

Learning Generative Models of Sentences and Images

Richard Zemel



A woman with a Mohawk mask is in front of her .
A blonde woman with a colorful costume .
A female performer with a rainbow wig .

Microsoft Research
Faculty Summit
2015



CIFAR
CANADIAN
INSTITUTE
FOR
ADVANCED
RESEARCH

Learning Generative Models of Sentences and Images



A woman with a Mohawk mask is in front of her .
A blonde woman with a colorful costume .
A female performer with a rainbow wig .

Richard Zemel

July 8, 2015

Microsoft Faculty Summit



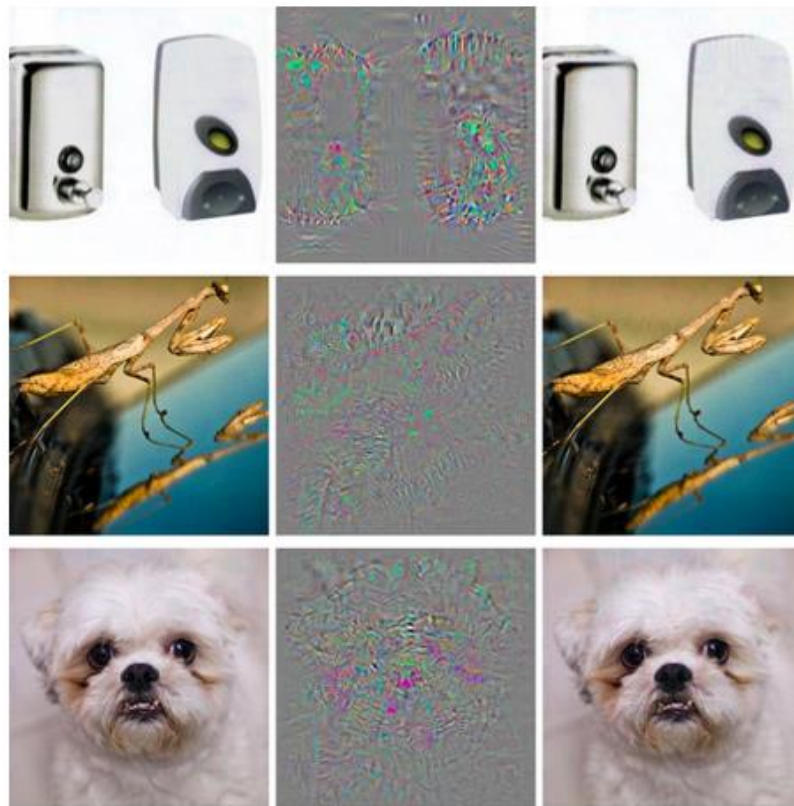
CIFAR
CANADIAN
INSTITUTE
FOR
ADVANCED
RESEARCH

Building Strong Models

Current successes of *deep networks*: classification problems (object recognition, speech recognition)

Standard supervised learning scenario with single correct response (class) for given input example

Current Models are Brittle



Szegedy, et al., ICLR, 2014

Building Strong Models

Current successes of deep networks: classification problems (object recognition, speech recognition)

Key aim: learn high quality generic representations, of images and text

Devise new objectives, based on image/text statistics, co-occurrence

Objective 1: Predict Context

When input consists of pairs (sets) of items, a sensible objective is to predict one item from the other

Standard setup:

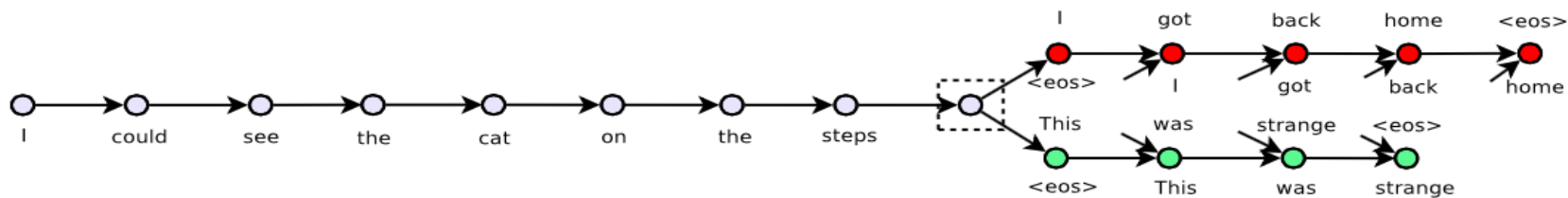
encoder maps first input to a vector

decoder maps vector to second input

Example: each word predicts the two words before and two words after it in a sentence (skip-gram [Mikolov et al., 2013])

Skip-Thought Vectors

Abstract the encoder-decoder model to whole sentences



Decode by predicting next word given generated words

Skip-Thought Vectors

Train on sentence triplets extracted from books

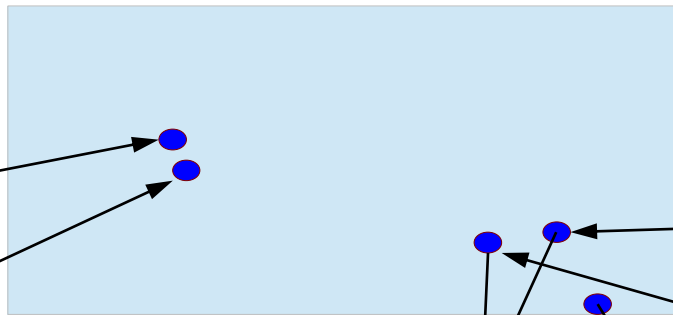
# of books	# of sentences	# of words	# of unique words	mean # of words per sentence
11,038	74,004,228	984,846,357	1,316,420	13

Demonstrate utility on 5 different NLP tasks (semantic relatedness, paraphrase detection)

