Gordon Lu

CS 1674

Dr.Hwang

Essay 2:

Part I:

Paper I: VQA: Visual Question Answering

1) What is this paper trying to accomplish? (Think about what the current limitations of prior approaches are, and why these limitations are important.)

For a system to succeed at VQA, it typically requires a detailed understanding of the image and complex reasoning than a system producing generic image captions. Recent papers have begun study on VQA, but operate under fairly restrictive settings with small datasets, such as only considering questions whose answers come from a predefined closed world of 16 basic colors or 894 object categories. In contrast, this paper seeks to involve open-ended, free-form questions and answers provided by humans with the goal of increasing the diversity of knowledge and kinds of reasoning needed to provide correct answers. As current standards only analyze images at a broader or coarser level and are not "AI-complete." The proposed VQA approach ultimately attempts to formulate natural-language answers to natural-language questions on a given input image.

2) What is the high-level idea of **how** the paper will accomplish its goal?

The paper will present a large dataset containing over 760K questions with around 10M answers. The paper will also compare the performance of VQA to baseline VQA performance as well as human performance. In addition, VQA exceeds previous standards by incorporating knowledge outside a single domain and using a quantitative evaluation metric. By explaining these methods, this paper will discuss why a novel VQA approach is necessary.

Paper II: VizWiz Grand Challenge: Answering Visual Questions from Blind People

1) What is this paper trying to accomplish? (Think about what the current limitations of prior approaches are, and why these limitations are important.)

Prior approaches to this paper involved VQA datasets arising from a natural VQA setting that were not goal-oriented. This paper proposes VizWiz, a goal-oriented dataset arising from a natural VQA setting. VizWiz is not an ordinary dataset, it consists of images captured by blind photographers that are often poor quality, contains questions that are spoken and are more conversational and often visual questions that cannot be answered. This paper seeks to introduce VizWiz to encourage a larger community to develop more generalized algorithms that can assist blind people.

2) What is the high-level idea of **how** the paper will accomplish its goal?

The authors will build off previous work which established a mobile application that supported blind people to ask over 70,000 visual questions by taking a photo and asking a question about it. Furthermore, the authors implemented a rigorous filtering process to remove visual questions that could compromise the safety or privacy of any individuals associated with them. The authors then crowdsourced answers to support algorithm and training evaluation. The team was able to create more than 31,000 visual questions by asking blind people to take pictures and record spoken questions, and additionally, through the findings of running numerous experiments and algorithms for predicting answers and if a visual question can be answered, the authors were able to conclude VizWiz is a difficult dataset for modern vision algorithms and offered new perspectives about the VQA problem. The authors were able to also use these findings to simultaneously educate people about the technological needs of blind people, and provide new exciting opportunities for researchers to develop assistive technologies that eliminate accessibility barriers for blind people.

Paper III: Speech2Action: Cross-modal Supervision for Action Recognition

1) What is this paper trying to accomplish? (Think about what the current limitations of prior approaches are, and why these limitations are important.)

This paper seeks to demonstrate how human action can be guessed from dialogue alone. With prior approaches, attempts at aligning screenplays to the videos themselves and using that as a source of weak supervision is challenging due to a lack of explicit correspondence between scene elements in video and their textual descriptions in screenplays. Thus, the authors seek to show that at scale, supervision obtained from the speech-action correlation provides sufficient weak supervision to train visual classifiers, and obtain weak labels for action recognition, using speech alone.

2) What is the high-level idea of **how** the paper will accomplish its goal?

As mentioned above, a drawback of prior approaches is attempts to align screenplays to videos is challenging. The authors instead learn from unaligned movie screenplays First, by learning the correlation between speech and actions from written material alone, and using it to train a classifier, called Speech2Action. Furthermore, this classifier is applied to the speech in an unlabeled, unaligned set of videos to obtain visual samples corresponding to the actions confidently predicted from the speech. The authors notably trained a BERT-based Speech2Action classifier on the unaligned movie screenplays. The model was then applied on speech segments from unlabeled movies, taking the resulting predations to make weak action labels. The paper ultimate demonstrates that the proposed model performs better than the standard benchmarks without manual labeling.

Part II:

Paper I: Unsupervised Visual Representation Learning by Context Prediction

1) Summarize what this paper aims to do (what gap in science is it trying to address), and what its main contribution is, compared to what prior methods have already accomplished.

