CLIPPING

ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM

JUMPING

ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

SPRAYING

ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	FIREMAN
SOURCE	SPRAY CAN	SOURCE	HOSE
SUBSTANCE	PAINT	SUBSTANCE	WATER
DESTINATION	WALL	DESTINATION	FIRE
PLACE	ALLEYWAY	PLACE	OUTSIDE

Situation Recognition: Visual Semantic Role Labeling for Image Understanding

By Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi

Presentation by Rishub Jain

Outline

- Problem statement
- Dataset
- Baseline model
- Experiments

Task Definition

CLIPPING			
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM

JUMPING			
ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

SPRAYING			
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	FIREMAN
SOURCE	SPRAY CAN	SOURCE	HOSE
SUBSTANCE	PAINT	SUBSTANCE	WATER
DESTINATION	WALL	DESTINATION	FIRE
PLACE	ALLEYWAY	PLACE	OUTSIDE

- Input: Image
- Output: (verb, realized frame) pair, where each realized frame is a list of pairs of (role, noun)
- For a given verb, its set of roles come directly from FrameNet
- The set of possible nouns are the 80,000 synsets in WordNet

Related Work

- Many other similar datasets (Stanford-40)
  - None are comprehensive in types of situations
- Work has been done on sentence generation
  - This approach can create simple sentences
  - Avoids evaluation challenges
  - Can better aid captioning
  - 20% of Visual Question Answering (VQA) tasks ask about a semantic role

The Dataset - imSitu

- 126,102 images
- 205,095 distinct situations
- 504 unique verbs
- 3.5 average roles per verb
- 1,788 unique roles

- 2 out of 3 annotators provided the same synset for over 75% of roles

verbs	504
images	126,102
realized frames / image	3
total annotations	1,481,851
unique entities (>= 3)	11,538 (6794)
semantic roles / verb (range)	3.55 (1 - 6)
semantic roles (types)	1788 (190)
images / verb (range)	250.2 (200 - 400)
unique realized frames (>= 3)	205,095 (21,505)
out of vocabulary rate (range)	3.5% (0% - 15.8%)
train / dev / test	75,702 / 25,200 / 25,200

Table 1. Summary statistics of imSitu.

Dataset Collection - Creating Verb and Role set

1. Extracted only visually related and recognizable verbs and roles from FrameNet
2. Created a sentence for each verb to define roles for annotators
    - "An AGENT clips an ITEM from a SOURCE using a TOOL in a PLACE."
3. Filtered out verbs for which 3 images could not be easily found through Google Image Search

Dataset Collection - Image Collection and Annotation

1. Mined phrases from Google Syntactic N-Grams that focused on verb-argument structure
2. Selected phrases that had dependencies on things like the object of the sentence
3. Through Google Image Search collected full-color medium-sized images that pass safe search
4. Workers filtered out images that were computer generated or didn’t match the activity searched
5. Given the image, the verb with its definition, and the roles with their sentence summary, workers assigned WordNet synsets to each role

Dataset Collection - Diversity and Coverage

1. Generated and annotated 200 images per verb
2. Calculated out of vocabulary (OOV) rate of each verb
    - Separated data into train and test sets
    - Found percentage of values for each role that appear in the test set but not training set
    - “putting” has a 15.5% rate while “flossing” has a 0.7% rate
3. Continue collecting more images if OOV rate > 5%, until a max of 400 images

Larger words have a larger rate of unseen value-role combinations

Dataset Statistics

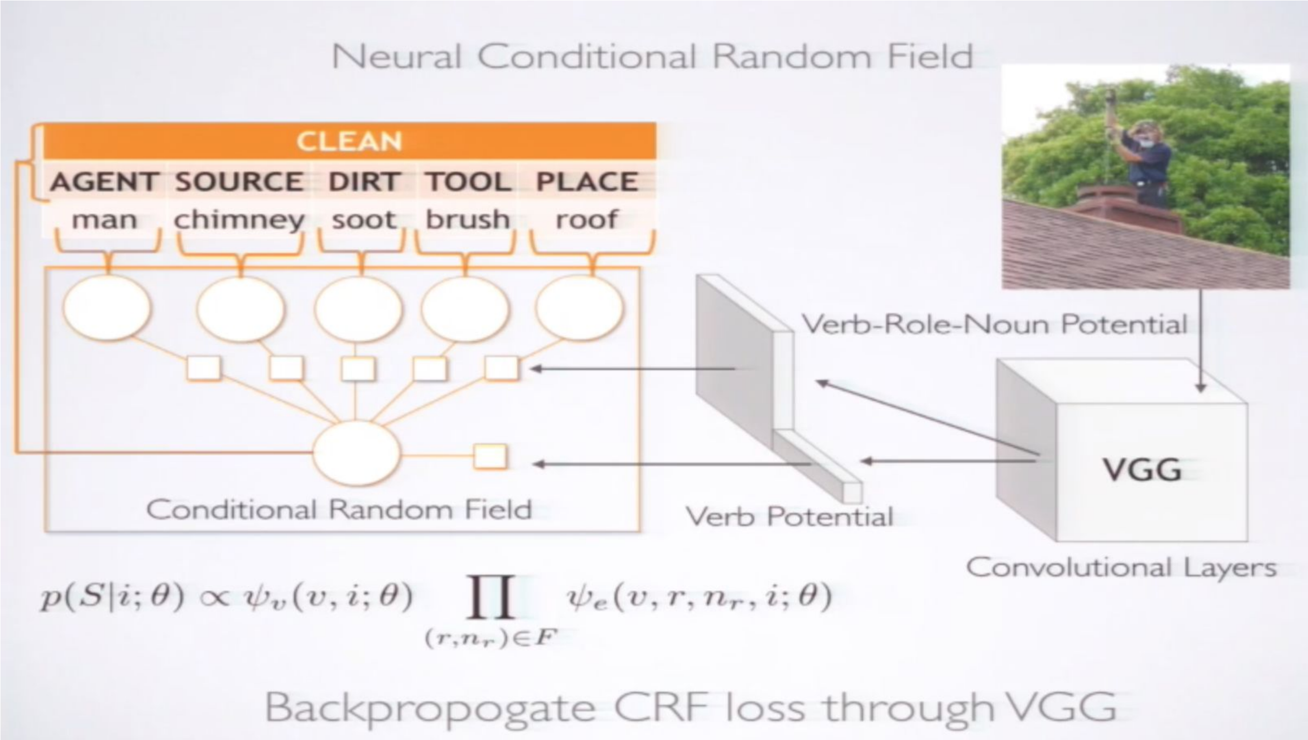
- 2 roles are in agreement if their sysnet values are within 3 links in the WordNet hierarchy
  - Ex: “musical instrument” and “trumpet” are 3 links away