Image Captioning as Context Prediction



A castle and reflecting water



Joint space



A ship sailing in the ocean

Minimize the following objective:

$$\begin{aligned} \text{images} \longrightarrow & \sum_{\mathbf{x}} \sum_k \max\{0, \alpha - s(\mathbf{x}, \mathbf{v}) + s(\mathbf{x}, \mathbf{v}_k)\} + \\ \text{text} \longrightarrow & \sum_{\mathbf{v}} \sum_k \max\{0, \alpha - s(\mathbf{v}, \mathbf{x}) + s(\mathbf{v}, \mathbf{x}_k)\} \end{aligned}$$

[Kiros et al, 2014]

Objective 2: Learning Generative Models

Another objective is to construct a model that can generate realistic inputs – ideally generalize beyond the training set

Difficult to formulate: cannot just directly match the training examples (over-fitting)

Learning Adversarial Models

One recently popular option: train model to fool adversary

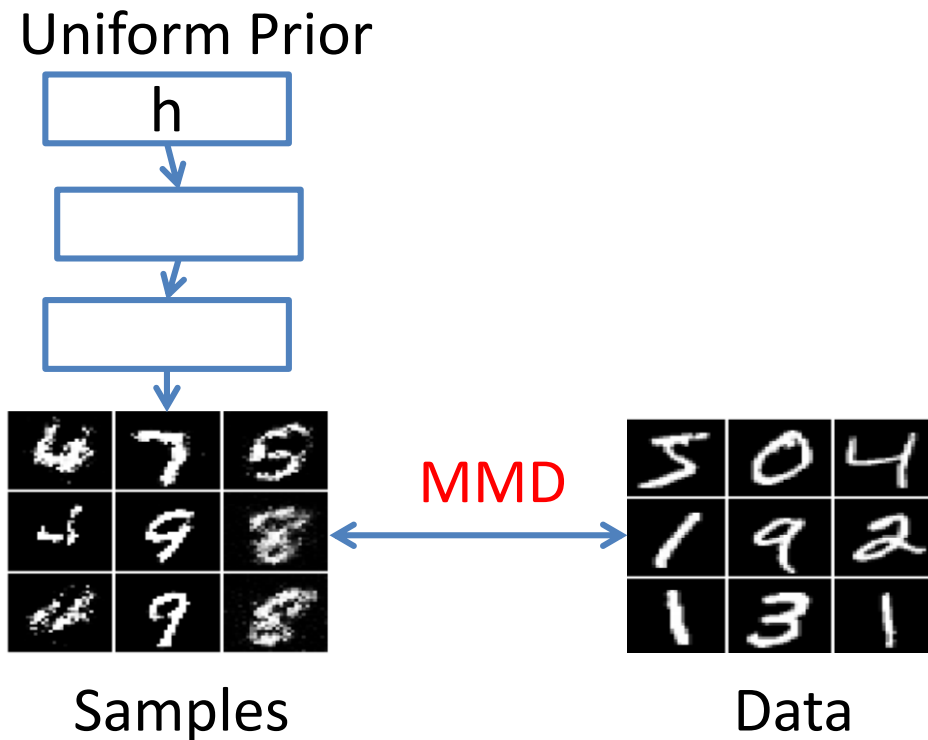
The adversary attempts to discriminate samples from the model from data samples

[MacKay 1995, 1996; Magdon-Ismael and Atiya, 1998;
Goodfellow et al. Generative Adversarial Nets. 2014]

Problem: min-max formulation makes optimization difficult

Generative Moment Matching Networks

Make model codes close to data codes



MMD

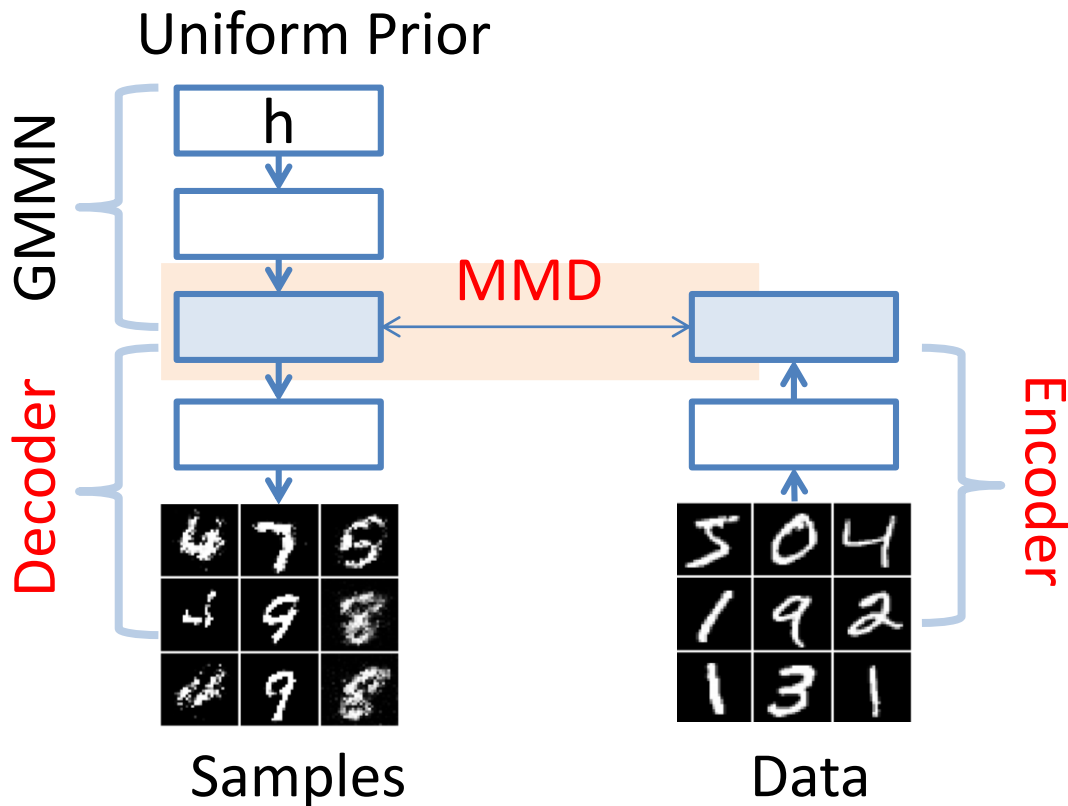
- Suppose we have access to samples from two probability distributions $X \sim P_A$ and $Y \sim P_B$, how can we tell if $P_A = P_B$?
- **Maximum Mean Discrepancy (MMD)** is a measure of distance between two distributions given only samples from each. [Gretton 2010]

