Overview

Adversarial Domain Adaptation

Learning end-to-end driving models from crowdsourced dashcams

Vision and Language: Learning to reason to answer and explain
Adversarial Domain Adaptation

Eric Tzeng
UC Berkeley

Judy Hoffman
UC Berkeley/Stanford

Trevor Darrell
UC Berkeley
Adapting across domains?
Adapting across domains?

- Applying source classifier to target domain can yield inferior performance...
Adapting across domains?

- Fine tune?
  .....Zero or few labels in target domain
- Siamese network?
  .....No paired / aligned instance examples!
Adapting across domains: minimize discrepancy

$\theta_c$

object classifier

$X_i$
Adapting across domains: minimize discrepancy

$\theta_c$ object classifier

$x_i$
Adapting across domains: minimize discrepancy

$\theta_c$  
object classifier

$x_i$  

$\theta_D$  
domain classifier

[ICCV 2015]
Adapting across domains: minimize discrepancy

\[ q^s_i = [1, 0] \quad \text{and} \quad q^t_j = [0, 1] \]

\[ \theta_D \]

domain classifier

[ICCV 2015]
Adapting across domains: minimize discrepancy

[ICCV 2015]
Adapting across domains: minimize discrepancy
Adapting across domains: \textbf{minimize discrepancy}

$\theta_c$

object classifier

$X^*_i$
Deep domain confusion

Adversarial Training of domain label predictor and domain confusion loss:

\[
\min_{\theta_D} \mathcal{L}_D(x_S, x_T, \theta_{\text{repr}}; \theta_D) = \sum_{y_D = 1}^{D} \log q_{d}
\]

\[
\min_{\theta_{\text{repr}}} \mathcal{L}_{\text{conf}}(x_S, x_T, \theta_D; \theta_{\text{repr}}) = \sum_{d} \frac{1}{D} \log q_d.
\]

Domain Label Cross-entropy with uniform distribution

[Tzeng ICCV15]
Deep domain confusion

Train a network to minimize classification loss AND confuse two domains

\[ \mathcal{L}_D(x_S, x_T, \theta_{\text{repr}}; \theta_D) = - \sum_d \mathbb{1}[y_D = d] \log q_d \]

\[ q = \text{softmax}(\theta_D^T f(x; \theta_{\text{repr}})) = p(y_D = 1|x) \]

\[ \mathcal{L}_{\text{conf}}(x_S, x_T, \theta_D; \theta_{\text{repr}}) = - \sum_d \frac{1}{D} \log q_d \]

(cross-entropy with uniform distribution)

[Source: Tzeng ICCV15]
Adversarial Discriminative Domain Adaptation (ADDA) (in submission)

Generative or discriminative? √

Which loss?

Shared or not? √

<table>
<thead>
<tr>
<th>Method</th>
<th>Base model</th>
<th>Weight sharing</th>
<th>Adversarial loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient reversal [16]</td>
<td>discriminative</td>
<td>shared</td>
<td>minimax</td>
</tr>
<tr>
<td>Domain confusion [12]</td>
<td>discriminative</td>
<td>shared</td>
<td>confusion</td>
</tr>
<tr>
<td>CoGAN [13]</td>
<td>generative</td>
<td>unshared</td>
<td>GAN</td>
</tr>
<tr>
<td>ADDA (Ours)</td>
<td>discriminative</td>
<td>unshared</td>
<td>GAN</td>
</tr>
</tbody>
</table>
Adversarial Discriminative Domain Adaptation (ADDA) (in submission)

Pre-training

Source images + labels → Source CNN → Classifier → class label

Adversarial Adaptation

Source images → Source CNN → Discriminator → domain label

Target images → Target CNN → Classifier → class label

Testing

Target image → Target CNN → Classifier → class label

min \(_{M_s,C}\) \(\mathcal{L}_{\text{cls}}(X_s, Y_s) = -\mathbb{E}_{(x_s, y_s) \sim (X_s, Y_s)} \sum_{k=1}^{K} \mathbb{1}_{[k=y_s]} \log C(M_s(x_s))\)

min \(_{D}\) \(\mathcal{L}_{\text{adv}_D}(X_s, X_t, M_s, M_t) = -\mathbb{E}_{x_s \sim X_s} \log D(M_s(x_s)) - \mathbb{E}_{x_t \sim X_t} \log (1 - D(M_t(x_t)))\)

min \(_{M_s, M_t}\) \(\mathcal{L}_{\text{adv}_M}(X_s, X_t, D) = -\mathbb{E}_{x_t \sim X_t} \log D(M_t(x_t))\).
## ADDA: Adaptation on digits