- The “Place” role is ambiguous
- Number of roles a noun can take varies
  - “man” takes 798 roles, “basin” takes 1 role
- Number of nouns a role can take varies
  - “putting item” vs “surfing tool”
- Number of entities each verb can take varies
  - “putting” vs “flossing”

	Majority	1-link	2-link	3-link
all Roles	76.8	81.5	84.8	86.5
w/o Place	81.5	84.6	88.2	89.9

Percentage of role annotations that have 2 out of 3 annotators agree

Baseline Model



Baseline Model

$$p(S|i;\theta) \propto \psi_v(v,i;\theta) \prod_{(e,n_e)\in R_f} \psi_e(v,e,n_e,i;\theta)$$

- Situation S = (v=verb, R_f=realized frame) pair, where each realized frame is a list of pairs of (e=role, n_e=noun)
- E_f is the frame corresponding to the verb, and $e \in E_f$
- i is the image
- θ is the parameters for the CRF
- ψ_v is potential for verbs, and ψ_e is the potential for roles

Baseline Model

$$\psi_v(v, i; \theta) = e^{\phi_v(v,i)\theta}$$

$$\psi_e(v, e, n_e, i; \theta) = e^{\phi_e(v,e,n_e,i)\theta}$$

- ϕ_e and ϕ_v are the outputs of a VGG CNN pretrained on ImageNet
- A_i is the set of possible true situations of the image
- Optimize the log-likelihood of observing at least one situation $S \in A_i$

$$\sum_{i \in D} \log \left(1 - \prod_{S \in A_i} (1 - p(S|i; \theta)) \right)$$

Experiments - Situation Recognition

		top-1 predicted verb				top-5 predicted verbs				ground truth verbs		
		verb	value	value-any	value-full	verb	value	value-all	value-full	value	value-all	value-full
dev	Discrete Classifier	26.4	4.0	0.4	0.2	51.1	7.8	0.6	0.4	14.4	0.9	0.6
dev	CRF	32.2	24.6	14.3	11.2	58.6	42.7	22.7	17.5	65.9	29.5	22.3
test	Discrete Classifier	26.8	4.1	0.3	0.2	51.2	7.8	0.5	0.4	14.4	0.8	0.6
test	CRF	32.3	24.6	14.2	11.2	58.9	42.8	22.5	17.5	65.7	29.0	22.0

Table 3. Situation prediction results in imSitu. Structured prediction outperforms classification of ten most common situations per activity.

- Included a Discrete Classifier model for comparison
    - VGG-like CNN that selects one of the 10 most frequent realized frames for each verb (5040-class problem)
- “value” - percentage of perfectly predicted verb-role-noun triplets
- “value-any” - realized frame is “correct” if each pair in the frame matches an annotation, percentage of “correct” realized frames
- “value-full” - percentage of perfect predicted full structures triplets
- “ground truth verbs” - accuracy of roles given the correct verb

Figure 7. Example realized situations from imSitu. Below each image is a table where the first row is the activity, the left column is semantic roles, and the right column is values for those roles. On the left outlined in gold are examples of gold standard annotated data. On the right is random output from our CRF model when it correctly predicted the activity. Incorrect semantic role values are highlighted in red, whereas correct ones are green.

Generalize to Unseen Combinations

Train

FEEDING	
AGENT	MAN
EATER	BABY
FOOD	MILK
SOURCE	BOTTLE
PLACE	ROOM

Instances in train : 35

FEEDING	
AGENT	GIRL
EATER	HORSE
FOOD	CARROT
SOURCE	HAND
PLACE	PEN

Instances in train : 7

Test

FEEDING	
AGENT	WOMAN
EATER	HORSE
FOOD	MILK
SOURCE	BOTTLE
PLACE	BARN

Instances in train : 0

Experiments - Activity and Object Recognition

- Situations help give context for activity and object recognition
- Activity recognition - same setup but only predicting verb
- Object recognition - same setup but predicting a single synset value from the annotated frame

		activity		object	
		top-1	top-5	top-1	top-5
dev	Activity	30.6	57.4	-	-
	Object	-	-	64.9	94.1
	Situation	32.25	58.6	72.9	95.0
test	Activity	31.1	57.7	-	-
	Object	-	-	64.1	94.2
	Situation	32.3	58.9	72.7	94.8

Table 4. Object and activity recognition results in imSitu. Joint prediction of object and activity through situation recognition improves over independently predicting either object or activity.