Abstract and Intro Only

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

* Using a large, unlabeled dataset
* Take a random pair of patches from the image and guess where those patches would be located relative to each other. CNNs
* Argue that doing well requires the model to learn to recognize objects and their parts.
* Talks about how the issue with having a lot of data (hundreds of billions of images) are hampered by the cost (time and money and resources) to annotate them
  + Wants to use unsupervised learning to save time
    - However, no research within the past few decades have shown that this method has not been able to extract useful information
* Workflow

1. Sample random pairs in one of the eight spatial configurations (i.e. center patch and a patch from either above, below, left, right, above to the right, above to the left, below to the left, below to the right)
2. Present to a machine learner with no information of the original position for the patch
3. Algorithm must guess the position of one patch relative to the other.

* Hypothesis: Doing well on this task requires understanding scenes and objects, i.e. a good visual representation for this task will need to extract objects and their parts in order to reason about their relative spatial location
* Demonstrates that the visual representation is good for both object detection. Means that it generalizes across images, despite being trained using an objective function that operates on a single image at a time.

[The Curious Robot: Learning Visual Representations via Physical Interactions](http://www.cs.cmu.edu/~ylpark/publications/Pinto_ECCV_2016.pdf)

* Argue that biological agents use physical interactions with the world to learn visual representation unlike current vision systems which just use passive observations
* Built a system on a Baxter platform (?) that allows pushing, poking, grasping, and observations
* This system uses these interactions to collect more than 130k datapoints, with each datapoint providing supervision to a shared ConvNet architecture allowing it to learn visual representations
* To evaluate the learned ConvNet, they compared and saw improvements after the system interacted with the object on image classification tasks. On the task of instance retrieval, the network outperforms ImageNet network on recal@1 by 3%
* Learned from nature and mimics biology
* Uses this physical interaction instead of the passive observation all other systems utilize

[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)

* Goal: Create a system that has learned the mapping between input and output images using a training set of aligned image pairs.
* Issue: paired training data is usually not available
* System can learn and capture the special characteristics of one image collection and figure out how those characteristics could be translated into another image collection
* Image-to-image translation
* Needed to translate between domains without paired input-output examples. However, some relationship between the given examples is needed
* Open source and called CycleGAN

Entire Paper

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

* Propose a new framework for estimating GANs where they train two models simultaneously. One for data distribution and one that estimates the probability that a sample came from the training data rather than the previous model
* The first model wants to increase the chances that the second model predicts wrong
* Adversarial network is a generative model pitted against a discriminative model
* Other work done utilized variational autoencoders (VAEs) which are similar to GANs but the discriminative model in a VAE is a recognition model that performs approximate interference. GANs require differentiation through visible units while VAEs require differentiation through hidden units.
* Both generative model and discriminative model follow a two-player minimax game where each ”compete” against each other in order to either generate data nearly indistinguishable from training data or where the discriminative model becomes too advanced for the generative model to be successful (thus making it a cycle of improvement)
* Theorem 1: the global minimum of the virtual training criterion C(G) is achieved if and only if pg=pdata. At that point, C(G) achieves the value -log4
* Trained the adversarial nets on a range of datasets.
  + Generator model utilized rectifier linear activations and sigmoid activations
  + Discriminator model utilized maxout activations and Dropout was applied.
  + Used noise as the input to only the bottommost layer of the generator network
* The experimented method of estimating the likelihood has somewhat high variance and does not perform well in high dimensionality but it is the best method available to their knowledge.
* Advances in generative models that can sample but not estimate likelihood are what motivate further research
* Pros and Cons
  + Cons
    - No explicit representation of Pg(x)
    - D must be synchronized well with G during training
      * In particular, G must not be trained too much without updating D
  + Pros
    - Markov chains are never needed
    - Only backprop is used to obtain gradients
    - No inference is needed during learning
    - Wide variety of functions can be incorporated into the model
  + Pros were primarily computation

[Image-to-Image Translation with Conditional Adversarial Networks](https://openaccess.thecvf.com/content_cvpr_2017/papers/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.pdf)