(in submission)

![Image of digits ]

<table>
<thead>
<tr>
<th>Method</th>
<th>MNIST $\rightarrow$ USPS</th>
<th>USPS $\rightarrow$ MNIST</th>
<th>SVHN $\rightarrow$ MNIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source only</td>
<td>$0.752 \pm 0.016$</td>
<td>$0.571 \pm 0.017$</td>
<td>$0.601 \pm 0.011$</td>
</tr>
<tr>
<td>Gradient reversal</td>
<td>$0.771 \pm 0.018$</td>
<td>$0.730 \pm 0.020$</td>
<td>$0.739$ [16]</td>
</tr>
<tr>
<td>Domain confusion</td>
<td>$0.791 \pm 0.005$</td>
<td>$0.665 \pm 0.033$</td>
<td>$0.681 \pm 0.003$</td>
</tr>
<tr>
<td>CoGAN</td>
<td>$0.912 \pm 0.008$</td>
<td>$0.891 \pm 0.008$</td>
<td>did not converge</td>
</tr>
<tr>
<td>ADDA (Ours)</td>
<td>$0.894 \pm 0.002$</td>
<td>$0.901 \pm 0.008$</td>
<td>$0.760 \pm 0.018$</td>
</tr>
</tbody>
</table>
ADDA: Adaptation on RGB-D

Train on RGB

Test on depth

<table>
<thead>
<tr>
<th># of instances</th>
<th>bathtub</th>
<th>bed</th>
<th>bookshelf</th>
<th>box</th>
<th>chair</th>
<th>counter</th>
<th>desk</th>
<th>door</th>
<th>dresser</th>
<th>garbage bin</th>
<th>lamp</th>
<th>monitor</th>
<th>night stand</th>
<th>pillow</th>
<th>sink</th>
<th>sofa</th>
<th>table</th>
<th>television</th>
<th>toilet</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source only</td>
<td>0.000</td>
<td>0.010</td>
<td>0.011</td>
<td>0.124</td>
<td>0.188</td>
<td>0.029</td>
<td>0.041</td>
<td>0.047</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.039</td>
<td>0.587</td>
<td>0.000</td>
<td>0.008</td>
<td>0.010</td>
<td>0.000</td>
<td>0.000</td>
<td>0.0139</td>
<td></td>
</tr>
<tr>
<td>ADDA (Ours)</td>
<td>0.000</td>
<td>0.146</td>
<td>0.046</td>
<td>0.229</td>
<td>0.344</td>
<td>0.447</td>
<td>0.025</td>
<td>0.023</td>
<td>0.000</td>
<td>0.018</td>
<td>0.292</td>
<td>0.081</td>
<td>0.029</td>
<td>0.021</td>
<td>0.116</td>
<td>0.143</td>
<td>0.091</td>
<td>0.000</td>
<td>0.211</td>
<td></td>
</tr>
<tr>
<td>Train on target</td>
<td>0.105</td>
<td>0.531</td>
<td>0.494</td>
<td>0.295</td>
<td>0.619</td>
<td>0.573</td>
<td>0.057</td>
<td>0.120</td>
<td>0.291</td>
<td>0.576</td>
<td>0.189</td>
<td>0.235</td>
<td>0.630</td>
<td>0.362</td>
<td>0.248</td>
<td>0.357</td>
<td>0.303</td>
<td>0.647</td>
<td>0.468</td>
<td></td>
</tr>
</tbody>
</table>
Autonomous Driving Paradigms

1) Learn affordances to predict state; apply rules or learned classic controllers

2) Abandon engineering principles, learn “end-to-end” policy
Autonomous Driving Paradigms

1) Learn affordances to predict state; apply rules or learned classic controllers

   How can visual sensing be robust to new environments?

