Emma Akbari

**Essay 2**

**Part I:**

Paper 1: [Deep Visual-Semantic Alignments for Generating Image Descriptions](https://cs.stanford.edu/people/karpathy/cvpr2015.pdf)

1. What is this paper trying to accomplish? (Think about what are the current limitations of prior approaches, and why these limitations are important.)
   1. This paper seeks to present a new model for generating natural language descriptions for images. It further generates descriptions of smaller regions within images. This model is necessary to improve upon previous standards of visual recognition, which only had fixed categories with closed vocabularies (much more restrictive than human recognition capabilities).
2. What is the high-level idea of **how** the paper will accomplish its goal?
   1. The paper will explain how a deep neural network model can infer the relationship between sentence fragments and image regions. Furthermore, it will describe how a multimodal Recurrent Neural Network can take input images then generate a text description. By comparing the generated descriptions with baseline descriptions, it will show the effectiveness of the suggested deep visual model.

Paper 2: [VQA: Visual Question Answering](https://openaccess.thecvf.com/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf)

1. What is this paper trying to accomplish? (Think about what are the current limitations of prior approaches, and why these limitations are important.)
   1. VQA needs a system to understand the intricacies of an image through complex reasoning; this is because it has to target areas of an image. This paper attempts explain that this approach is necessary because current standards only analyze images at a broader or coarser level and are not “AI-complete.” The VQA approach ultimately attempts to formulate natural-language answers to natural-language questions on an input image.
2. What is the high-level idea of **how** the paper will accomplish its goal?
   1. The paper will discuss the results of a dataset with around millions of images, questions, and answers. It will also compare this VQA performance with baseline VQA performance and human performance. Additionally, VQA reaches beyond previous standards by incorporating knowledge outside a single domain and using a quantitative evaluation metric. By explaining these methods, the paper will why a novel VQA approach is necessary.

Paper 3: [Making the V in VQA Matter](https://openaccess.thecvf.com/content_cvpr_2017/papers/Goyal_Making_the_v_CVPR_2017_paper.pdf)

1. What is this paper trying to accomplish? (Think about what are the current limitations of prior approaches, and why these limitations are important.)
   1. This paper attempts to fix the fact that vision and language tasks use models that ignore visual information. The way our language is built is simpler than the visual modalities, so the approach in this paper emphasizes the visual aspect of VQA to account for this difference. Thus, the imbalance of language priors would be abolished.
2. What is the high-level idea of **how** the paper will accomplish its goal?
   1. In order to account for language priors, this paper will explain how collecting complementary images and mapping pairs of different images to different answers. This balanced dataset is therefore more comprehensive than the original VQA dataset. In the end, this novel data collection method will allow the creation of a new model in which similar images with different answers will be identified, creating a more trustable machine.

Paper 4: [VizWiz Grand Challenge: Answering Visual Questions from Blind People](https://openaccess.thecvf.com/content_cvpr_2018/papers/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.pdf)

1. What is this paper trying to accomplish? (Think about what are the current limitations of prior approaches, and why these limitations are important.)
   1. Prior to this paper, VQA approaches were not goal-oriented; this paper proposes VizWiz, a VQA approach incorporating goals. The VizWiz dataset is a step up from previous datasets because of its photos which originate from blind people. This means that its applications can be more helpful to assisting blind people.
2. What is the high-level idea of **how** the paper will accomplish its goal?
   1. The study gathers images from blind people so that they are of poorer quality, forcing VizWiz to be more robust. Furthermore, the questions are crowdsourced from these same blind participants and are therefore more conversational than traditional VQA questions. The team achieved the creation of more than 31,000 visual questions by asking blind people to take pictures and record spoken questions.

