Project ideas -- brief descriptions

Adriana Kovashka

December 16, 2019

*Please do not share these ideas with others beyond the class.*

1. **Logical representations for common-sense reasoning**

The community has made great progress on purely visual tasks such as object recognition, but vision is just part of general intelligence and lives in a broader context. Humans can perform higher-order reasoning about the relationships and intents behind objects, e.g. two people hugging might be friends, a person wearing a suit and speaking in front of an audience may be a lawyer, the sidewalk turns a darker color when it rains, water can be poured in concave objects, etc. Being able to make these kinds of inferences is important for developing the future generation of agents that can interact with human users and reason about goals and intents.

There are some recent datasets developed on common-sense reasoning. In computer vision, one dataset is VCR (<https://visualcommonsense.com/> ). Another is the Pitt Ads dataset (<http://people.cs.pitt.edu/~kovashka/ads/> ) where we ask what action an advertisement prompts, and what arguments it provides for taking the suggested action (also referred to as action-reason, question-answering).

Several approaches exist that use graph convolutional networks and their variants, to deduce an answer using some of the evidence in the image (e.g. recognized objects and attributes). However, complete reasoning involves multi-hop reasoning paths from evidence to answers.

One option to enable such reasoning is through first-order logic which allows us to make an inference using a set of premises. Recent work proposes a differentiable approach called Neural Logic Programming. However, this approach has only been used for symbolic (non-vision) data.

The goal of this project is to adapt the Neural Logic Programming (Neural LP) approach to the VCR/Ads dataset. What will the nodes that are represented in logic be? For example, they can be entities recognized in the images using standard visual classifiers; they can be entities from knowledge bases (e.g. ConceptNet), etc.

Reading:

* VCR paper (see website, also check out leaderboard and some methods there)
* <https://deanplayerljx.github.io/tabvcr/> (recent method for VCR)
* <https://papers.nips.cc/paper/7531-out-of-the-box-reasoning-with-graph-convolution-nets-for-factual-visual-question-answering.pdf> (related paper using graph convolutional networks for visual question answering, a related task)
* <https://arxiv.org/abs/1702.08367> (Neural Logic Programming)

1. **Vision to aid in common sense**

Datasets also exist for commonsense reasoning purely in the textual space, such as the Sense Making dataset (<https://arxiv.org/abs/1906.00363> ). The dataset asks an agent to predict which of two statements make sense, e.g. “put a turkey in the fridge” or “put an elephant in the fridge”. One cue humans use to do the task is they can visually imagine these statements, and they can infer which ones are possible. To an extent, the Sense Making dataset relies on understanding physics (e.g. sizes of objects relative to each other). Some physical relationships can be learned from images, by measuring sizes of detected objects and relating them to each other (e.g. dogs are typically smaller than persons). Exploratory work on this was done in VisKE (<https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Sadeghi_VisKE_Visual_Knowledge_2015_CVPR_paper.pdf> ) but this approach did not leverage neural representations.

The goal of this project is to explore different ways of using visual data to help with the tasks in Sense Making. Start with the most direct approach, i.e. the easiest adaptation of VisKE to Sense Making. Then think of a few improvements over the basic approach. For example, one can look up phrases in Google Images, compare counts, train classifiers, etc.

1. **Geometric/physical relationships for VCR**

Several tasks in VCR (see above) rely on reasoning about relationships between objects, even purely within the image. For example, one needs to reason about proximity of people to different objects, orientation of people (i.e. are they looking at each other), etc. This relates to techniques for detecting objects based on the context of surrounding objects (see readings below); the difference is here we aim to answer questions about images, not detect objects, but the modeling is similar (reason about regions and relationships between them). Prior approach model relationships between regions through proximity (distance), visual similarity, etc. Here, we need to model additional characteristics; e.g. for people, we should model orientation of the head, body pose (for which plenty of off-the-shelf techniques exist). Physical properties of objects (such as “heavy”, which can be obtained from a knowledge base) can be incorporated. Optical flow (direction of movement) can be “hallucinated” to aid in the common sense reasoning (see references below). Human-object interactions (HOI) are also helpful; see some readings below.

The goal of this project is thus to adapt context-based detection methods, by adding new features to them, for the task of visual commonsense reasoning (VCR).