$$\begin{aligned} & \left\| \frac{1}{N} \sum_{n=1}^N \phi(X_n) - \frac{1}{M} \sum_{m=1}^M \phi(Y_m) \right\|^2 \\ &= \frac{1}{N^2} \sum_{n=1}^N \sum_{n'=1}^N \phi(X_n)^\top \phi(X_{n'}) + \frac{1}{M^2} \sum_{m=1}^M \sum_{m'=1}^M \phi(Y_m)^\top \phi(Y_{m'}) - \frac{2}{NM} \sum_{n=1}^N \sum_{m=1}^M \phi(X_n)^\top \phi(Y_m) \\ &= \frac{1}{N^2} \sum_{n=1}^N \sum_{n'=1}^N k(X_n, X_{n'}) + \frac{1}{M^2} \sum_{m=1}^M \sum_{m'=1}^M k(Y_m, Y_{m'}) - \frac{2}{MN} \sum_{n=1}^N \sum_{m=1}^M k(X_n, Y_m) \end{aligned}$$

- Our idea: learn to make two distributions indistinguishable
→ small MMD!

Generative Moment Matching Networks

Direct backpropagation through MMD, no adversary required!



GMNN: Experiments

Model	MNIST	TFD
DBN	138 \pm 2	1909 \pm 66
Stacked CAE	121 \pm 1.6	2110 \pm 50
Deep GSN	214 \pm 1.1	1890 \pm 29
Adversarial nets	225 \pm 2	2057 \pm 26
GMMN	147 \pm 2	2085 \pm 25
GMMN+AE	282 \pm 2	2204 \pm 20

GMNN: Generalizing?



Independent Samples

Generated faces vs. Nearest Neighbors in Training Set



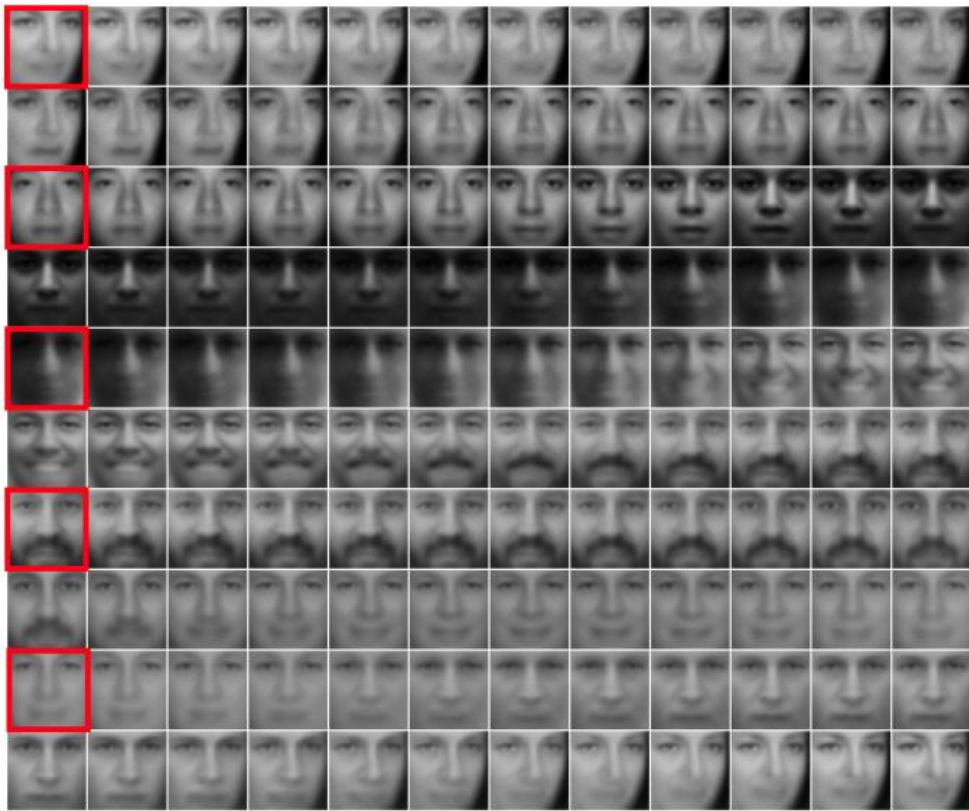
Exploring Latent Space

Interpolating between 5 random points (highlighted in red)



Exploring Latent Space

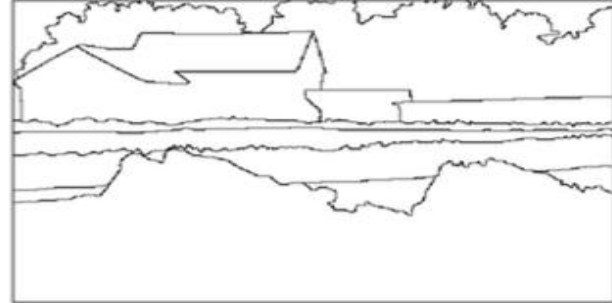
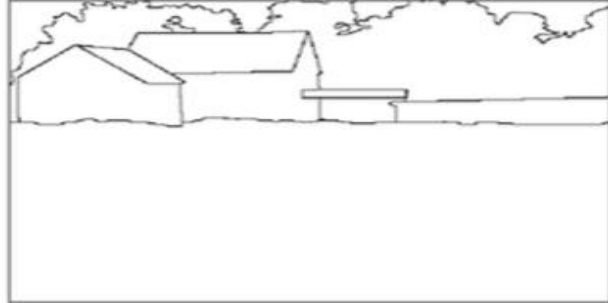
Interpolating between 5 random points (highlighted in red)



Objective 3: Learning One:Many Problems

Interesting tasks are often inherently ambiguous:

Segmenting image into coherent regions: What level of granularity?