* Investigating conditional adversarial networks as a general-purpose solution to image-to-image translation problems.
* Due to the way these structures function, it is possible to apply the same generic approach to problems that would usually require drastically different loss functions.
* Define “image-to-image translation” as the problem of translating one possible representation of a scene into another, given sufficient training data.
* Setting is always the same but wants to develop a common framework for a multitude of problems
* Current CNNs can work well but require a lot of manual intervention and attention to detail as it is possible to mess up the output. Wants an easier, more hands-off solution. GANs do exactly what they want.
* Utilizes conditional GANs (cGANs) as they condition on an input image and generate a corresponding output image (user gives some input and the model gives some output)
* Current work with Image-to-image translation treat the output space as “unstructured” in the sense that each output pixel is considered conditionally independent from all others given the input image. cGANs instead learn a structured loss, which penalize the joint configuration of the output.
* Not the first people to utilize cGANs as others had used them to produce normal maps, future frame prediction, product photo generation, and image generation from sparse annotations. Some other papers have utilized GANs for image-to-image mappings but the GANs were applied unconditionally. The approach of the paper differs in that nothing is application specific making it simpler than the other setups.
  + It also differs as it utilizes a “U-Net” based architecture for the generator and a convolutional “PatchGAN” classifier for the discriminator.
* Both generator and discriminator use modules of the form convolution-BatchNorm-ReLu.
* Assumes that input and output differ in surface structure but both are renderings of the same underlying structure. Therefore, structure in the input is roughly aligned with structure in the output.
  + Previous solutions to similar problems used an encoder-decoder network. For this problem, they implemented skip connections between some layers to circumvent the bottleneck the previous solutions would impose.
* Designed “PatchGAN” to maximize sharpness of the image and only penalizes structure at the scale of patches. It tries to classify if each NxN patch in an image is real or fake.
* Experiments
  + Wants to test the generality of cGANs including both photo generation and semantic segmentation.
  + Results shown both in paper and on a GitHub. Shows their results compared to other methods. Shows a hand drawn demo (sketch->photo) which utilizes an edge detector.
  + <https://phillipi.github.io/pix2pix/>
* Evaluations
  + Run “real vs fake” perceptual studies on Amazon Mechanical Turk (AMT).
    - Testers were presented with a series of trials that pitted a “real” image against a “fake” image generated by our algorithm. Each image appeared for 1 second and the testers were given unlimited time to select “real” or “fake”.
  + Measure if synthesized cityscapes are realistic enough that off-the-shelf recognition system can recognize the objects in them.
    - Idea: If off-the-shelf classifiers, which are trained on real images, can detect objects within the generated images, the generated images are realistic.
* Analysis
  + Compared encoder-decoder method to their U-Net skip method and found that the encoder-decoder is unable to learn to generate realistic images in their experiments. The advantages of U-Net does not appear to be specific to cGANs.
* Results in this paper suggest that conditional adversarial networks are a promising approach for many image-to-image translation tasks.

**CS1674: Essay 2**

**Due:** 4/27/2021, 11:59pm  
  
This assignment is worth 26 points.  
  
There are 17 papers from recent conferences (in the format "LastName CONF Year") posted on the course website, for the last three topics (dates 4/6 to 4/22). Your task is to:

1. choose 2 of these to read carefully (the entire paper) to answer Part II,
2. choose 3 (different than the first 2) for which you read the Introduction section (feel free to skim the rest but no need to) to answer Part I

You will not be penalized for providing more than the suggested number of sentences.  
  
Part I (6 points):  
  
For each of the 3 papers whose introduction you choose to read, answer the following questions (2 points total per paper):

1. [1 pt] What is this paper trying to accomplish? (Think about what are the current limitations of prior approaches, and why these limitations are important.) (2-3 sentences)
2. [1 pt] What is the high-level idea of **how** the paper will accomplish its goal? (1-3 sentences)

The high-level answers from the Introduction section are sufficient.  
  
Part II (20 points):  
  
For each of the 2 papers you chose to read in detail, answer the following questions (10 points total per paper):

1. [2 pts] 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. (2-3 sentences)
2. [3 pts] Summarize the proposed approach. (3-5 sentences)
3. [2 pts] Summarize the experimental validation of the approach-- how is the proposed method tested, and what are the major observations and conclusions about its effectiveness? (2-3 sentences)
4. [1 pt] What is one advantage of the proposed approach, beyond strong performance/accuracy? (1-3 sentences)
5. [1 pt] What is one disadvantage/weakness/limitation of the approach or experimental validation? (1-3 sentences)
6. [1 pt] Suggest one possible extension of this approach, i.e. one idea for future work. (1-3 sentences)

You do not need to understand the entire mathematical details of the work; high-level descriptions (e.g., no need to provide equations) are sufficient.  
  
**Submission:**

* essay2.pdf/.docx

Part I

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

1. This paper is trying to have a CNN be able to determine relative positions of two patches of a larger image. The data used for training this CNN were unlabeled and the training was done in an unsupervised fashion. Wants to show that if the CNN does well, it has gained an understanding of scenes and objects and has a good visual representation.
2. They take a sample of random pairs that are in one of the eight spatial configurations. The present those to a machine learner with no prior information of the original position for the pairs. The algorithm will then make a prediction in of where it things the pairs are relative to each other.

[The Curious Robot: Learning Visual Representations via Physical Interactions](http://www.cs.cmu.edu/~ylpark/publications/Pinto_ECCV_2016.pdf)

1. The paper is trying to prove that using a more biological approach to Visual Representations (i.e. similar to how animals and humans approach it) is a more effective way than a passive observation approach. They want to see if having datapoints from not only visual observation but also grabbing, pushing, and poking with a haptic sensor will improve the system’s Visual Representation.
2. The team built a robot that observes an object from a static point. That robot has a moveable arm with different ends, a grabbing head, a pushing head, and a poking head. The system will then utilize datapoints gathered from all four of these interactions in order to better broadly classify objects.

