Towards a Visual Turing Challenge (May 5, 2015)

Malinowski et al. (Max Planck Institute for Informatics, Germany)

Intro

- Combination of perception/language progress in nets (references described below)
    - Grounding
        - Map sentence+object set to object subset in sentence
        - Image=>knowledge base, sentence=>logical form, together produces grounding/denotation
    - Language gen. From image/video
    - Image to sentence alignment
    - Recent question-answering problems
        - Auto-learn logical trees from (q,a) pairs
        - Resolve mismatch of question/knowledge base with paraphase model
        - Answer quiz bowl style questions with few named entities
        - Combine both “curated” and “extracted” knowledge bases
        - DAQUAR paper
- Aside: more from original DAQUAR paper (description of dataset)
    - Motivation
        - Vision type of tasks: labeling pixels, regions, bounding boxes
            - Uncertainty comes from limitations in features, data, ambiguity
        - Many question-answer methods
            - Uncertainty about fact correctness
        - Combine these two domains to test the whole chain of reasoning
            - “Perception, language understanding, deduction”
    - Dataset
        - Build on top of NYU-Depth V2 (1449 RGBD image, 894 classes), indoor scenes like bathrooms, kitchens, offices
        - Mapped all pixels into 40 classes (37+other structure/furniture/prop)
        - Fit 3D cuboids to segmentations
        - 795 training, 654 test
        - Synthetic QA, based on templates: counting, room type, colors, superlatives, negations
        - Human QA (12468 pairs gathered from 5 participants probably grad students), answers must be basic colors, etc. no constraint/error correction on questions
            - 9 per image, >400 tables/chairs in answers, but average is (14.25,4) for training/(22.485.75) total (using (mean,trimean) notation)

- With greater task complexity how do we craft a good benchmark?
    - Good evaluation requires human judgment
    - Large scale/domain is hard
    - Inherent ambiguities
- Essentially, use multiple ground truths (“social consensus”)

Challenges

Vision and language

Scalability: instances, cartegories, spatio-temporal concepts in the thousands
Concept ambiguity: inevitable with more categories (“armchair” or “sofa”)
Attributes: sometimes cannot be learned, but inferred from noun (“white elephant vs. white snow”)
Ambiguity in reference resolution: culture/context dependent, research for this is using symbolic+vector approaches

Common sense knowledge

Co-occurrence, parts, associations, etc.

Defining a benchmark/quantifying performance

- VQA is harder than grounding (less limitations)

DAQUAR

- 1088 nouns in Q, 803 nouns in A, 1586 together (573 categories of singular nouns in Q)
- 10.5 word questions, var 5.5, max 30
- Language errors
- They hypothesize common sense helps answer questions
    - “Which object on the table is used for cutting?” (knowing cutting helps)
    - “What is above the desk in front of scissors?”
    - Some annotators infer missing object parts
- Grounding is a subgoal of understanding intent of question asker

Quantifying performance

- Best evaluation requires deep understanding of language/intention
- Multiple correct answers from ambiguity
- We need similar scores for equivalence class of answers outside of some lexical database

WUPS scores

min(how good humans match architecture (product over all architecture answers), how good architecture matches humans (product over all human answers))

metric is motivated by the development of a 'soft' generalization of accuracy that takes ambiguities of different concepts into account via the set membership measure μ:

$$\frac{1}{N}\sum_{i=1}^{N}\min\{\prod_{a\in A^i}\max_{t\in T^i}\mu(a,t),\ \prod_{t\in T^i}\max_{a\in A^i}\mu(a,t)\}\cdot 100 \quad (1)$$

where for each i-th question, A^i and T^i are the answers produced by the architecture and human respectively, and they are represented as bags of words. The authors of [27] have proposed using WUP similarity [62] as the membership measure μ in the WUPS score. Such choice of μ suffers from the aforementioned coverage problem and the whole metric takes only one human interpretation of the question into account.

WUP is 2*depth(least common ancestor)/(depth(s1)+depth(s2)) (Wu Palmer similarity)

Future

- Alternate metrics (run max of eq1 over al human answers (similar to one human answer), or consider mean (agree with most human answers))
- Subtask without auxiliary data/with explicitly listed aux. data

[VQA: Visual Question Answering](#) (April 20, 2016)

Agrawal et al. (Virginia Tech, Microsoft Research, FAIR)
Antol et al. in some citations (but he's demoted in the latest arXiv version??)

Intro

- Image captioning is not very AI-complete, just need coarse image understanding+caption
- Need a well-posed task for "next gen of AI algs"
    - Multi-modal knowledge (beyond just CV)
    - Well-defined quantitative evaluation metric
- Paper introduces the task of "free-form" and "open-ended" VQA
    - Free-form: natural language answer as output, nat lang Q and image as input
    - Open ended: fine-grained recognition, object detection, activity recog, knowledge based reasoning, common sense reasoning
- Paper defines open-ended and multiple choice task
- New dataset of 204,721 images from COCO
- Dataset of abstract scenes to remove need to parse images
- 760K images, 10M images

Related work

- DAQUAR is "closed world"
- 2 orders of magnitude more than DAQUAR (I guess only when you count the synthetic abstract scenes)
- Text-based Q&A, captioning images
- Other vision+language tasks: coreference resolution, generating referring phrases