Objective 3: Learning One:Many Problems

Interesting tasks are often inherently ambiguous:

Segmenting image into coherent regions: What level of granularity?

Generating caption for an image: What is relevant content?



generat
e →

A car on a beach with a boat in the background .

Two hilly islands in the water.

Can think of problem as one of diversity – what is the appropriate level of diversity in one → many mapping?

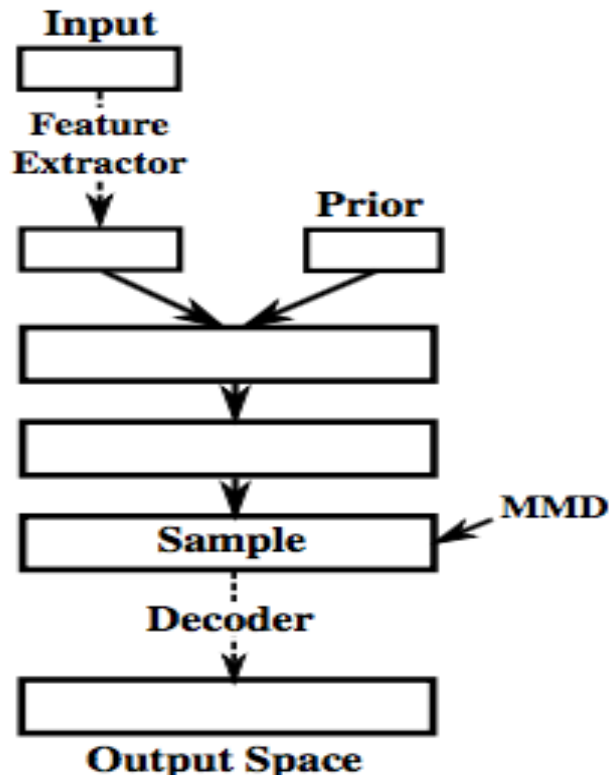
Luckily, data becoming available – can learn appropriate level of diversity for given input

Conditional Generative Moment Matching Networks

Include input as bias during generation

- Makes generation image-dependent
- Apply MMD on model/data samples per input

Idea: Generate outputs whose statistics match the statistics of multiple outputs for given input



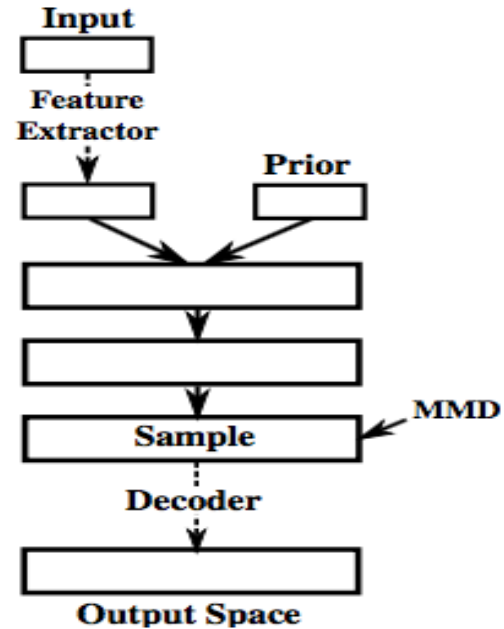
CGMMN: Image Captioning

Train joint embedding on Flickr8K dataset:

- 8000 images, 5 captions each
- 6000 training, 1000 each validate/test
- Images & sentences encoded in sentence space (skip-thought vectors)