Paper 5: [Speech2Action: Cross-modal Supervision for Action Recognition](https://openaccess.thecvf.com/content_CVPR_2020/papers/Nagrani_Speech2Action_Cross-Modal_Supervision_for_Action_Recognition_CVPR_2020_paper.pdf)

1. What is this paper trying to accomplish? (Think about what are the current limitations of prior approaches, and why these limitations are important.)
   1. This paper seeks to demonstrate how human action can be guessed from dialogue alone. The authors leveraged the already existing movies with screenplay actions and speech. In this way, they could train an effective model with weak supervision and already collected data.
2. What is the high-level idea of **how** the paper will accomplish its goal?
   1. The authors trained a BERT-based Speech2Action classifier on the screenplays I mentioned above. Then, they apply the model on speech segments from unlabeled movies, taking the resulting predictions to make weak action labels. The paper ultimately demonstrates that their proposed model performs better than standard benchmarks without manual labeling.

**Part II:**

Paper 1: [Unsupervised Visual Representation Learning by Context Prediction](https://openaccess.thecvf.com/content_iccv_2015/papers/Doersch_Unsupervised_Visual_Representation_ICCV_2015_paper.pdf)

1. Summarize what this paper aims to do (what gap in science it is trying to address), and what its main contribution is, compared to what prior methods have already accomplished.
   1. This paper seeks to show that spatial context is a solid source of training for understanding and categorizing images. By using a representation within-images using separate object parts, the paper aims to improve upon previous algorithms’ performance. The new technique provides state-of -the-art performance while using unsupervised training, so it requires less man power than other techniques.
2. Summarize the proposed approach in 3-5 sentences.
   1. The approach starts with taking random pairs of patches within an image and training a CNN to predict the second patch’s position. By training based on space/context, the model can be trained in a “self-supervised” manner. Then, a visual representation can be extracted from the model via a Conv-Net. Finally, the paper shows how this method of visual representation compares to previous (unsupervised) object detection and visual data mining standards. The approach also works across images and proves to be effective on a categorical level, making it more useful.
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?
   1. 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?
   1. One advantage of the proposed approach is that it uses unsupervised learning. This saves a lot of human resources (time, money) that would be spent on human annotation during labeling.
5. What is one disadvantage/weakness/limitation of the approach or experimental validation?
   1. One disadvantage is that when implementing the approach, you have to be careful about avoiding low-level cues which allow the algorithm to take shortcuts (e.g. textures, chromatic aberration). If not careful, the training could be accurate, but during test time would not perform well.
6. Suggest one possible extension of this approach, i.e. one idea for future work.
   1. If the algorithm can match image patches within the same image, and then be further applied to categorical images, perhaps with more research a similar methodology would be applicable to videos. For example, an algorithm would identify images within the same (probably short) video, and then a ConvNet could create a representation that then allows for cross-video categorization. This would allow for sorting of a video database without human input or labels.

Paper 2: [Learning Image Representations Tied to Ego-Motion](https://openaccess.thecvf.com/content_iccv_2015/papers/Jayaraman_Learning_Image_Representations_ICCV_2015_paper.pdf)

1. Summarize what this paper aims to do (what gap in science it is trying to address), and what its main contribution is, compared to what prior methods have already accomplished.
   1. This paper aims to bridge the gap between ego-motions and visual development. Other methods of visual learning are unable to extend their capabilities to transformations via ego-motion. This novel approach will allow visual development to be robust in response to transformations via ego-motion (i.e. equivariance rather than the narrower category of invariance), and it utilizes motor sensory cues while other approaches have not. It also performs better than previous visual recognition tests.
2. Summarize the proposed approach in 3-5 sentences.
   1. This approach uses proprioceptive motor signals (ego-motion yi and pixels xi in parallel) with CNN to gain understanding of visual representations with egocentric video. The feature learning is unsupervised so that the model can build more accurate visual representations. Then, a feature map with ego motion-based equivariance is generated using a feature mapping function z0.
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?
   1. The experimenters apply their approach to three datasets against two methods to verify its effectiveness. In all cases, this ego-motion approach performs better than previous related methods in terms of classification accuracy. Additionally, the results show that unsupervised learning feature methods perform better than supervised (performing around 30% better than baseline approaches). When qualitatively graphing the results in a feature difference space, the experimenters also found that ego-motion spaces show accurate transformations while other images did not.
4. What is one advantage of the proposed approach, beyond strong performance/accuracy?
   1. The proposed approach achieves equivariance that is more complex than in previous studies as it leverages 3D ego-motion rather than hand-crafted data.
5. What is one disadvantage/weakness/limitation of the approach or experimental validation?
   1. One limitation of this approach is that it requires accurate motion-sensing technology (this may be limited by budget, logistics, or other resources to the experimenter).
6. Suggest one possible extension of this approach, i.e. one idea for future work.
   1. Maybe this ego-motion approach could be extended to predicting actions: with a large data collection of ego-centric videos, a new motion input could be used to predict the next motion as an output.