2) Abandon engineering principles, learn “end-to-end” policy

   How to learn generic driving policies from diverse data?
Autonomous Driving Paradigms

1) Learn affordances to predict state; apply rules or learned classic controllers

   How can visual sensing be robust to new environments?
   ...Fully Convolutional Domain Adaptation “in the wild”

2) Abandon engineering principles, learn “end-to-end” policy

   How to learn generic driving policies from diverse data?
   ...Learning end-to-end driving policy/model from crowdsourced videos
BDD Dataset

BDD Video
- 720p 30fps 40s video clips
- ~50K clips
- GPS + IMU

BDD Segmentation
- 720p images
- Fine instance segmentation
- Compatible with Cityscapes
In-domain fully supervised FCN

Train on Cityscapes, Test on Cityscapes
Domain shift: Cityscapes to SF

Train on Cityscapes, Test on San Francisco Dashcam
No tunnels in CityScapes?...
Medium Shift: Cross Seasons Adaptation
Small Shift: Cross City Adaptation
Effect of domain confusion loss

Before domain confusion

After domain confusion
Effect of domain confusion loss

Before domain confusion

After domain confusion
Effect of domain confusion loss

Before domain confusion

After domain confusion
Effect of domain confusion loss

Before domain confusion

After domain confusion
FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation

Judy Hoffman, Dequan Wang, Fisher Yu, Trevor Darrell

(Submitted on 8 Dec 2016)

Fully convolutional models for dense prediction have proven successful for a wide range of visual tasks. Such models perform well in a supervised setting, but performance can be surprisingly poor under domain shifts that appear mild to a human observer. For example, training on one city and testing on another in a different geographic region and/or weather condition may result in significantly degraded performance due to pixel-level distribution shift. In this paper, we introduce the first domain adaptive semantic segmentation method, proposing an unsupervised adversarial approach to pixel prediction problems. Our method consists of both global and category specific adaptation techniques. Global domain alignment is performed using a novel semantic segmentation network with fully convolutional domain adversarial learning. This initially adapted space then enables category specific adaptation through a generalization of constrained weak learning, with explicit transfer of the spatial layout from the source to the target domains. Our approach outperforms baselines across different settings on multiple large-scale datasets, including adapting across various real city environments, different synthetic sub-domains, from simulated to real environments, and on a novel large-scale dash-cam dataset.
Overview

Adversarial Domain Adaptation

Learning end-to-end driving models from crowdsourced dashcams

Vision and Language: Learning to reason to answer and explain
Learning and Adapting from Large-Scale Driving Data

• Fully Convolutional Domain Adaptation “in the wild”

• Learning end-to-end driving policy/model from dashcam videos
End-to-End Paradigm

• ALVINN
• DAVE
• NVIDIA
• BDD RC Cars
• BDD WebCam
AVLINN
(1989)

ALVINN:
An Autonomous Land Vehicle In a
Neural Network

Dean A. Pomerleau
January 1989
CMU-CS-89-107-
Yann LeCun, Eric Cosatto, Jan Ben, Urs Muller, Beat Flepp: *End-to-End Learning of Vision-Based Obstacle Avoidance for Off-Road Robots*. Delivered at the Learning@Snowbird Workshop, April 2004.
model driving car, ‘direct’ mode
10 future timepoints

conv1
96 channels

metadata input 4 channels

conv2 384 channels

fully connected 512 channels

steering 10 future timepoints

motor 10 future timepoints

camera input 4 channels

fully connected 512 channels

conv2 384 channels

metadata input 4 channels

conv1 96 channels

camera input 4 channels
Driving Policy
Learning a universal driving policy

• Self driving as egomotion prediction

• Learn general driving policy that is applicable to all car models.

• Use a large number of easily accessible dashcam videos as self-supervision.
FCN-LSTM

Visual Encoder
- Dilated Fully Convolutional Nets could provide more spatial details than CNN

Temporal Fusion
- Fuse the visual information, vehicle state (speed and angular velocity) from each frame
FCN-LSTM

Privileged Learning

• The model should implicitly know what objects are in the scene
• We use the semantic segmentation mask from Cityscapes as extra source of supervision
• It ultimately improves the learnt representation of the dilated FCN

Dataset

- Real first person driving videos
- Diverse
  - City
  - Highway
  - Rainy days
  - Nights and evenings
  - Construction zones

Sample frames from the dataset
Scene and Trajectory Reconstruction of Crowd-sourced Driving Videos using Semantic Filtered SfM

Yang Gao*, Huazhe Xu*, Christian Hane, Fisher Yu, Trevor Darrell
Challenging Driving Videos in the Wild

Challenges

Moving Objects
Subtle behaviors
  Lane changing
  Slight Steering
Unknown Camera Calibration
Rolling Shutters
Existing Motion-Based Method Failed to Reject Moving Object from the Scene

Keypoints from motion-based keypoints rejection methods

Keypoints from our Semantic Filtered SfM pipeline. Most moving keypoints have been filtered out.
Semantically Filtered SfM: $(Sf)^2 M$

Classical keypoints matching as points pair preference ranking

$$M(i_1, i_2) = \frac{1}{\|d(l_1, i_1), d(l_2, i_2)\|_2}$$

$M$ is the preference score over point pair $(i_1, i_2)$, defined by distance between two low level descriptors $d(\cdot, \cdot)$.