The same features can also be applied on the ads understanding task i.e. the Pitt Ads dataset (see above). The task there is to select the correct action/reason statement with the help of inferred optical flow features, human-object interactions, etc.

Reading:

* VCR dataset, paper, leaderboard and methods in leaderboard (see above)
* Pitt Ads dataset (see above)
* Structure inference net: <https://arxiv.org/abs/1807.00119> , <https://github.com/choasup/SIN>
* Reasoning RCNN: <http://openaccess.thecvf.com/content_CVPR_2019/papers/Xu_Reasoning-RCNN_Unifying_Adaptive_Global_Reasoning_Into_Large-Scale_Object_Detection_CVPR_2019_paper.pdf>
* Region proposals: Faster RCNN, EdgeBoxes
* Optical flow hallucination: <https://arxiv.org/abs/1505.00295> , <http://vision.cs.utexas.edu/projects/im2flow/>
* Human-object interactions <http://www-personal.umich.edu/~ywchao/hico/> , <https://gkioxari.github.io/InteractNet/index.html>

1. **VCR / Sense Making / Ads two tasks**

The VCR dataset (see above) contains two prediction tasks, one main task and one “explanation” task. The same is true for the Sense Making dataset (also see above). It is true in the Pitt Ads dataset as well (<http://people.cs.pitt.edu/~kovashka/ads/> ) where we ask what action an advertisement prompts (the “what” task), and what arguments it provides for taking the suggested action (the “why” task). While the second task is supposed to help the former, most methods only do this in the form of multi-task learning (i.e. train jointly for both tasks, which means that the tasks indirectly help each other). What ways are there to directly force the model for the first task to use the explanation task, explicitly? What are different ways to connect the tasks, beyond simply joint training?

1. **Domain adaptation in VQA and common sense**

Common sense should transfer across dataset boundaries. Visual question answering (VQA) is an earlier form of common sense, for which there are several datasets (see below). There are also several datasets for common sense in language and beyond (see below). How well do models trained on one dataset transfer to another dataset? Do accuracies remain steady without additional training? If not, how much training on the target dataset is needed to achieve comparable training to that on the source dataset? (Source dataset is the one in which plentiful data is available, whereas target is a disjoint dataset on which we have sparser data.)

The goal of this project is to experiment with domain generalization and domain adaptation approaches in the relatively novel context of VQA and commonsense reasoning (very limited work exists for domain adaptation in VQA). You want to analyze drops in performance as a function of similarities between datasets—it is up to you how you define similarity, as long as it can be automatically computed. You also want to explore ways to regularize and ensure generalization across domains, whether novel or from existing approaches. Is there a benefit to training on multiple source VQA datasets? What are the reasons for variations across datasets?

Reading:

* Domain adaptation survey (not on VQA): <https://arxiv.org/pdf/1802.03601.pdf>
* Different VQA datasets: <https://visualqa.org/> , <https://vizwiz.org/tasks-and-datasets/vqa/> , <http://www.cs.toronto.edu/~mren/research/imageqa/data/cocoqa/> , <https://arxiv.org/abs/1606.05433> , <http://openaccess.thecvf.com/content_CVPR_2019/papers/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.pdf> , <http://dosa.cds.iisc.ac.in/kvqa/KVQA-AAAI2019.pdf> , <https://arxiv.org/abs/1511.03416> , <https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/vision-and-language/visual-turing-challenge/>
* Different common sense datasets: many can be found authored by <https://homes.cs.washington.edu/~yejin/>

1. **Image-text complementarity, whole is bigger than parts**

Modern media contain multimodal data, i.e. both images and text. Understanding the relationship between the two is not easy. Unlike image captioning approaches, which generate a textual caption that literally describes the image, the text that appears with images in news articles does not literally describe those images. Further, the interaction of image and text conveys new meaning that goes beyond the simple combination of the two separate modalities—i.e. “the whole is bigger than the sum of its parts”. Consider a public service announcement warning the audience about the dangers of human trafficking, that shows a woman “dressed” in cardboard boxes, with the words “winter fashion” near it. The image symbolically refers to homelessness (through the cardboard boxes) and juxtaposes the drama of homelessness with the luxury of fashion. Importantly, the image alone, and text alone, do not convey the same dramatic point.