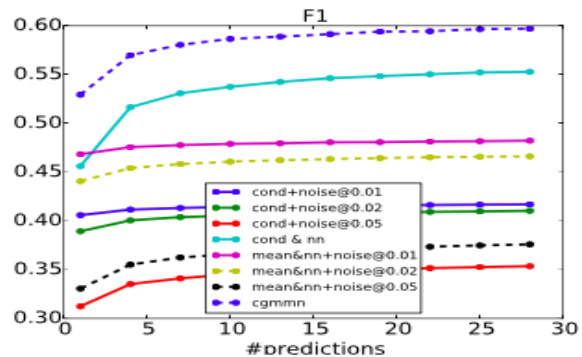
Projected down to 300 dimensional space

- CGMMN: 10-256-256-1024-300
- Minimize multiple kernel MMD loss

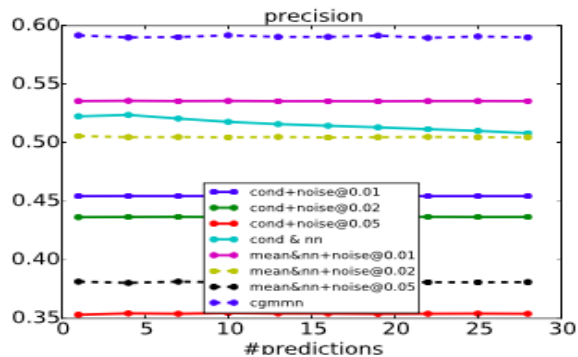


Aim: capture multi-modal distribution in code space

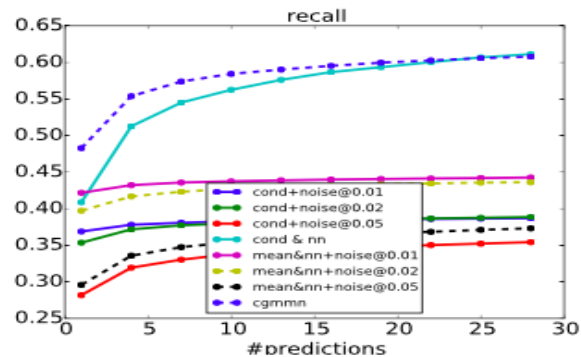
CGMMN: Image Captioning



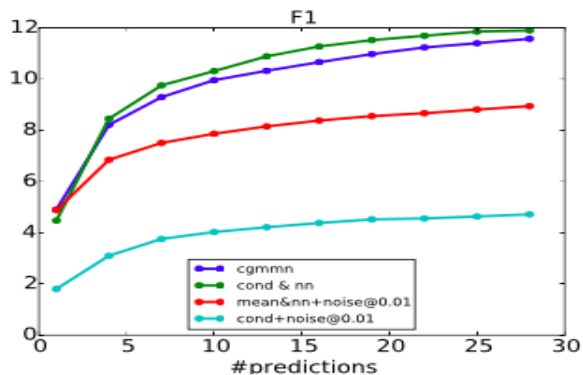
(a) F1



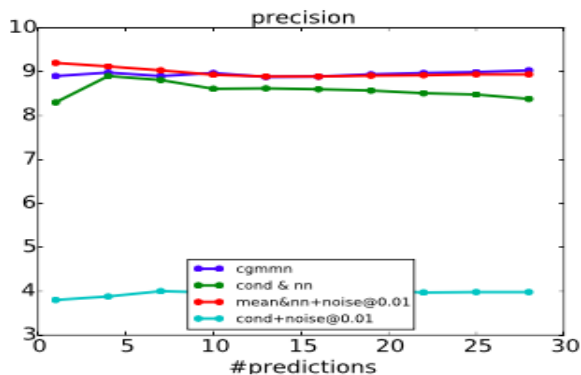
(b) precision



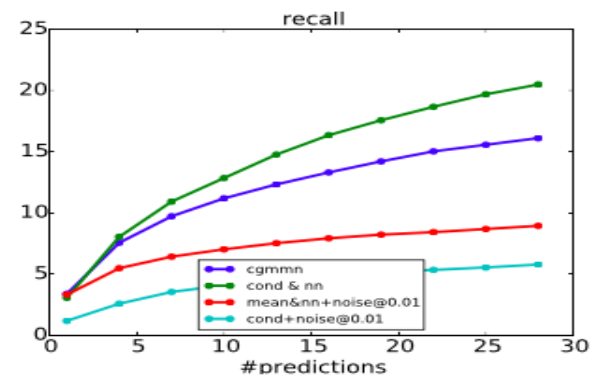
(c) recall



(a) F1




(b) precision




(c) recall

CGMMN: Image Captioning

Image	Method	Captions
	CGMMN	A small dog chases a ball in a green yard . A small dog plays with a yellow ball ball . A brown dog plays with a ball on the green grass .
	cond&nn	A black dog jumping to get a soccer ball . a small dog with a blue color fetching a yellow ball A dog catching a tennis ball at sunset in a yard .
	mean&nn+noise	A brown dog chasing a red ball . A brown dog chasing a ball . A brown dog chasing a ball .
	cond+noise	Brown dog with orange ball over chasing chasing each other . Brown dog chasing after jumping to catch orange balls . Brown dog chasing after jumping in orange hoop

CGMMN: Image Captioning

	CGMMN	A woman with a Mohawk mask is in front of her . A blonde woman with a colorful costume . A female performer with a rainbow wig .
	cond&nn	an asian woman holds her fur scarf . A person with face paint is staring at something from within a crowd . A young women with a pink mask over her face .
	mean&nn+noise	A woman wears a red and white necklace , smile . A woman wears a red , black and white headscarf , scarves . A woman dressed in a white dress and purple beads .
	cond+noise	a woman with red hair and making makeup getting off . Two women with red hair and a crowd . Two woman with red hair and a crowd of them .

CGMMN: Image Captioning



CGMMN	<p>A man with a basketball player is about to make the back .</p> <p>A basketball player takes a shot from the opposite team .</p> <p>A man playing a basketball , one has his arm around the number seven .</p>
cond&nn	<p>A basketball player wearing a white uniform is dribbling the ball on the court .</p> <p>A school basketball game is in progress .</p> <p>a basketball player wearing a red and white jersey while running down the court .</p>
mean&nn+noise	<p>An SMU basketball player in a white jersey dribbles the basketball .</p> <p>The basketball player dribbles the basketball in a Miami .</p> <p>The basketball player is dropping the basketball in a game .</p>
cond+noise	<p>basketball player in uniform .</p> <p>basketball player in uniform .</p> <p>basketball player with arms turned .</p>

Conclusions & Open Problems

Claim: Strong representation is not only capable of recognizing objects and words, but can model inputs themselves and their context

Developed 3 objectives for learning strong representations:

- Predict context (sentence-sentence; sentence-image)