Classical matchings could be formulated as ranking based on $M(\cdot, \cdot)$

Semantics should be incorporated in SfM to be robust to moving objects

$$M(i_1, i_2) = \frac{\text{Semantic}(I_1, I_2)[i_1, i_2]}{\|d(l_1, i_1), d(l_2, i_2)\|_2}$$

Use the FCN as a semantic term

$$\text{Semantic}(I_1, I_2)[i_1, i_2] = FCN(I_1)[i_1] \cdot FCN(I_2)[i_2]$$
City Turning Example
Lots of Moving Vehicles Example
Recover the subtle car backing behavior
Experiments – Continuous Actions

Lane following: left and right
Experiments – Continuous Actions

Intersection
Experiments – Continuous Actions

Side Walk
End-to-end Learning of Driving Models from Large-scale Video Datasets

Huazhe Xu, Yang Gao, Fisher Yu, Trevor Darrell

(Submitted on 4 Dec 2016)

Robust perception-action models should be learned from training data with diverse visual appearances and realistic behaviors, yet current approaches to deep visuomotor policy learning have been generally limited to in-situ models learned from a single vehicle or a simulation environment. We advocate learning a generic vehicle motion model from large scale crowd-sourced video data, and develop an end-to-end trainable architecture for learning to predict a distribution over future vehicle egomotion from instantaneous monocular camera observations and previous vehicle state. Our model incorporates a novel FCN-LSTM architecture, which can be learned from large-scale crowd-sourced vehicle action data, and leverages available scene segmentation side tasks to improve performance under a privileged learning paradigm.
Overview

Adversarial Domain Adaptation

Learning end-to-end driving models from crowdsourced dashcams

Vision and Language: Learning to reason to answer and explain
Explainable AI (XAI): Visual Explanations

Western Grebe
This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

Laysan Albatross
This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.
Explainable Models with Implicit Capabilities

• Translate DNN hidden state into
  • human-interpretable language
  • visualizations and exemplars

Can you park here?
Visual Question Answering
How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Who is wearing glasses?
man
woman

Where is the child sitting?
fridge
arms

Is the umbrella upside down?
yes
no

How many children are in the bed?
2
1
NAACL 2016: MCB with Attention

- Predict spatial attentions with MCB

[Diagram with MCB and related operations]
Winner VQA Challenge 2016 (real open ended)
Attention Visualizations

What is the woman **feeding** the giraffe?

**Carrot**

[Groundtruth: Carrot]
What color is her shirt?

Purple

[Groundtruth: Purple]
Attention Visualizations

What is her hairstyle for the picture?

**Ponytail**

[Groundtruth: Ponytail]
What color is the chain on the red dress?

**Pink**

[Groundtruth: Gold]

- Correct Attention, Incorrect Fine-grained Recognition
Attention Visualizations

Is the man going to fall down?
No
[Groundtruth: No]
Attention Visualizations

What is the surface of the court made of?
Clay
[Groundtruth: Clay]
Attention Visualizations

What sport is being played?

Tennis

[Groundtruth: Tennis]
Attentive Explanations: Justifying Decisions and Pointing to the Evidence
Q: What is the person doing?

A: Skiing

Because:
They are on skis and going down a mountain

Q: What is the person doing?

A: Skiing

Because:
He is on a snowy hill wearing skis and clothing appropriate for skiing
Human ground truth for the textual justification task.

Description
A man on a snowboard is on a ramp.

A gang of biker police riding their bikes in formation down a street.

Description
A man in a black shirt and blue jeans is holding a glowing ball.

A man standing wearing a pink shirt and grey pants near a ball.

Explanation
Q: What is the person doing?
A: Snowboarding
Because... they are on a snowboard in snowboarding outfit.

Q: Can these people arrest someone?
A: Yes
Because... they are Vancouver police.

