Similarity Learning and Data Generators

Reed Graff[1]

[1] Undergraduate Student at The University of Texas at Austin
Rangergraff@gmail.com

Abstract

Applying similarity learning methods in classification problems allows for extremely accurate classification with the downside of longer classification times. This paper explores many different methods in which similarity learning algorithms can be applied to classic classification problems, primarily regarding the Siamese NN architecture.

Background

As a point of reference, this paper will be focusing on siamese neural networks and its respective data generators, however, this data paper may still hold value to other fields of research, void of any connection to siamese neural networks. This paper has utilized some common terms, idiomatic expressions, and lingo specific to both data generation and machine learning, which will be addressed now:

Table 1: *Common Terms and Definitions*

Short	Long
AI/ML	Artificial Intelligence/Machine Learning
SiNN	Siamese Neural Network
Scrape	To aggregate information

What Is AI/ML?

AI was originally coined by John McCarthy in 1955, before being further defined as "the construction of computer programs that engage in tasks that are currently more satisfactorily performed by human beings because they require high-level mental processes such as: perceptual learning, memory organization and critical reasoning"[1, 2].

Now AI is a broad term used to describe just about any learning process being completed by a computer, however, in simplest terms it is a computer task that creates a function which is trained to predict outputs based on inputs.

What Is Deep Learning?

Deep Learning is a more specific branch of AI that is typically associated with multi-layer neural networks (more than 3) also with the purpose of training values to better predict outputs dependent upon inputs.

What Is Similarity Learning?

Similarity is a very specific branch of ML that is used for determining the similarity between different data sets[3].

What Are Siamese Neural Networks?

A SiNN is a kind of similarity learning approach to comparing images (or other 2D data), and determining their similarity. SiNNs leverage one neural network which is used to classify each individual image, which can then be used to find the euclidean distance and the overall similarity of the images.

Introduction

Requirements of The Data Generator

Many libraries exist now for generating, changing, and augmenting data, however, there has been some amount of underdevelopment in the area of SiNNs, which can in large part be attributed to the lack of data generation for such architecture.

The purpose of this paper is to expose possible methods of data generation for SiNNs, both as a stand alone generator (one that isn't dependent on other AI/ML libraries), as well as one that may interface with the Tensorflow library.

Table 2: *Tensorflow arguments for image_dataset_from_directory()*

Short	Long
directory	Directory where the data is located.If labels is "inferred", it should containsubdirectories, each containing images for a class.Otherwise, the directory structure is ignored.
labels	Either "inferred"(labels are generated from the directory structure),None (no labels),or a list/tuple of integer labels of the same size as the number ofimage files found in the directory. Labels should be sorted accordingto the alphanumeric order of the image file paths(obtained via os . walk(directory) in Python).
label _ mode	String describing the encoding of labels . Options are: ’int’: means that the labels are encoded as integers(e.g. for sparse_categorical_crossentropy loss). ’categorical’ means that the labels areencoded as a categorical vector(e.g. for categorical_crossentropy loss). ’binary’ means that the labels (there can be only 2)are encoded as float32 scalars with values 0 or 1(e.g. for binary_crossentropy). None (no labels).
class _ names	Only valid if "labels" is "inferred". This is the explicitlist of class names (must match names of subdirectories). Usedto control the order of the classes(otherwise alphanumerical order is used).
color _ mode	One of "grayscale", "rgb", "rgba". Default: "rgb".Whether the images will be converted tohave 1, 3, or 4 channels.
batch _ size	Size of the batches of data. Default: 32.If None , the data will not be batched(the dataset will yield individual samples).
image _ size	Size to resize images to after they are read from disk,specified as (height , width) . Defaults to (256 , 256) .Since the pipeline processes batches of images that must all havethe same size, this must be provided.
shuffle	Whether to shuffle the data. Default: True.If set to False, sorts the data in alphanumeric order.
seed	Optional random seed for shuffling and transformations.
validation _ split	Optional float between 0 and 1,fraction of data to reserve for validation.
subset	Subset of the data to return.One of "training", "validation" or "both".Only used if validation _ split is set.When subset="both" , the utility returns a tuple of two datasets(the training and validation datasets respectively).
interpolation	String, the interpolation method used when resizing images.Defaults to bilinear . Supports bilinear , nearest , bicubic , area , lanczos3 , lanczos5 , gaussian , mitchellcubic .
follow _ links	Whether to visit subdirectories pointed to by symlinks.Defaults to False.
crop _ to _ aspect _ ratio	If True, resize the images without aspectratio distortion. When the original aspect ratio differs from the targetaspect ratio, the output image will be cropped so as to return thelargest possible window in the image (of size image _ size) that matchesthe target aspect ratio. By default (crop _ to _ aspect _ ratio=False),aspect ratio may not be preserved.
*kwargs	Legacy keyword arguments.