- Generate distribution of images

- Generate distribution of sentences, specific to an image

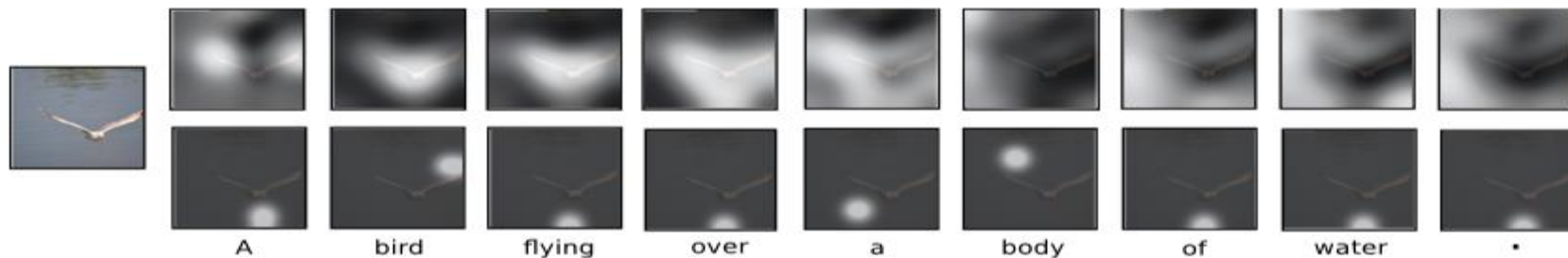
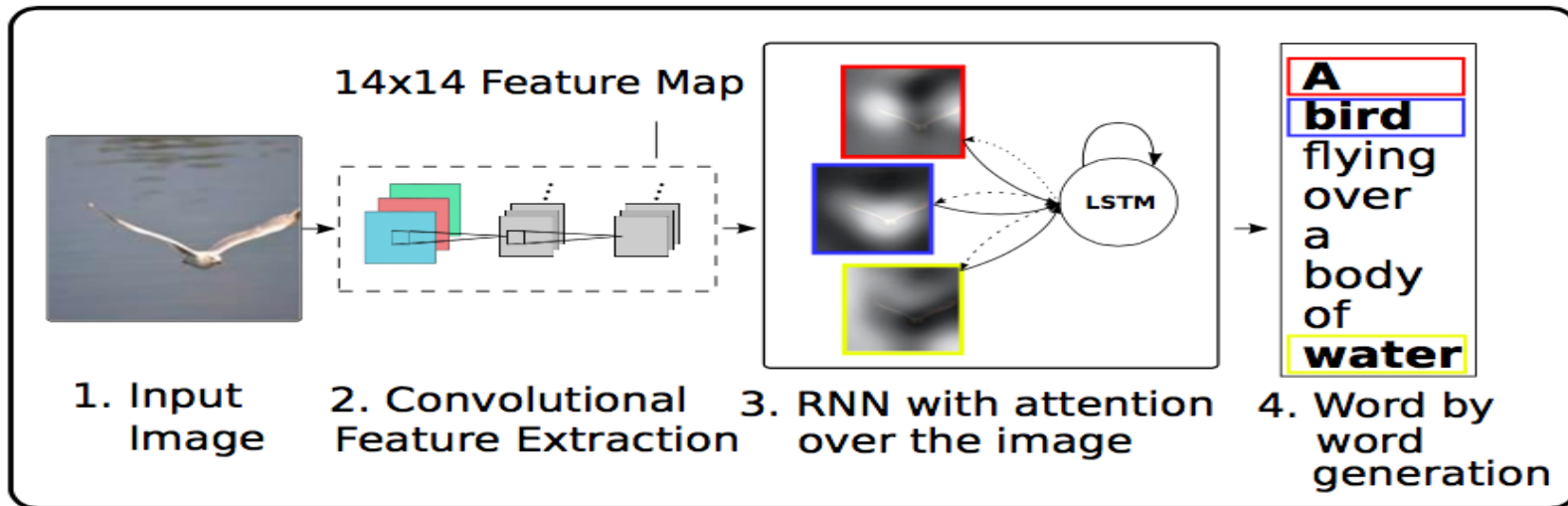
Leverage generative models?

- Gain insight into model behavior

- Improve standard classification tasks, esp. when labels scarce

- Generalize beyond static images to video

Generating Captions with Attention



Thanks!



A woman with a Mohawk mask is in front of her .
A blonde woman with a colorful costume .
A female performer with a rainbow wig .

Extra Slides

CGMMN: Image Segmentation

- Ambiguity in segmentation important
 - Different attentional foci, granularity
 - Valuable in interactive setting: user can choose from candidate segmentations to suit need
- Formulate problem: generate edge maps
 - CGMMN produces distribution over edge maps, sample to get different maps
 - Post-processing system constructs region hierarchy, threshold to form output regions
- Compare to strong baselines that produce single edge map
 - sample to get diverse map, sampling distribution optimized
 - apply same post-processing