Collection

- 123,287 train/val, 81,434 test images from COCO
- Abstract scenes 50k, 20 posable humans+100 objects+31 animals
  - Also get 5 captions to match COCO
  - test-dev/-standard/-challenge/-reserve
- Questions must require image
- Ask people what a toddler/alien/smart robot would have trouble answering
- 3 questions from unique workers, shown previous questions
- 10 answers from unique workers (matter-of-fact, no opinion, not sentence)
  - Also asked if they answered question correctly (yes/no/maybe)
  - Accuracy=min(humans with that answer/3, 1)
- Multiple choice: 18 candidate answers
  - Correct (1): most popular human answer
  - Plausible (3): humans don't see image and answer, if <3 then drawn from bag of words nearest neighbor questions
  - Popular (10): y/n, red/white/blue/green, 1-4
  - Random to fill the rest

Analysis

- 614,163 questions/7,984,119 answers for 204,721 images from COCO
- 150k questions/1,950,000 answers for 50k abstract scenes
- These numbers are 13 answers/question??
- Question diversity similar for both abstract/real (4-10 words mostly)
- 90-6-3 roughly for 1 word/2 word/3 word answers, 23,234 unique one-word answers for real/3,770 for abstract
- ~40% are y/n questions, 58% yes for real
- ~12% are number questions (2 is 26% for real/40% for synthetic)
- Good interhuman agreement (exact string matching), 2.7 unique answers/real (2.39/abstract)
  - 95% for y/n, <76% for others
- Evaluate human performance on question, question+caption, question+image on 3k question train subset (1k images)

- Study on which need common sense: 10k questions, asked whether or not they believed you need common sense to answer, also youngest age group that could answer
- Human Q+I>Q+C, question N/V/A statistically different than captions

Baselines

- Random: random answer from top 1k answers
- Yes (prior): always pick yes
- Per Q-type prior (word2vec nearest if unavailable)
- Nearest neighbor (word2vec nearest to most common answer of nearest questions (measured by skip-thought feature space))

Methods

- Image embeddings: VGG 4096 (+normalized version)
- Question embeddings: bag of words (1000+30 for first 3 words), LSTM/deeper LSTM (1024)
- Perceptron: concat bag of words with image, or element wise multiply LSTM by image

Results

- Vision alone does worse than “yes”
- Perf is best for deep lstm + normalized embedding
- Perf is good for common objects, bad for high counts
- Model equivalent to 4.74 year old child for validation (8.98 is average needed)

VQA link

Making the V in VQA Matter (May 15, 2017)

Goyal et al. (Virginia Tech, Georgia Tech, Army Research Laboratory)

Abstract

Turns out models were ignoring visual information. Let’s create another VQA dataset with the same questions but different answers to make models unlearn that.

Intro

- Language prior can result in good perf so vision is ignored

- “Visual priming bias”: subjects see image while asking questions (“is there a clock tower”, “do you see a” is always yes)
- Counter this with a balanced dataset: given (I,Q,A) have human pick I’ (nearby in fc7 space) so A’ != A
- Increase entropy of P(A|Q)
- Created double size dataset, evaluated SOTA models which did much worse
- Create interpretable model (counter-example based explanation)

Related work

- Benchmark one baseline model, attention based model, 2016 winner, language only model
- This work is collecting hard negatives
  - Hodosh et al. used machine rules to generate two similar captions
  - Zhang et al. allowed modification of VQA abstract scenes to change binary question answer, but this paper allows for all question types/real images
- Other “models with explanation”
  - Hendricks et al. generates natural language explanation
  - Other models have spatial maps highlighting important regions of focus
  - Counter-example is the novelty here

Dataset

- Given (I,Q,A), show 24 nearest neighbor images (VGG fc7 L2 dist.) and ask for which I’ does Q make sense and answer is not A, get 10 new answers for it
- Also allow selection of “not possible”: usually when object is small or opposite concept is rare (bananas that are not yellow) -- 22% of all questions in VQA
- 135K questions had “not possible”
- Overall new dataset is 443k train/214k val/453k test (question, image) pairs
- A=A’ about 9% of the time
- Entropy (averaged across question types) increased 56%, yes/no very balanced

Benchmark

- Deep LSTM + normalized embedding from VQA paper (1k answers)
- Hierarchical co-attention: hierarchical modeling of question and image (1k answers)
- Multimodal Compact Bilinear Pooling: efficient approximate bilinear pooling (picks from 3k answers)
- Baselines: yes only, language only
- Models trained on unbalanced are worse on balanced, trained on balanced does better, 2x data does even better
- We can analyze (I,I’) statistics: training on the balanced dataset reduces the amount of time the guessed A’=A

- Most improved categories are yes/no (4.5%/3% for MCB/HieCoAtt), number (3%/2%), whereas previously minimal improvements seen, suggesting that language priors result in similar accuracies

Counterexample explanations

- Model takes in (Q,I), outputs A_pred
- We could take nearest neighbors, pick one with lowest P(A_pred), however Q might not make sense
- Do supervised training using the info that I' is a counterexample for (Q,A) from the set of neighboring questions I_NN
- Test time, you predict A, then use Q/A/I_NN to get an I'
- Model contains trunk and two heads (for creating A and picking I')
  - Trunk creates QI embedding (elem-wise mult) for I+I_NN
  - Answer head: fc+softmax to get A
  - Explaining head: Transforms QI and A to same space, computes inner products fed to FC layer to get scores
  - Hinge loss term to force human-selected I' to rank higher by some margin

Results

- Compare to picking random from I_NN, picking closest in I_NN, picking lowest P(A_pred) from I_NN
- Human evaluation using Top 5 recall (I' in top 5), this is only slightly better than picking distance