Regarding the development of the generator, this paper seeks to mimic the pre-existing tensorflow function “tf.keras.utils.image_dataset_from_directory”, and match the arguments (Table 2) that are supported by the aforementioned function[4] in the Tensorflow integration explained later on.
However, prior to this there will be a stand alone generator developed for the same purpose. The focus, however, of the standalone is to provide a more fundamental understanding of the generator and will use the following limited list arguments:

- directory
- labels

Species Identification

Proin lobortis efficitur dictum. Pellentesque vitae pharetra eros, quis dignissim magna. Sed tellus leo, semper non vestibulum vel, tincidunt eu mi. Aenean pretium ut velit sed facilisis. Ut placerat urna facilisis dolor suscipit vehicula. Ut ut auctor nunc. Nulla non massa eros. Proin rhoncus arcu odio, eu lobortis metus sollicitudin eu. Duis maximus ex dui, id bibendum diam dignissim id. Aliquam quis lorem lorem. Phasellus sagittis aliquet dolor, vulputate cursus dolor convallis vel. Suspendisse eu tellus feugiat, bibendum lectus quis, fermentum nunc. Nunc euismod condimentum magna nec bibendum. Curabitur elementum nibh eu sem cursus, eu aliquam leo rutrum. Sed bibendum augue sit amet pharetra ullamcorper. Aenean congue sit amet tortor vitae feugiat.

Mauris interdum porttitor fringilla. Proin tincidunt sodales leo at ornare. Donec tempus magna non mauris gravida luctus. Cras vitae arcu vitae mauris eleifend scelerisque. Nam sem sapien, vulputate nec felis eu, blandit convallis risus. Pellentesque sollicitudin venenatis tincidunt. In et ipsum libero. Nullam tempor ligula a massa convallis pellentesque.

Data Analysis

Vestibulum sodales orci a nisi interdum tristique. In dictum vehicula dui, eget bibendum purus elementum eu. Pellentesque lobortis mattis mauris, non feugiat dolor vulputate a. Cras porttitor dapibus lacus at pulvinar. Praesent eu nunc et libero porttitor malesuada tempus quis massa. Aenean cursus ipsum a velit ultricies sagittis. Sed non leo ullamcorper, suscipit massa ut, pulvinar erat. Aliquam erat volutpat. Nulla non lacus vitae mi placerat tincidunt et ac diam. Aliquam tincidunt augue sem, ut vestibulum est volutpat eget. Suspendisse potenti. Integer condimentum, risus nec maximus elementum, lacus purus porta arcu, at ultrices diam nisl eget urna. Curabitur sollicitudin diam quis sollicitudin varius. Ut porta erat ornare laoreet euismod. In tincidunt purus dui, nec egestas dui convallis non. In vestibulum ipsum in dictum scelerisque.

Mauris interdum porttitor fringilla. Proin tincidunt sodales leo at ornare. Donec tempus magna non mauris gravida luctus. Cras vitae arcu vitae mauris eleifend scelerisque. Nam sem sapien, vulputate nec felis eu, blandit convallis risus. Pellentesque sollicitudin venenatis tincidunt. In et ipsum libero. Nullam tempor ligula a massa convallis pellentesque. Mauris interdum porttitor fringilla. Proin tincidunt sodales leo at ornare. Donec tempus magna non mauris gravida luctus. Cras vitae arcu vitae mauris eleifend scelerisque. Nam sem sapien, vulputate nec felis eu, blandit convallis risus. Pellentesque sollicitudin venenatis tincidunt. In et ipsum libero. Nullam tempor ligula a massa convallis pellentesque.