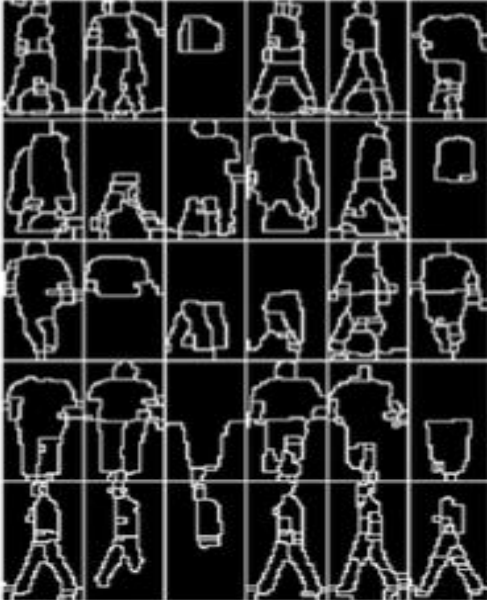
CGMMN: Image Segmentation



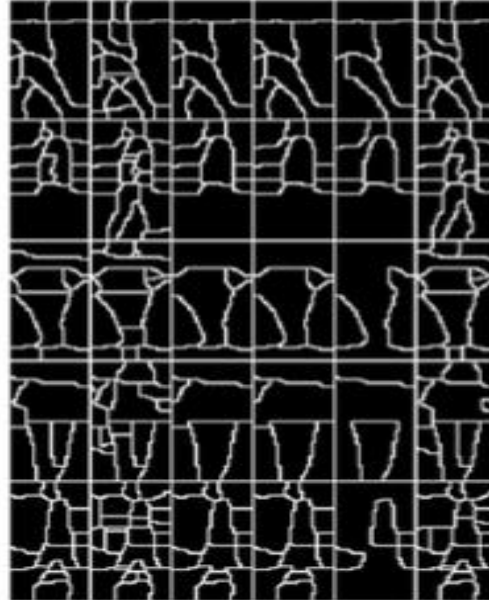
Image

Ground-Truth Segmentations

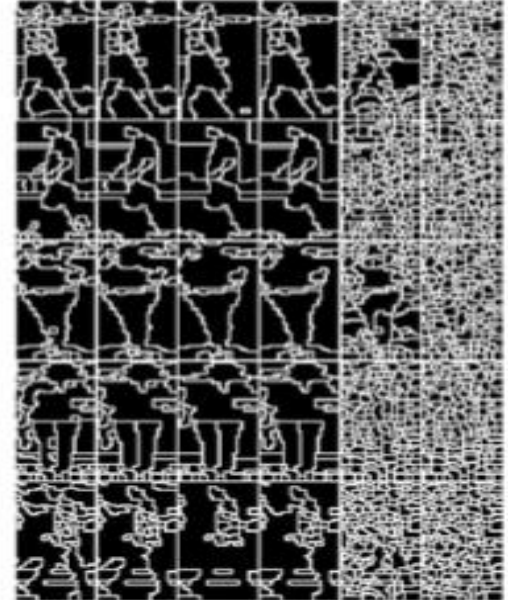
CGMMN: Image Segmentation



CGMMN

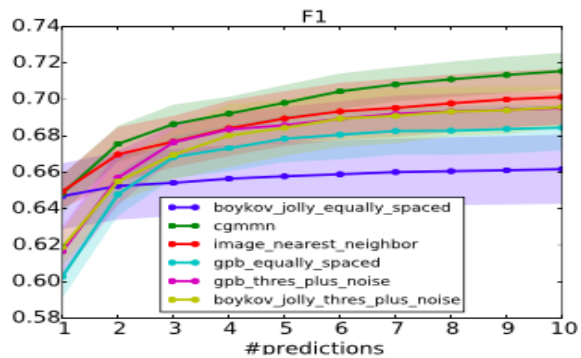


gPb

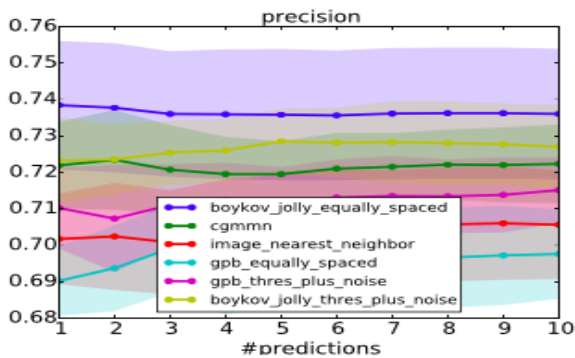


Boykov-Jolly

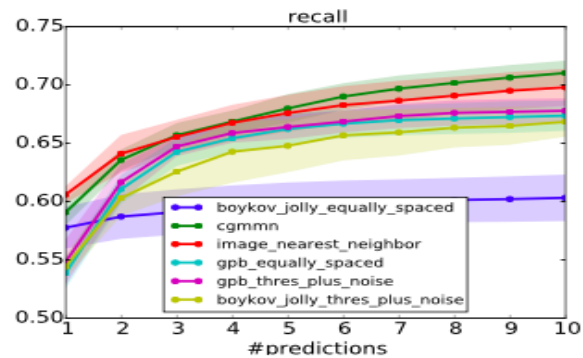
CGMMN: Image Segmentation



(a) F1



(b) precision



(c) recall

Image Q&A

Developed new Question/Answer dataset:

- Based on descriptions in COCO (COCO-QA)
- Use parse tree to make coherent questions
- Single word answers
- ~80K Q&A pairs (Object, Number, Color, Location)

