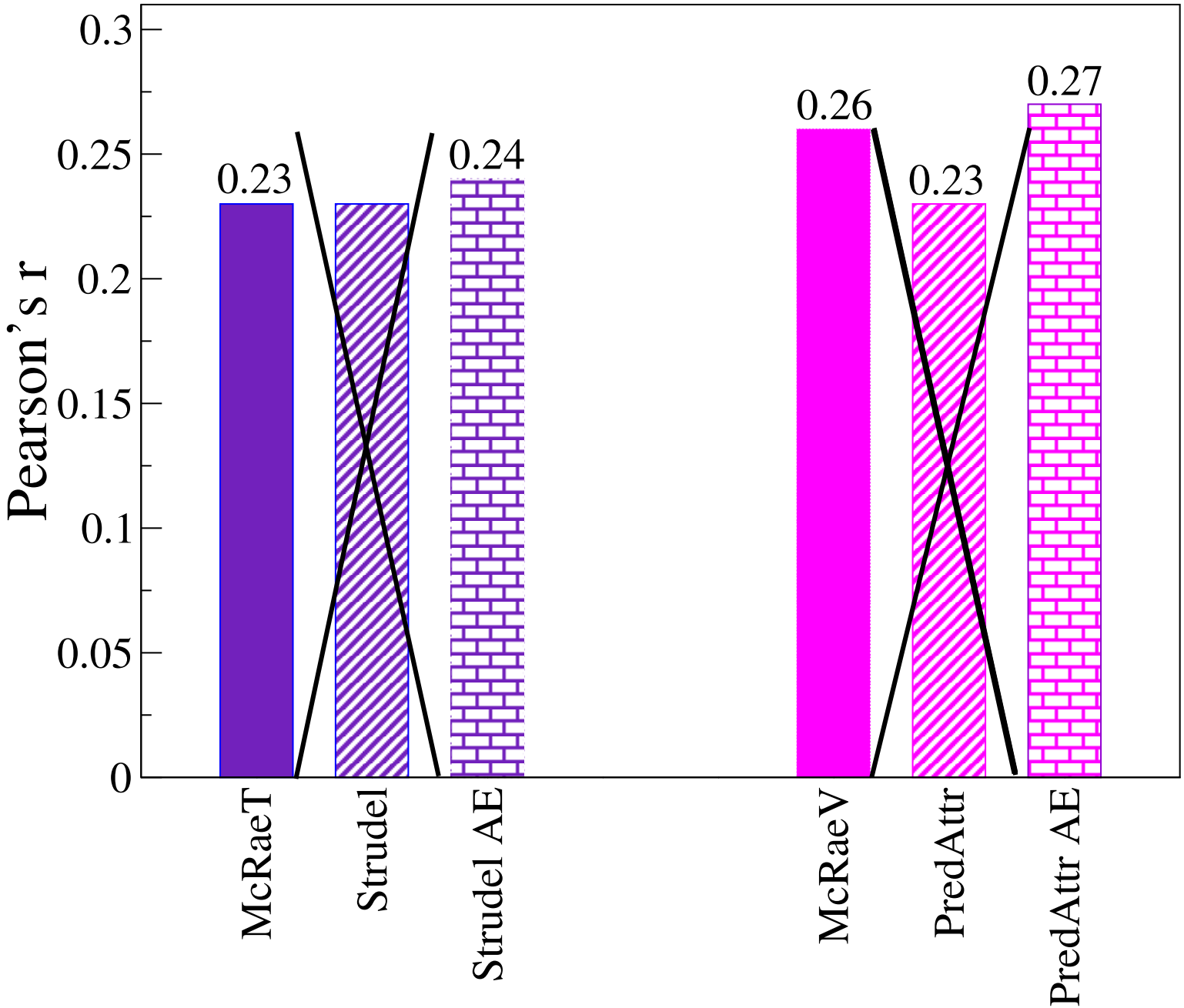
Results: Single Modalities



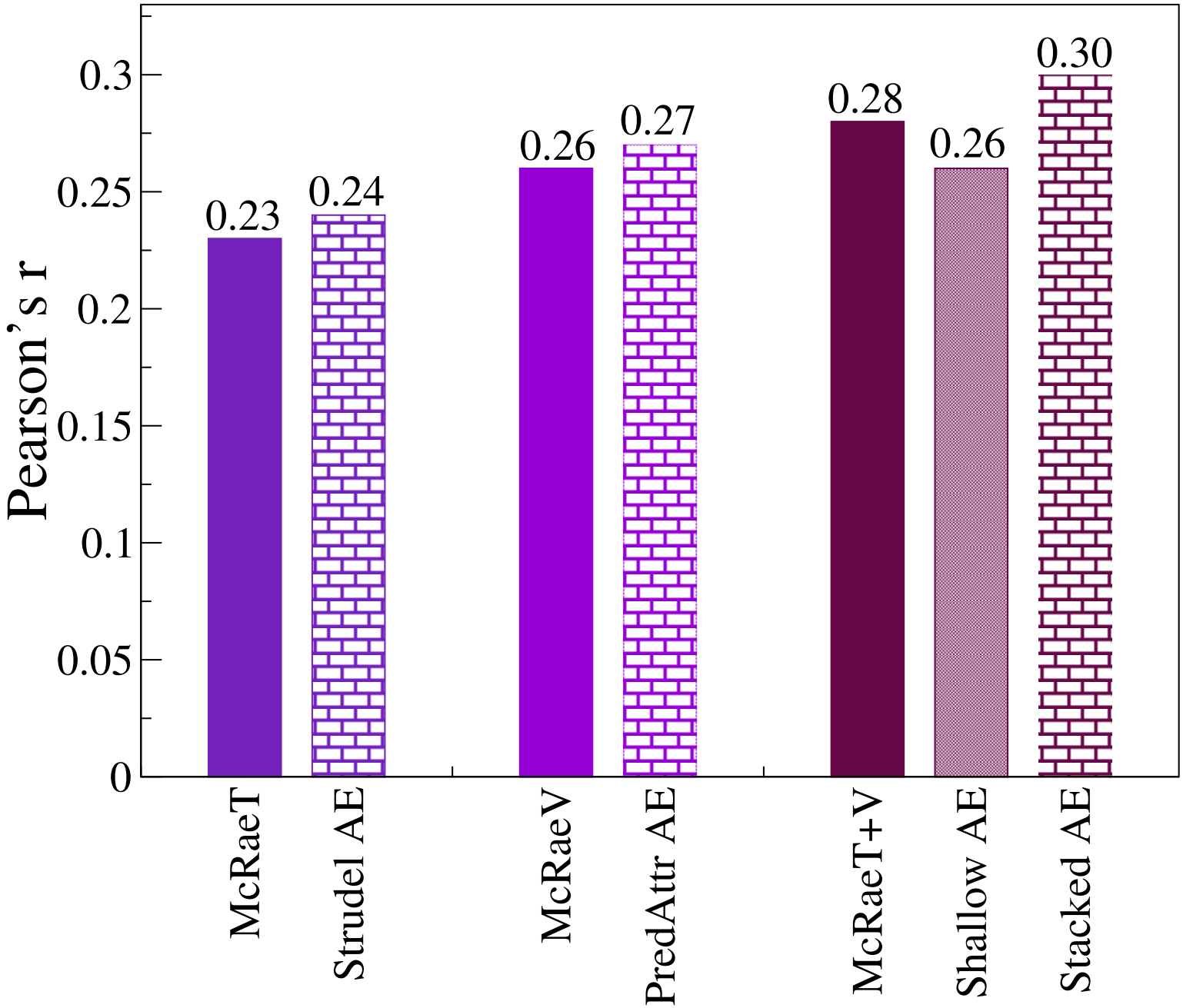
Results: Single Modalities



Results: Single Modalities



Results: Two Modalities



Discussion

Q_1: How should we represent perceptual information?
Proposal: visual attributes as means of grounding meaning.

Discussion

Q_1: How should we represent perceptual information?
Proposal: visual attributes as means of grounding meaning.

Q_2: How do we integrate perceptual and textual information?
Autoencoders show promise; attribute-based representation is not model specific.

Discussion

Q_1: How should we represent perceptual information?
Proposal: visual attributes as means of grounding meaning.

Q_2: How do we integrate perceptual and textual information?
Autoencoders show promise; attribute-based representation is not model specific.

Q_3: Are two modalities better than one?
Yes! Visual modality more prominent for association task.

Discussion

Q_1: How should we represent perceptual information?
Proposal: visual attributes as means of grounding meaning.

Q_2: How do we integrate perceptual and textual information?
Autoencoders show promise; attribute-based representation is not model specific.

Q_3: Are two modalities better than one?
Yes! Visual modality more prominent for association task.

Q_4: How do we deal with unseen information?
Attributes generalize to unseen images (500 nouns, concepts with non-perceptual correlates).

Discussion

Q_1: How should we represent perceptual information?
Proposal: visual attributes as means of grounding meaning.

