Conditional-GAN Research Paper summary

Mohammed Ikram C C
CE24B086

INTRODUCTION:-

- Different users may describe the same image in different ways using synonyms or personal preferences
  e.g., one person might tag a beach photo as “ocean,” while another calls it “vacation”
- To tackle this, researchers explore conceptual word embeddings, which group similar words together based on meaning.
- that uses a Conditional Adversarial Network (cGAN) to predict a set of relevant words based on an image.
- This research introduces a machine learning model that is trained using:
  A convolutional neural network (CNN) to extract image features.
  A skip-gram word embedding model to understand text-based relationships.

Related Work

Multi-modal Learning For Image Labelling

Traditional image labeling models mostly rely on visual features alone, but real-world image descriptions often involve both images and text.

Previous research has explored using deep learning models to process images and text together. Some approaches involve training neural networks on large datasets where images are linked to descriptions, tags, or captions.

By doing so, these models learn the relationships between visual content and language. It allow them to predict better, more human-like labels for images.

Conditional Adversarial Nets

3.1 Generative Adversarial Nets

A GAN consists of two competing neural networks:
Generator – This network creates synthetic data (e.g., fake images or labels) that resemble real data.
Discriminator – This network evaluates whether the generated data is real or fake.

The two networks play a game where the generator continuously improves its ability to produce realistic data, while the discriminator gets better at distinguishing real from fake. Over time, the generator produces outputs that are nearly indistinguishable from real data.

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

3.2 Conditional Adversarial Nets

The standard GANs generate data randomly. cGANs allow control over the output by providing extra inputs, such as class labels, image features, or other contextual data.

By using cGANs, the model can learn relationships between images and descriptive text more effectively. It allows to produce realistic and relevant multi-label tags for images

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x}|\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}|\boldsymbol{y})))].$$

The process going on(the below text is exactly copy pasted from chatgpt)

Step 1: Generator (G) Creates Fake Data
The Generator (G) takes in two inputs:

- A random noise vector (z) → Adds randomness to generate different variations of data.
- A conditional input (y) → Provides guidance for controlled data generation (e.g., image features for image labeling).

Using this input, the Generator produces a synthetic data sample (e.g., a predicted image label).

Step 2: Discriminator (D) Evaluates the Data
The Discriminator (D) receives two types of inputs:

- Real data (x) paired with its actual condition (y) → Genuine data samples from the dataset.
- Fake data (G(z, y)) generated by the generator, paired with the condition y → Synthetic samples created by the generator.

The Discriminator's goal is to classify whether the input data is real or fake based on how well it matches the given condition (y).

Step 3:
Training the Networks Generator's Goal → Improve at creating realistic data that fools the Discriminator.
Discriminator's Goal → Get better at distinguishing between real and fake data.
This adversarial process continues until the Generator produces highly realistic samples that the Discriminator can no longer easily distinguish from real data.

Experimental Results

4.1 Unimodal

In this section, the researchers test their Conditional Adversarial Network (cGAN) on a single type of data (unimodal) to evaluate its performance before applying it to more complex

The experiment is conducted on the MNIST dataset, which contains handwritten digit images (0–9) along with their corresponding labels.
The goal is to train the cGAN to generate digit images conditioned on a given digit label (y).

The model successfully generates realistic digit images corresponding to the given label (y).

4.2 Multimodal

Multimodal learning combines multiple data sources (images + text) to generate better, context-aware predictions.

The dataset used for this experiment is MIR Flickr 25,000, which contains images with user-generated metadata (UGM), such as tags, titles, and descriptions.

The process:-
Image Features are extracted using a convolutional neural network (CNN) pre-trained on ImageNet.
Word Representations: Trained using a skip-gram model on metadata from the YFCC100M dataset
Generator (G): Maps image features + noise to a 200-dimensional word vector representing possible tags.
Discriminator (D): Evaluates whether the generated words match the image content.
Used stochastic gradient descent (SGD) with momentum and dropout.

Future Work

Now, CNN (image feature extractor) and the skip-gram model (word embeddings) were kept fixed during adversarial training. Future work will involve backpropagating through these models.

Further experiments with different architectures for the generator and discriminator will be conducted.

Model will perform on larger and more complex datasets is planned.