Description
I can tell the person is juggling
Because... he holds two balls in one hand, while another ball is aloft just above the other hand.

Because... he has two balls in his hands while two are in the air.
Human ground truth for the pointing task.

- **Q:** What is the person doing?  
  **A:** Skiing

- **Activity:** Mowing Lawn

- **Q:** What is the boy doing?  
  **A:** Skateboarding

- **Activity:** Planting, Potting

- **Q:** What game are they playing?  
  **A:** Baseball

- **Activity:** Bicycling, Mountain
Discussing different evidence for different images.

Q: Where is this picture taken? A: Airport
Because there are planes and trucks parked on the tarmac

Because there is a baggage carousel
Discussing different evidence for different questions.

*Q: Is this a social event? A: Yes*
Because they are many people gathered together

*Q: What game are they playing? A: Soccer*
Because they are kicking a soccer ball
Discussing different evidence for different questions.

*Q:* Is this a social event? *A:* Yes
Because they are many people gathered together

*Q:* What game are they playing? *A:* Soccer
Because they are kicking a soccer ball
Differentiating between some activities requires understanding special equipment.

I can see that he is windsurfing
Because he is standing on a windsurfing board and holding on to the sail

I can see that he is kayaking
Because he is sitting in a kayak and using a paddle in his hands

I can see that he is canoeing
Because he is sitting in a canoe and paddling with a paddle in the water
Differentiating between some activities requires recognizing specific context.

*I can see that he is bicycling, BMX*
Because he is riding a bmx bike and doing a trick on a low wall

*I can see that he is bicycling, racing and road*
Because she is wearing a bicycling uniform and riding a bicycle down the road

*I can see that he is bicycling, stationary*
Because he is sitting on a stationary bike with his feet on the pedals
Differentiating between some activities requires recognizing specific context.

I can see that he is bicycling, BMX
Because he is riding a bmx bike and doing a trick on a low wall

I can see that he is bicycling, racing and road
Because she is wearing a bicycling uniform and riding a bicycle down the road

I can see that he is bicycling, stationary
Because he is sitting on a stationary bike with his feet on the pedals
Explanations when the model predicts the wrong answer.

**Q: What is the bear doing? GT = Swimming, P = Eating**

Because it is hungry and likes food

**Q: Should we stop? GT = Yes, P = No**

Because the light is green

**GT = Piano, Sitting, P = Carpentry, General**

Because he is standing in a workshop with many tools on the table

**GT = Manual or Unskilled Labor, P = Yoga, Power**

Because he is sitting on a yoga mat and holding a yoga pose
Explainable Models with Explicit Capabilities

Explain higher-level reasoning in DNNs

- Explainable decision path for multi-task, control and planning
- Provide structure and intermediate state

Can you park here?
Explainable Models with Explicit and Implicit Capabilities

(a) NMN for the question *What color is his tie?*

(b) NMN for the question *Is the bus empty?*
   No, because there is a person in the bus.
Is there a red shape above a circle?

Yes
Is there a red shape above a circle? Yes

[Iyyer et al. 2014, Bordes et al. 2014, Yang et al. 2015, Malinowski et al., 2015]
Is there a red shape above a circle?

Neural module networks learn both!

Is there a red shape above a circle?

yes
Neural module networks

Is there a red shape above a circle?

- red
- exists
- above

↦  true
Neural module networks

Is there a red shape above a circle?

- `red` => true
- `exists` => true
- `above` => false
Neural module networks

Is there a red shape above a circle?

- red
- exists
- above

\[ \text{true} \]

\[ \text{yes} \]
Questions in CLEVR test various aspects of visual reasoning including attribute identification, counting, comparison, spatial relationships, and logical operations.
Learning to Reason: End-to-End Module Networks for Visual Question Answering

R. Hu, J. Andreas, M. Rohrbach, T. Darrell, K. Saenko
Background

Natural language is **compositional**: the meaning of a sentence comes from the meanings of its components.

Different questions may require significantly different reasoning procedure.