The goal of this project is to explore the types of relationships between image and text—are image and text redundant (as in the case of captioning), complementary (as in the case of the PSA example), contradictory, etc? In other words, the goal is to predict the type of relationship, using one or both of the following datasets: Pitt (<http://people.cs.pitt.edu/~mzhang/ads_parallelity/> ) and SRI/Stanford (<https://github.com/karansikka1/documentIntent_emnlp19> ). One suggested approach is to represent regions in the image and words in the text as nodes in a graph, then jointly reason over the graph to eventually produce a label for the relationship. An additional approach is to hide parts of the text or image, and check whether a side task (for which data exists, e.g. action/reason in the Pitt Ads dataset) can still be predicted – if so, the channels (image and text) might be redundant.

1. **Differences in captions**

Captions are short sentences crowdsourced from human annotators that describe an image. For example, the COCO dataset has five captions per image, written by different annotators. The captions for the same image are not the same. What can a model learn from the *differences* in captions? Further, what can it learn from the differences between what is seen (e.g. captured through image labels) and what is mentioned in a caption? What content do humans always agree on and mention, and what content is omitted by some/all annotators? How can we learn to better detect objects from this week supervision, or better produce captions, by explicitly capturing inter-annotator differences? Can we learn better image or text representations by training a model to predict not just an “average” over the different annotations, but the “differences” between them?

Reading:

* Captioning: <https://cs.stanford.edu/people/karpathy/cvpr2015.pdf> (and many more)
* Human reporting bias: <https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Misra_Seeing_Through_the_CVPR_2016_paper.pdf>
* Object detection from captions: <http://openaccess.thecvf.com/content_ICCV_2019/papers/Ye_Cap2Det_Learning_to_Amplify_Weak_Caption_Supervision_for_Object_Detection_ICCV_2019_paper.pdf>
* Modeling human uncertainty in labels: <https://arxiv.org/abs/1908.07086>

1. **Curriculum learning for weak multimodal supervision**

****

As discussed above, modern media contains plentiful multimodal data, specifically images paired with text, and the relationship between image and text is complex, and the two channels are not redundant (see example above). How can we learn to retrieve matching text for an image, or matching images for a text? Captioning approaches do this for image-text pairs that are redundant. Given the challenge of the task, the goal of this project is to explore a curriculum learning approach, where the model learns easier image-word relations first (i.e. learn to predict easier words from images), then more complex ones. A variety of datasets exist for this task: for example, Conceptual Captions (<https://ai.google.com/research/ConceptualCaptions/> ), the Pitt Politics dataset (<http://people.cs.pitt.edu/~chris/politics/> ), the Good News dataset (<https://github.com/furkanbiten/GoodNews> ), the WLD dataset (<https://github.com/hellock/WLD> ). In all, you should aim to predict words from images. At its most basic, each word that co-occurs with an image or video can be assumed related to this video. However, some image-text relationships are more obvious and easier to learn; some words have a more visually concrete meaning (i.e. images corresponding to “tiger” are similar) while other words are less concrete (i.e. images corresponding to “mother” are less internally similar compared to “tiger” images). Thus, you need to develop a scoring mechanism for which words to learn to predict first, i.e. in a first stage of training. Once this training is complete, you can continue training the model with more abstract/challenging words.

1. **Unsupervised VQA/VCR**

To answer questions in VQA (see above), we normally require a dataset containing both questions and answers. However, the answers may not be fully necessary, as related information can be leveraged. For example, given a question “What is in this image” (for an image showing a dog), perhaps we only need an object detection model that can predict “dog”, rather than a question-answer pair from a VQA dataset. The same goes for the VCR dataset (also see above). One prior work that explores unsupervised image captioning exists (<http://openaccess.thecvf.com/content_CVPR_2019/papers/Feng_Unsupervised_Image_Captioning_CVPR_2019_paper.pdf> ). The goal of this project is to adapt this work in the VQA/VCR context, then come up with techniques to improve the work.

1. **WSOD from graph influence**

Weakly supervised object detection (WSOD) techniques aim to detect object boundaries in image, i.e. both recognize an object category in the image (e.g. “dog”) and localize the dog (with bounding box coordinates). Importantly, unlike standard object detection methods, WSOD methods do not have bounding box level supervision at training time, only image-level supervision (i.e. images but not regions labels with object classes). One new idea for WSOD is to examine the influence that a region has on other regions, when computing a prediction for some task of interest—e.g. classifying a whole image, predicting a caption, etc. We represent all regions proposed in an image with a graph, but we do not know which *region* corresponds to a particular class. All we have are image-level labels or captions. A graph convolutional network method (GCN) can learn the weights (contributions) of nodes (regions) on the overall performance. The weight can be leveraged to predict which regions carry the most importance, and thus are likely to contain the objects of interest.