Table 3: *Example single column table.*

Location		
East Distance	West Distance	Count
100km	200km	422
350km	1000km	1833
600km	1200km	890

Figure 1: *Anther of thale cress (Arabidopsis thaliana), fluorescence micrograph. Source: Heiti Paves, https://commons.wikimedia.org/wiki/File:Tolmukapea.jpg.*

Results

Referencing a table using its label: Table 3.

Aenean feugiat pellentesque venenatis. Sed faucibus tristique tortor vel ultrices. Donec consequat tellus sapien. Nam bibendum urna mauris, eget sagittis justo gravida vel. Mauris nisi lacus, malesuada sit amet neque ut, venenatis tempor orci. Curabitur feugiat sagittis molestie. Duis euismod arcu vitae quam scelerisque facilisis. Praesent volutpat eleifend tortor, in malesuada dui egestas id. Donec finibus ac risus sed pellentesque. Donec malesuada non magna nec feugiat. Mauris eget nibh nec orci congue porttitor vitae eu erat. Sed commodo ipsum ipsum, in elementum neque gravida euismod. Cras mi lacus, pulvinar ut sapien ut, rutrum sagittis dui. Donec non est a metus varius finibus. Pellentesque rutrum pellentesque ligula, vitae accumsan nulla hendrerit ut.

Table 4: *Example two column table with fixed-width columns.*

Location		
East Distance	West Distance	Count
100km	200km	422
350km	1000km	1833
600km	1200km	890

Referencing a figure using its label: Figure 1.

Aenean porttitor eros non pharetra congue. Proin in odio in dolor luctus auctor ac et mi. Etiam euismod mi sed lectus fringilla pretium. Phasellus tristique maximus lectus et sodales. Mauris feugiat ligula quis semper luctus. Nam sit amet felis sed leo fermentum aliquet. Mauris arcu dui, posuere id sem eget, cursus pulvinar mi. Donec nec lacus non lectus fermentum scelerisque et at nibh. Sed tristique, metus ac vestibulum porta, tortor lectus placerat lorem, et convallis tellus dolor eget ante. Pellentesque dui ligula, hendrerit a purus et, volutpat tempor lectus. Mauris nec purus nec mauris rhoncus pellentesque. Quisque quis diam sed est lacinia congue. Donec magna est, hendrerit sed metus vel, accumsan rutrum nibh.

Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Etiam cursus lectus purus, tempus iaculis quam dictum tristique. Nam interdum sapien nec tempor mattis. Quisque id sapien nisi. Mauris vehicula ornare eros vel efficitur. Nulla consectetur, turpis quis fringilla tincidunt, mi neque iaculis lectus, vel commodo elit odio non ex. Duis facilisis, purus ac viverra iaculis, turpis lectus ultrices ante, ac vestibulum ligula magna in libero. Etiam tristique maximus lacinia. Vestibulum hendrerit, lacus malesuada laoreet blandit, sapien velit sollicitudin nunc, eu porttitor urna ligula at lorem. Aliquam faucibus eros in fermentum venenatis. Fusce consectetur congue pellentesque. Suspendisse at nisi sit amet est porttitor cursus. Cras placerat faucibus nunc, a laoreet justo dignissim sit amet.

International Support

àáâäãåèéêëìíîïòóôöõøùúûüÿýñçčšž
ÀÁÂÄÃÅÈÉÊËÌÍÎÏÒÓÔÖÕØÙÚÛÜŸÝÑ
ßÇŒÆČŠŽ

Links

This is a clickable URL link: LaTeX Templates. This is a clickable email link: vel@latextemplates.com. This is a clickable monospaced URL link: `https://www.LaTeXTemplates.com`.