Paper 3: [General Adversarial Nets](https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf)

1. Summarize what this paper aims to do (what gap in science it is trying to address), and what its main contribution is, compared to what prior methods have already accomplished.
   1. Past generative models have had little success due to the complexity of probabilities and using piecewise linear units. The approach in this paper proposes an effective generative model that avoids those obstacles and leverages discriminative models. Its main contribution is showing how to create a visually accurate generative model.
2. Summarize the proposed approach in 3-5 sentences.
   1. The proposed approach, known as the adversarial nets framework, plays a minimax game between a generative and discriminative model (both multilayer perceptrons). The discriminative model must determine whether a sample is generated or from the data distribution; the feedback between generative and discriminative is positive in which both improve based on it. Both models are trained using backpropagation and dropout algorithms.
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?
   1. The paper describes how they trained the adversarial nets on multiple datasets, such as MNIST[21] and the Toronto Face Database (TFD). They measure the probability of the test data using a Gaussian Parzen window then recording log-likelihood. According to their testing, the generated samples are competitive with other well-working generative models. Thus, the adversarial nets framework is effective and has much room for potential extension applications.
4. What is one advantage of the proposed approach, beyond strong performance/accuracy?
   1. One advantage of this approach is a wide variety of functions can be worked into the model, expanding its capabilities.
5. What is one disadvantage/weakness/limitation of the approach or experimental validation?
   1. One disadvantage of this approach is that D and G have to be synchronized well during training, which may require a lot of engineering effort.
6. Suggest one possible extension of this approach, i.e. one idea for future work.
   1. A possible extension of this approach is improving the efficiency of training through better coordinating G and D.

Paper 4: [Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks](https://openaccess.thecvf.com/content_ICCV_2017/papers/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.pdf)

1. Summarize what this paper aims to do (what gap in science it is trying to address), and what its main contribution is, compared to what prior methods have already accomplished.
   1. Image-to-image translation changes an input image into an output image based on training image pairs. The data for these training pairs, however, is often unavailable. This approach accommodates for this unavailable data by showing how to translate images and create a mapping even without this data.
2. Summarize the proposed approach in 3-5 sentences.
   1. The approach seeks to create a mapping G: X 🡪 Y from the source domain to the target domain and an inverse mapping F : Y 🡪 X. The mapping for G will be distributed identically to the set in Y to be an ideal and accurate mapping; additionally, F will be an inverse of G in order to ensure that the mapping is correct. The mappings are trained at the same time and their loss computed with cycle consistency and adversarial loss (using discriminators). The approach is applied to a variety of applications such as style transfer.
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?
   1. Both qualitative and quantitative results show this approach (CycleGAN) improves upon prior methods. First, the approach is compared against recent methods: in prior methods, participants were not fooled in direction of translation, but in CycleGAN, participants were fooled in a quarter of trials. Then, they compare the full method (including adversarial and cycle consistency loss) against other methods: removing loss degraded results by causing mode collapse. Lastly, they demonstrate the generalization of the method on applications without paired data: it works in object transfiguration, season transfer, collection style transfer, photo generation from paintings, and photo enhancement.
4. What is one advantage of the proposed approach, beyond strong performance/accuracy?
   1. A wide variety of applications (from object transfiguration to photo enhancement) can be accomplished without paired image training data.
5. What is one disadvantage/weakness/limitation of the approach or experimental validation?
   1. One disadvantage of this approach is that transformations involving geometric changes suffer.
6. Suggest one possible extension of this approach, i.e. one idea for future work.
   1. This approach could potentially be used as a security measure to hide identities in systems that monitor people on camera.