Reading:

* Cap2Det (see above)
* Ctrl-f “weakly supervised object detection”: <http://openaccess.thecvf.com/CVPR2019.py> , <http://openaccess.thecvf.com/ICCV2019.py>
* GCNs: <https://tkipf.github.io/graph-convolutional-networks/>

1. **Detection with word embeddings**

Image classification and detection models aim to predict class IDs, but classes are not just IDs—they have semantics. Thus, predictions for “bicycle” and “motorcycle” should not be considered completely disjoint. Semantics in words are captured by word embedding models (look up word2vec, glove, bert, elmo). Imagine a classification method that predicts not IDs as in classification, but rather word embeddings for the class of interest (i.e. an L2 loss over the predicted and expected embedding). The predicted embedding can then ultimately map to an ID, but rely on an embedding prediction (loss) at the layer before the last one. You can experiment with this in the context of image classification, object detection, weakly supervised object detection, etc. An alternative is to still predict IDs, but have the second-to-last layer that aims to predict an embedding, and leverage pre-trained embedding models for this embedding as a secondary loss. Some language/word embeddings: BERT, ELMO, GPT-2, ROBERTA.

1. **Task ordering and meta learning**

Computer vision methods greatly benefit from transfer learning, i.e. pretraining on a dataset or task disjoint from the main dataset/task of interest. However, there is limited research on what source tasks are appropriate for what target tasks. Further, pre-training on some set of source tasks does not need to be done all at once; it can be sequentially ordered, similar to curriculum learning, to maximize gains. This project aims to learn the best order of tasks, using the Taskonomy dataset (<http://taskonomy.stanford.edu/> ). In particular, imagine there are three tasks, A, B, and C. In what order should a model learn them? How to ensure that an estimated order generalizes well to another set of tasks? This project is related to meta-learning. The figure below shows one possible order of learning tasks: object detection (in dark green), attribute recognition (in light green), semantic segmentation (in blue), depth estimation (in red) and surface normal estimation (in yellow).

. 

1. **Stick-figure story generation**

Human audiences are moved by stories. However, to maximize engagement, stories being told should contain some modicum of novelty. Since stories take effort to write, can we leverage generative approaches to tell visual stories? Generative approaches require massive data and a story has many elements, so generating all of them is infeasible. However, comic books show that stick-figure story telling is also engaging. If we work with stick figures, we can just generate the joints of the body, of which there is a sparse number (e.g. 26x2, 26 joints with x/y location each). Can we extract joints from massive video datasets to learn how humans move, then learn to represent stories/films at the joint level, and use that to generate new stories (new stickfigure motions)?

Relevant reading:

* Learning Individual Styles of Conversational Gesture (also leverages stick figures) <https://people.eecs.berkeley.edu/~shiry/projects/speech2gesture/index.html>
* Visual Storytelling <http://visionandlanguage.net/VIST/>

1. **Self-supervised learning: does dynamic prediction help for static recognition?**

Self-supervised learning is an important mechanism to leverage data without labels; according to many, it is the way of the future. The goal of this part of the project is to compare existing techniques for self-supervised learning, and examine what possible prediction tasks have been overlooking or not leveraged fully. For example, do tasks for dynamic prediction help to learn representations that are useful in a static setting as well. If we represent a video on the frame level, and learn to predict the next frame (or learn to predict whether two frames are in the correct order), can we leverage the learned CNN as a pre-training mechanism for static settings, e.g. recognition on images? How can we do dynamic prediction without the need to predict all pixels—does it make sense to infer locations or names of objects, rather than the full objects?

Relevant reading:

* Recent comparison papers <http://openaccess.thecvf.com/content_CVPR_2019/papers/Kolesnikov_Revisiting_Self-Supervised_Visual_Representation_Learning_CVPR_2019_paper.pdf> , <https://arxiv.org/abs/1905.01235> , <https://arxiv.org/abs/1912.01991>
* Classic dynamics prediction as pre-training <https://arxiv.org/abs/1603.08561>