Discussion

This statement requires citation [**Smith:2023qr**]. This statement requires multiple citations [**Smith:2023qr**, **Smith:2024jd**]. This statement contains an in-text citation, for directly referring to a citation like so: **Smith:2024jd**.

Subsection One

Suspendisse potenti. Vivamus suscipit dapibus metus. Proin auctor iaculis ex, id fermentum lectus dapibus tristique. Nullam maximus eros eget leo pretium dapibus. Nunc in auctor erat, id interdum risus. Suspendisse aliquet vehicula accumsan. In vestibulum efficitur dictum. Sed ultrices, libero nec fringilla feugiat, elit massa auctor ligula, vehicula tempor ligula felis in lectus. Suspendisse sem dui, pharetra ut sodales eu, suscipit sit amet felis. Donec pretium viverra ante, ac pulvinar eros. Suspendisse gravida consectetur urna. Pellentesque vitae leo porta, imperdiet eros eget, posuere sem. Praesent eget leo efficitur odio bibendum condimentum sit amet vel ex. Nunc maximus quam orci, quis pulvinar nibh eleifend ac. Quisque consequat lacus magna, eu posuere tellus iaculis ac. Sed vitae tortor tincidunt ante sagittis iaculis.

Subsection Two

Nullam mollis tellus lorem, sed congue ipsum euismod a. Donec pulvinar neque sed ligula ornare sodales. Nulla sagittis vel lectus nec laoreet. Nulla volutpat malesuada turpis at ultricies. Ut luctus velit odio, sagittis volutpat erat aliquet vel. Donec ac neque eget neque volutpat mollis. Vestibulum viverra ligula et sapien bibendum, vel vulputate ex euismod. Curabitur nec velit velit. Aliquam vulputate lorem elit, id tempus nisl finibus sit amet. Curabitur ex turpis, consequat at lectus id, imperdiet molestie augue. Curabitur eu eros