- Developed a variety of models, baselines

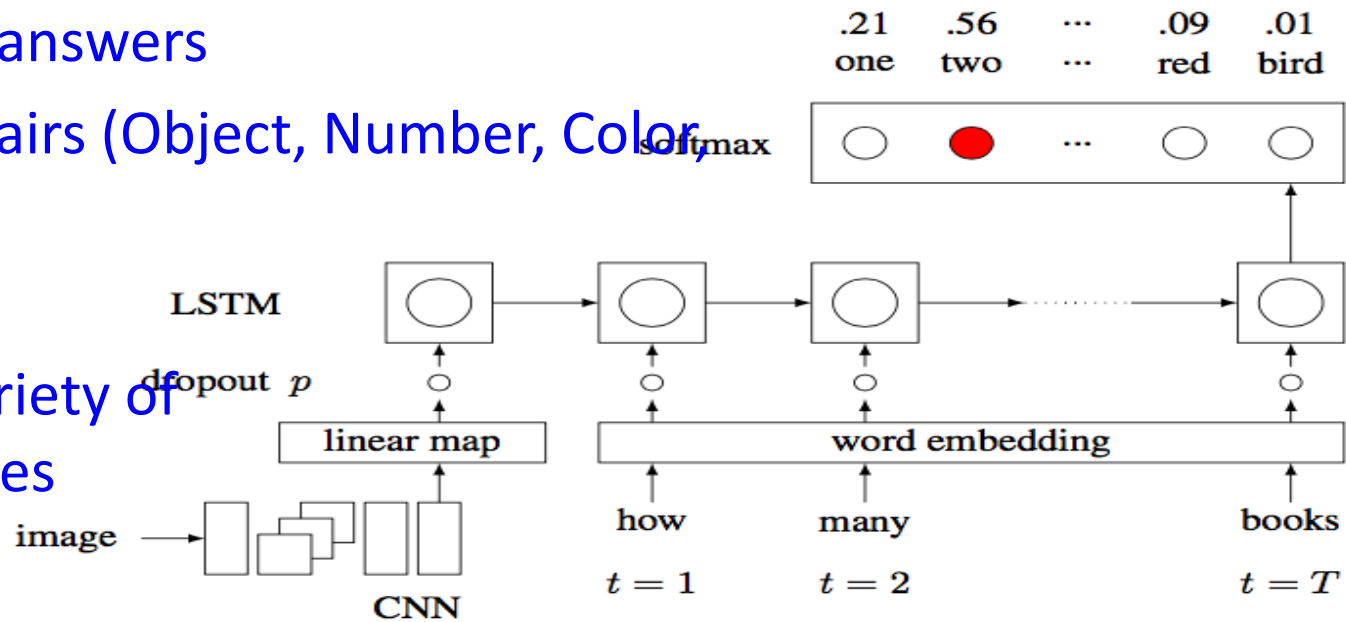


Image Q&A



COCOQA 4018

What is the color of the bowl?

Ground truth: blue

IMG+BOW: blue (0.49)

2-VIS+LSTM: blue (0.52)

BOW: white (0.45)

COCOQA 4018a

What is the color of the vest?

Ground truth: red

IMG+BOW: red (0.29)

2-VIS+LSTM: orange (0.37)

BOW: orange (0.57)



DAQUAR 1522

How many chairs are there?

Ground truth: two

IMG+BOW: four (0.24)

2-VIS+BLSTM: one (0.29)

LSTM: four (0.19)

DAQUAR 1520

How many shelves are there?

Ground truth: three

IMG+BOW: three (0.25)

2-VIS+BLSTM: two (0.48)

LSTM: two (0.21)

Image Q&A



COCOQA 14855

Where are the ripe bananas sitting?

Ground truth: basket

IMG+BOW: **basket (0.97)**

2-VIS+BLSTM: **basket (0.58)**

BOW: **bowl (0.48)**

COCOQA 14855a

What are in the basket?

Ground truth: bananas

IMG+BOW: **bananas (0.98)**

2-VIS+BLSTM: **bananas (0.68)**

BOW: **bananas (0.14)**



DAQUAR 585

What is the object on the chair?

Ground truth: pillow

IMG+BOW: **clothes (0.37)**

2-VIS+BLSTM: **pillow (0.65)**

LSTM: **clothes (0.40)**

DAQUAR 585a

Where is the pillow found?

Ground truth: chair

IMG+BOW: **bed (0.13)**

2-VIS+BLSTM: **chair (0.17)**

LSTM: **cabinet (0.79)**

Multiple Ground Truths: Caption Generation



(a)

Reference Sentences

- R1:** A bicyclist makes a gesture as he rides along
- R2:** A cyclist posing on his bicycle while riding it.
- R3:** A disabled biker rides on the road.
- R4:** A man in racing gear riding a bike and making a funny face.
- R5:** The man is riding his bike on the street.
- R6:** A man riding his bike outside.
- R7:** A man riding his bike.

(b)

Candidate Sentences

- C1:** A man rides a bike with one hand.
- C2:** A male biker dressed in white rides on pavement with a landscape of tree and grass behind him.

Triplet Annotation

Which of the sentences, B or C, is more similar to sentence A?

- Sentence A :** Anyone from R1 to R50
- Sentence B :** C1
- Sentence C :** C2

(c)

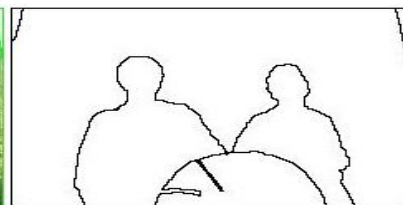
Multiple Ground Truths: Image Segmentation

Berkeley Segmentation Dataset: Test Image #229036 [color]

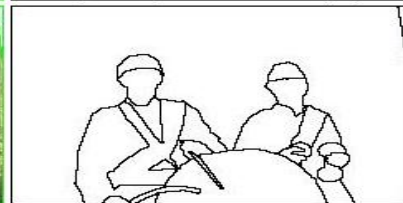
7 Color Segmentations



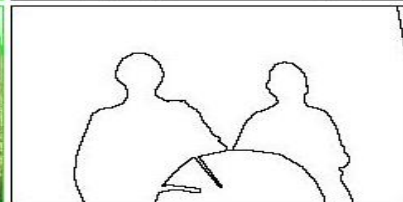
Multiple Ground Truths: Image Segmentation



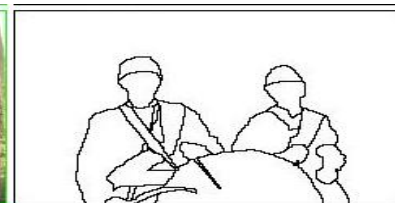
User #1104
7 Segments



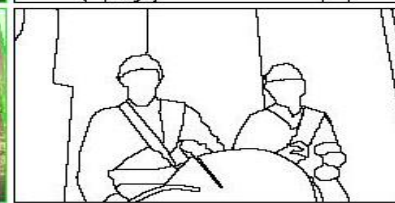
User #1105
26 Segments



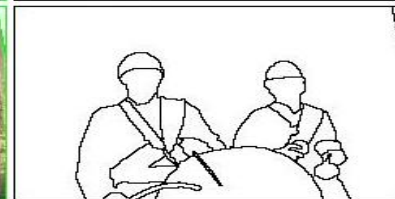
User #1109
5 Segments



User #1117
41 Segments



User #1123
39 Segments



User #1124
30 Segments