- *What kind of vehicle is the one on the left of the brown car that is next to the building?*
- *Why is the person running away?*
Background

- Neural Module Networks: **dynamic** inference structure for each question
- Previous work: structure from NLP parser or parser re-ranking
- This work: **learned** layout policy to dynamically build a network

There is a **shiny** object that is right of the **gray metallic cylinder**; does it have the **same size** as the **large rubber sphere**?
End-to-End Module Networks (N2NMN)

Components

- Layout policy $p(l \mid q)$ with sequence-to-sequence RNN
- Neural modules with co-attention, dynamically assembled into a network
End-to-end Training

Loss $L(\theta) = E_{l \sim p(l|q; \theta)}[\tilde{L}(\theta, l; q, I)]$

where $\tilde{L}(\theta, l; q, I)$ is the softmax loss of the answer

Optimization: policy gradient method

$$\nabla_\theta L = E_{l \sim p(l|q; \theta)} \left[ \tilde{L}(\theta, l) \nabla_\theta \log p(l|q; \theta) + \nabla_\theta \tilde{L}(\theta, l) \right]$$

$$\approx \frac{1}{M} \sum_{m=1}^{M} \left( \tilde{L}(\theta, l_m) \nabla_\theta \log p(l_m|q; \theta) + \nabla_\theta \tilde{L}(\theta, l_m) \right)$$

Easier: behavior cloning from expert layout policy
Experiments on the CLEVR dataset

How many other things are of the same size as the green matte ball?

Does the blue cylinder have the same material as the big block on the right side of the red metallic thing?
<table>
<thead>
<tr>
<th>Method</th>
<th>Overall</th>
<th>Exist</th>
<th>Count</th>
<th>Compare Integer</th>
<th>Query Attribute</th>
<th>Compare Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN+BoW [25]</td>
<td>48.4</td>
<td>59.5</td>
<td>38.9</td>
<td>Equal: 50</td>
<td>Size: 56</td>
<td>Size: 52</td>
</tr>
<tr>
<td>CNN+LSTM [4]</td>
<td>52.3</td>
<td>65.2</td>
<td>43.7</td>
<td>Less: 57</td>
<td>Color: 32</td>
<td>Color: 54</td>
</tr>
<tr>
<td>CNN+LSTM+MCB [9]</td>
<td>51.4</td>
<td>63.4</td>
<td>42.1</td>
<td>More: 60</td>
<td>Material: 59</td>
<td>Material: 51</td>
</tr>
<tr>
<td>CNN+LSTM+SA [24]</td>
<td>68.5</td>
<td>71.1</td>
<td>52.2</td>
<td>Size: 68</td>
<td>Shape: 87</td>
<td>Shape: 52</td>
</tr>
<tr>
<td>ours - cloning expert</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ours - policy search</td>
<td>78.9</td>
<td>83.3</td>
<td>63.3</td>
<td>68.2</td>
<td>90.5</td>
<td>89.4</td>
</tr>
<tr>
<td>after cloning</td>
<td>83.7</td>
<td>85.7</td>
<td>68.5</td>
<td>73.8</td>
<td>93.1</td>
<td>92.6</td>
</tr>
</tbody>
</table>

**Question:** Do the small cylinder that is in front of the small green thing and the object right of the green cylinder have the same material?

**Ground-truth answer:** No

**Cloning expert**

**End-to-end optimization after cloning**
Policy Search from scratch (no experts used)

Even without resorting to an expert policy during training, our method still achieves state-of-the-art performance with reinforcement learning from scratch.

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN+BoW [4]</td>
<td>48.4</td>
</tr>
<tr>
<td>CNN+LSTM [1]</td>
<td>52.3</td>
</tr>
<tr>
<td>CNN+LSTM+MCB [2]</td>
<td>51.4</td>
</tr>
<tr>
<td>CNN+LSTM+SA [3]</td>
<td>68.5</td>
</tr>
<tr>
<td>ours - policy search from scratch</td>
<td>68.5</td>
</tr>
<tr>
<td>ours - cloning expert</td>
<td>78.9</td>
</tr>
<tr>
<td>ours - policy search after cloning</td>
<td>83.7</td>
</tr>
</tbody>
</table>
Accuracy v.s. Question length

Even on long questions with 30+ words, our method still achieves relatively high accuracy (Figure a).
question: there is a shiny object that is right of the gray metallic cylinder; does it have the same size as the large rubber sphere?

ground-truth answer: "yes"  predicted answer: "yes"

output from find[0]
output from relocate[1]
output from filter[2]
output from find[3]
output from compare[4]

"yes"
Overview

Adversarial Domain Adaptation

Learning end-to-end driving models from crowdsourced dashcams

Vision and Language: Learning to reason to answer and explain