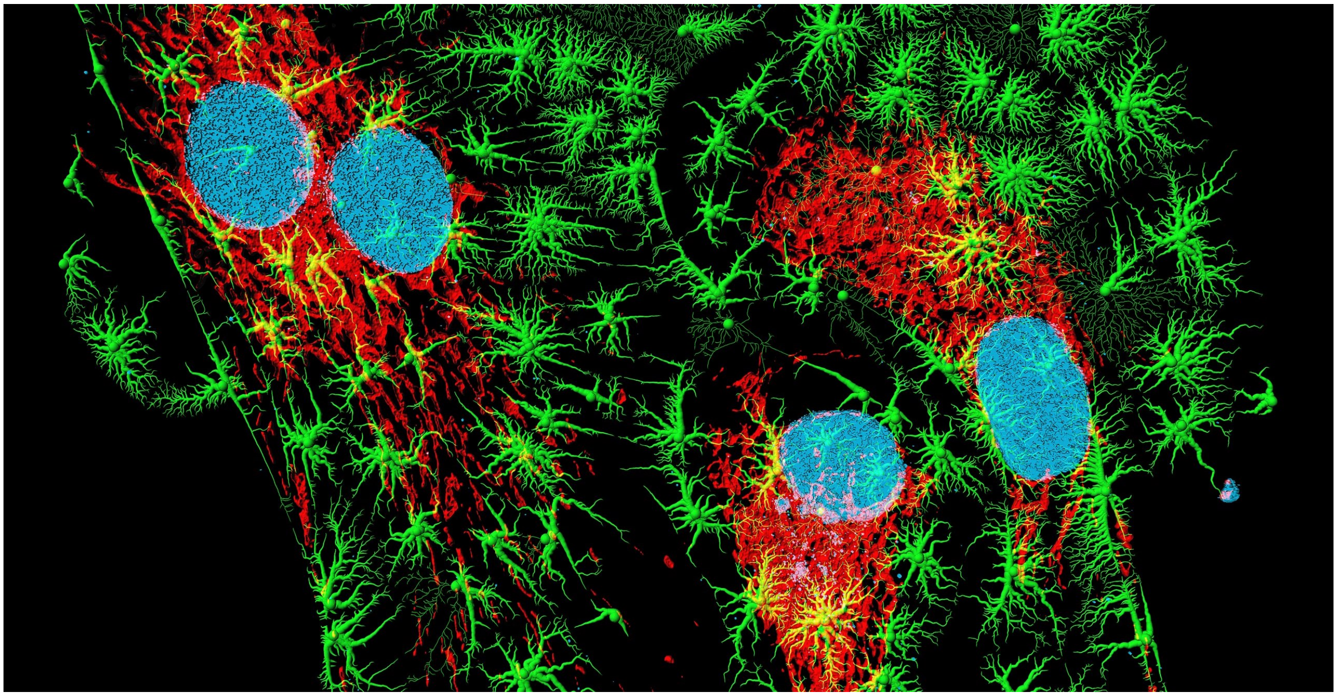

Figure 2: *Bovine pulmonary artery endothelial cells in culture. Blue: nuclei; red: mitochondria; green: microfilaments. Computer generated image from a 3D model based on a confocal laser scanning microscopy using fluorescent marker dyes. Source: Heiti Paves, https://commons.wikimedia.org/wiki/File:Fibroblastid.jpg.*

molestie purus commodo hendrerit. Quisque auctor ipsum nec mauris malesuada, non fringilla nibh viverra. Quisque gravida, metus quis semper pulvinar, dolor nisl suscipit leo, vestibulum volutpat ante justo ultrices diam. Sed id facilisis turpis, et aliquet eros.

Subsubsection Example Duis venenatis eget lectus a aliquet. Integer vulputate ante suscipit felis feugiat rutrum. Aliquam eget dolor eu augue elementum ornare. Nulla fringilla interdum volutpat. Sed tincidunt, neque quis imperdiet hendrerit, turpis sapien ornare justo, ac blandit felis sem quis diam. Proin luctus urna sit amet felis tincidunt, sed congue nunc pellentesque. Ut faucibus a magna faucibus finibus. Etiam id mi euismod, auctor nisi eget, pretium metus. Proin tincidunt interdum mi non interdum. Donec semper luctus dolor at elementum. Aenean eu congue tortor, sed hendrerit magna. Quisque a dolor ante. Mauris semper id urna id gravida. Vestibulum mi tortor, finibus eu felis in, vehicula aliquam mi.

Aliquam arcu turpis, ultrices sed luctus ac, vehicula id metus. Morbi eu feugiat velit, et tempus augue. Proin ac mattis tortor. Donec tincidunt, ante rhoncus luctus semper, arcu lorem lobortis justo, nec convallis ante quam quis lectus. Aenean tincidunt sodales massa, et hendrerit tellus mattis ac. Sed non pretium nibh.

Donec cursus maximus luctus. Vivamus lobortis eros et massa porta porttitor. Nam vitae suscipit mi. Pellentesque ex tellus, iaculis vel libero at, cursus pretium sapien. Curabitur accumsan velit sit amet nulla lobortis, ut pretium ex aliquam. Proin eget volutpat orci. Morbi eu aliquet turpis. Vivamus molestie urna quis tempor tristique. Proin hendrerit sem nec tempor sollicitudin.

References

[1] Gil Press. "A Very Short History Of Artificial Intelligence". In: *Forbes* (Oct. 2022). URL: `https://www.forbes.com/sites/gilpress/2016/12/30/a-very-short-history-of-artificial-intelligence-ai/?sh=a3401fc6fba2`.

[2] 2014. URL: `https://www.coe.int/en/web/artificial-intelligence/history-of-ai#:~:text=The%20term%20%22AI%22%20could%20be,because%20they%20require%20high%2Dlevel`.

[3] 2023. URL: `https://www.aiforanyone.org/glossary/similarity-learning`.

[4] 2023. URL: `https://www.tensorflow.org/api_docs/python/tf/keras/utils/image_dataset_from_directory`.