[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. They want to create a system that has learned the mapping between input and output images using a set of aligned image pairs. They need to do this in such a way that there is no paired training data necessary as that is often non-existent. The system would be able to translate some input image into the form or style of another image, i.e. photo -> painting and vis-versa.
2. They utilized an algorithm that can learn to translate between domains without paired input-output examples. They trained the mapping such that the input and output would be deemed indistinguishable by an adversary trained to distinguish them apart in order to find close enough pairs.

Part II

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

1. The paper is trying to propose a new framework for estimating generative networks. The way it hopes to do this is through the use of adversarial networks. Other work on deep generative networks had issues as they would require numerous approximations to the likelihood gradient. Other work has been done with variational autoencoders (VAEs) which are similar to GANs but the discriminative model in a VAE is a recognition model that performs approximate interference. GANs require differentiation through visible units while VAEs require differentiation through hidden units.
2. This proposed framework of utilizing adversarial networks would work as follows: two models would be trained simultaneously. The first model would be a generative model which is trained to generate things similar in form to the training data. The second model would be a discriminator model with the goal of predicting if a given sample is part of the training data utilized for the generative model or if it was generated by the first model. Both models would follow a two-player minimax game where each “compete” against each other in order to either generate data nearly indistinguishable from training data (at least to the discriminator model) or where the discriminative model becomes too advances for the generative model to be successful, thus making it a cycle of improvement for future iterations of training.
3. The adversarial nets were trained on a range of datasets including MNIST, the Toronto Face Database (TFD), and CIFAR-10. The generative model utilized rectifier linear activations and sigmoid activations. The discriminator model utilized maxout activations and dropout was applied. The use of noise was the input only to the bottommost layer of the generator network. The results of the experimented method of estimating the likelihood has somewhat high variance and does not perform well in high dimensionality, but it is the best method available at the time of the paper being written (2014).
4. All of the listed advantages were primarily computational based. Some examples being Markov chains are never needed, only backprop is used to obtain gradients, no inference is needed during learning, and a wide variety of functions can be incorporated into the model.
5. A drawback to the approach is that the two models, generative and discriminative, must be synchronized well during the training as if the generative model is trained too much, “the Helvetia scenario” may occur where it collapses too many values.
6. I do wonder if utilizing increasing the “game” from a two-player minimax game to a, say 4 player, 2 on each team, would provide benefits to this type of model. I could see the use of two separate models working to make data to fool the discriminators and using the result to both become better.

[Image-to-Image Translation with Conditional Adversarial Networks](https://openaccess.thecvf.com/content_cvpr_2017/papers/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.pdf)

1. The paper is trying to investigate and apply conditional adversarial networks as a general-purpose solution for image-to-image translation problems. Prior approaches have used CNNs which work well but require a lot of manual intervention and attention to detail in that intervention as it is possible to mess up the network after the intervention. cGANs (conditional Generative Adversarial Networks) provide a more hands-off solution to the issue CNNs provide. Current work with Image-to-image translation treat the output space as “unstructured” in the sense that each output pixel is considered conditionally independent from all others given the input image. cGANs instead learn a structured loss, which penalize the joint configuration of the output. Not the first people to utilize cGANs as others had used them to produce normal maps, future frame prediction, product photo generation, and image generation from sparse annotations. Some other papers have utilized GANs for image-to-image mappings but the GANs were applied unconditionally. The approach of the paper differs in that nothing is application specific making it simpler than the other setups.
2. As stated, the paper utilizes cGANs for their general use applications. Within the cGANs, the generative network utilizes a “U-Net” architecture with a skip element to it and the discriminator utilizes a developed “PatchGAN” architecture which checks NxN patches within the image to determine if they are real or fake. Both generator and discriminator will use modules of the form convolution-BatchNorm-ReLu.
3. They want to test the generality of cGANs with photo generation and semantic segmentation. Their results were shown in a paper and on a GitHub link which shows their results compared to other methods. [Link](https://phillipi.github.io/pix2pix/). They had two method of evaluating the results. First they would run “real vs fake” perceptual studies using the Amazon Mechanical Turk. Testers were presented with a series of trials that pitted a “real” image against a “fake” image and then were asked to determine which one was real and which one was fake. They also measured if their results were realistic enough that off-the-shelf recognition systems could recognize the generated objects within the output. The idea being, that if off-the-shelf classifiers, which are trained on real images, can detect objects within the generated images, the generated images are realistic. Compared to other methods, their results show that their method was significantly better. Compared to encoder-decoder method against their U-Net skip method, they found that encoder-decoder was unable to learn to generate realistic images in their experiments whereas the U-Net skip method was able to be generalized and achieves the superior results. Results in this paper suggest that conditional adversarial networks are a promising approach for many image-to-image translation tasks
4. As stated in the previous answer, the U-Net utilized for generation appear not to be specific to cGANs, or in other words, is able to be generalized whereas other methods were not.
5. They state that while cGANs achieve some success, that they are far from the best available method for solving some problems they tacked, specifically photo->labels. A simpler method of using a L1 regression gets better scores making it more sufficient.
6. If I had to expand some part of this paper, I would see if it could work for text-in-image (i.e. photo of text).