This paper aims to demonstrate that spatial context is a good source of training for understanding and categorizing images. By demonstrating that the feature representation learned using within-image context allows us to perform unsupervised visual discovery of objects. Additionally, the paper aims to improve upon the performance of a randomly-initialized ConvNet. By doing so, this new technique will allow for state-of-the-art performance while using unsupervised training, requiring less manpower than traditional techniques.

2) Summarize the proposed approach.

The paper provides a similar "self-supervised" formulation for image data, being a supervised task involving predicting the context for a patch. The self-supervised approach begins by sampling random pairs of patches in one of eight spatial configurations, and training a CNN to predict the second patch's position. By training based on space and context, the model is trained in a "self-supervised" manner. A visual representation can then be extracted from the model using a ConvNet. The authors then discuss the results of using a self-supervised approach for visual representation compares to prior unsupervised objection detection and visual data mining standards. The authors are able to conclude that this self-supervised approach works well across images and is effective on a categorical level, lending itself better than previous standards.

3) Summarize the experimental validation of the approach -- how is the proposed method tested, and what are the major observations and conclusions about its effectiveness?

Some of the pre-text concepts within this approach have already been explored, such as unsupervised representation and context prediction. Although these concepts have been explored in different domains (i.e. words rather than pixels), the concepts in general are considered valid and standard within the CV community. The approach aims to use ConvNets, which are widely-used, as a way of extracting an image representation from those pre-text tasks. The model is demonstrated to associate patch-pairs with k-nearest neighbors, then analyzed for performance on object detection and visual data mining.

Since the experimental applications perform relatively well in coverage and accuracy, the approach is valid.

4) What is one advantage of the proposed approach, beyond strong performance/accuracy?

One advantage of the proposed approach is that it uses unsupervised learning. Using unsupervised learning saves a lot of resources such as time and money, that would be spent on human annotation during labeling.

5) What is one disadvantage/weakness/limitation of the approach or experimental validation?

One disadvantage is that in implementing the approach, you have to be caregu about avoiding low-level cues that allow the algorithm to take shortcuts such as textures, and chromatic aberration. If this is done carelessly, the results could be overfit for training, but perform terribly during testing.

6) Suggest one possible extension of this approach, i.e., one idea for future work.

If the algorithm can accurately match image patches within the same image, and be applied to categorical images, a further extension of this work could be a similar task with videos. For example, an algorithm that would identify images within the same video, then having a ConvNet creating a representation of that allows for cross-video categorization would allow videos to be sorted without human labels or input.

Paper II: Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks

1) Summarize what this paper aims to do (what gap in science is it trying to address), and what its main contribution is, compared to what prior methods have already accomplished.

Image-to-image translation changes an input image to an output image based on training image pairs. However, more often than not, paired training data will not be available. This paper presents an approach for learning to translate an image from a source domain to a target domain in the absence of paired examples. Additionally, this paper will provide qualitative and quantitative results to demonstrate the superiority of this approach.

2) Summarize the proposed approach.

The approach seeks to learn a mapping $G : X \rightarrow Y$ from the source domain to the target domain such that the distribution of images from G(X) is indistinguishable from the distribution of Y using an adversarial loss, as well as an inverse mapping $F : Y \rightarrow X$. Furthermore, F will be the inverse of G in order to ensure that the mapping is correct. The mappings are trained at the same time, and their loss is computed with cycle consistency and adversarial loss using discriminators. The approach is applied to a variety of tasks where training data does not exist, including collection style transfer and object transfiguration.

3) Summarize the experimental validation of the approach -- how is the proposed method tested, and what are the major observations and conclusions about its effectiveness?

Both qualitative and quantitative results demonstrate that this approach, CycleGAN improves upon prior methods, namely those that require paired training data. This approach is compared to recent approaches, in which participants were not fooled in the direction of translation, but in CycleGAN, participants were fooled in a quarter of the trials. Then, the full method, which included adversarial and cycle consistency loss was compared against other methods which involved removing loss degraded results by causing mode collapse. Lastly, the authors demonstrated that the generalization of the method on applications without paired data such as object transfiguration and photo enhancement perform better than prior methods that do require paired training data.

4) What is one advantage of the proposed approach, beyond strong performance/accuracy?

One advantage is that a wide variety of applications such as object transfiguration and photo enhancement can be accomplished without paired image training data.

5) What is one disadvantage/weakness/limitation of the approach or experimental validation?

One disadvantage of this approach is that transformation involving geometric changes will suffer.

6) Suggest one possible extension of this approach, i.e., one idea for future work.

One possible extension of this approach could be that CycleGAN could be used as a security measure to hide identities in systems that monitor people on camera.