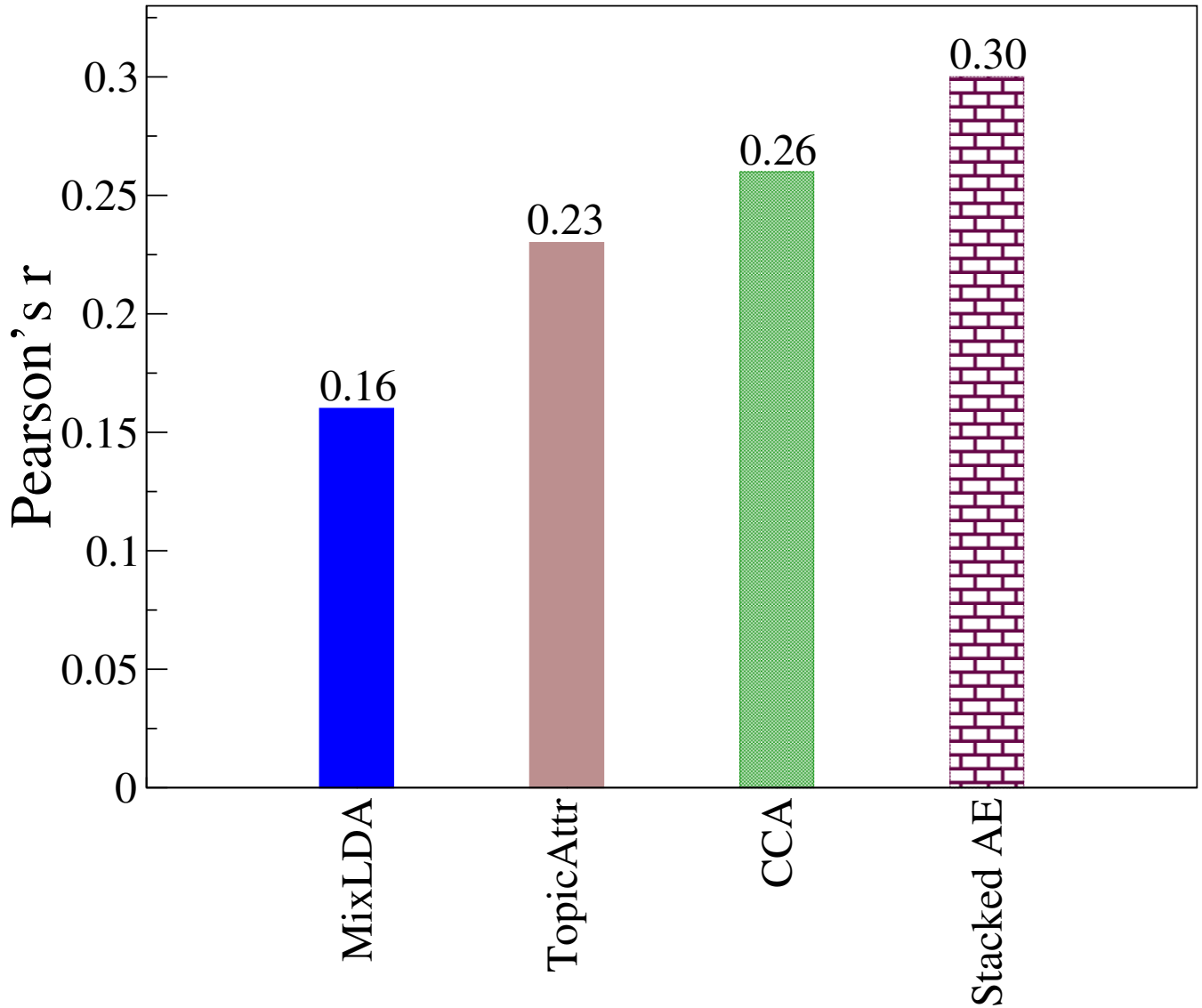
Q_2: How do we integrate perceptual and textual information?
Autoencoders show promise; attribute-based representation is not model specific.

Q_3: Are two modalities better than one?
Yes! Visual modality more prominent for association task.

Q_4: How do we deal with unseen information?
Attributes generalize to unseen images (500 nouns, concepts with non-perceptual correlates).

Q_5: How do we measure success?
Difficult! Need appropriate task such as categorization, object recognition, image retrieval.

Model Comparison



MixLDA (Feng and Lapata, 2010); TopicAttr (Andrews et al., 2009), CCA (Hardoon et al., 2004).