## Package 'causalimages'

## January 7, 2024

**Title** causalimages: An R Package for Causal Inference with Earth Observation, Biomedical, and Social Science Images

#### Version 0.0.1

**Description** A system for performing causal inference with earth observation, bio-medical, and social science images and image sequences (i.e., videos). The causal AI models use a JAX backend; the package provides functions for establishing the backend in a conda environment with GPU acceleration when possible.

Depends R (>= 3.3.3)

License GPL-3

Encoding UTF-8

LazyData false

Maintainer 'Connor Jerzak' <connor.jerzak@gmail.com>

Imports tensorflow, latex2exp, keras, reticulate, viridis, geosphere, raster, animation

RoxygenNote 7.2.3

## R topics documented:

Index		16
	WriteTfRecord	14
	SimulateImageSystem	
	LongLat2CRS	13
	image2	12
	GetImageRepresentations	10
	GetElementFromTfRecordAtIndices	9
	GetAndSaveGeolocatedImages	8
	BuildBackend	8
	AnalyzeImageMediation	6
	AnalyzeImageHeterogeneity	3
	AnalyzeImageConfounding	2

AnalyzeImageConfounding

Perform causal estimation under image confounding

## Description

Perform causal estimation under image confounding

## Usage

AnalyzeImageConfounding(obsW, obsY, imageKeysOfUnits, file, ...)

#### **Arguments**

٤	guments				
	obsW	A numeric vector where $\theta$ 's correspond to control units and 1's to treated units.			
	obsY	A numeric vector containing observed outcomes.			
	X	$(default = NULL) \ An \ optional \ numeric \ matrix \ containing \ tabular \ information \ used \ if \ orthogonalize = T. \ X \ is \ normalized \ internally \ and \ salience \ maps \ with \ respect to X \ are \ transformed \ back \ to \ the \ original \ scale.$			
	file	(default = NULL) Path to a tfrecord file generated by WriteTfRecord.			
	imageKeysOfUnit	S			
		$(default=1:length(obsY))\ A\ vector\ of\ length\ length(obsY)\ specifying\ the\ unique\ image\ ID\ associated\ with\ each\ unit.\ Samples\ of\ imageKeysOfUnits\ are\ fed\ into\ the\ package\ to\ call\ images\ into\ memory.$			
	nBoot	(default = 100L) Number of bootstrap iterations for uncertainty estimation.			
transportabilityMat					
		(optional) A matrix with a column named $imageKeysOfUnits$ specifying keys to be used by the package for generating treatment effect predictions for out-of-sample points.			
	long, lat	(optional) Vectors specifying longitude and latitude coordinates for units. Used only for describing highest and lowest probability neighbrhood units if specified.			
	conda_env	$(default = "CausalImagesEnv") \ A \ conda \ environment \ where \ computational \ environment \ lives, \ usually \ created \ via \ causalimages::BuildBackend()$			
	conda_env_required				
		(default = T) A Boolean stating whether use of the specified conda environment is required.			
	figuresTag	(default = "") A string specifying an identifier that is appended to all figure names.			
	figuresPath	(default = "./") A string specifying file path for saved figures made in the anal-			

age representation) should be plotted in the visual results. If a vector, plotBands should have 3 (and only 3) dimensions (corresponding to the 3 dimensions to be

ysis.

plotBands

plotResults

used in RBG plotting).
(default = T) Should analysis results be plotted?

strides (default = 2L) Integer specifying the strides used in the convolutional layers.

(default = 1L) An integer or vector specifying which band position (from the im-

dropoutRate (default = 0.1) Droppout rate used in training used to prevent overfitting (dropoutRate

= 0 corresponds to no dropout).

batchSize (default = 50L) Batch size used in SGD optimization. kernelSize (default = 5L) Dimensions used in convolution kernels.

nSGD (default = 400L) Number of stochastic gradient descent (SGD) iterations.

testFrac (default = 0.1) Fraction of observations held out as a test set to evaluate out-of-

sample loss values.

channelNormalize

(default = T) Should channelwise image feature normalization be attempted?

Default is T, as this improves training.

TfRecords\_BufferScaler

(default = 4L) The buffer size used in tfrecords mode is batchSize\*TfRecords\_BufferScaler. Lower TfRecords\_BufferScaler towards 1 if out-of-memory problems.

#### Value

Returns a list consisting of

- ATE\_est ATE estimate.
- ATE\_se Standard error estimate for the ATE.
- plotResults If set to TRUE, causal salience plots are saved to disk, characterizing the image confounding structure. See references for details.

#### References

• Connor T. Jerzak, Fredrik Johansson, Adel Daoud. Integrating Earth Observation Data into Causal Inference: Challenges and Opportunities. *ArXiv Preprint*, 2023.

## **Examples**

```
# For a tutorial, see
# github.com/cjerzak/causalimages-software/
```

 ${\tt AnalyzeImageHeterogeneity}$ 

Decompose treatment effect heterogeneity by image

#### **Description**

Implements the image heterogeneity decomposition analysis of Jerzak, Johansson, and Daoud (2023). Users input in treatment and outcome data, along with a function specifying how to load in images using keys referenced to each unit (since loading in all image data will usually not be possible due to memory limitations). This function by default performs estimation, constructs salience maps, and can optionally perform estimation for new areas outside the original study sites in a transportability analysis.

#### Usage

```
AnalyzeImageHeterogeneity(obsW, obsY, imageKeysOfUnits, file, kClust_est, ...)
```

#### **Arguments**

obsW A numeric vector where 0's correspond to control units and 1's to treated units.

obsY A numeric vector containing observed outcomes.

X (optional) A numeric matrix containing tabular information used if orthogonalize

= T.

orthogonalize (default = F) A Boolean specifying whether to perform the image decomposition

after orthogonalizing with respect to tabular covariates specified in X.

imageKeysOfUnits

(default = 1:length(obsY)) A vector of length length(obsY) specifying the unique image ID associated with each unit. Samples of imageKeysOfUnits are

fed into the package to call images into memory.

kClust\_est (default = 2L) Integer specifying the number of clusters used in estimation.

file (default = NULL) Path to a tfrecord file generated by WriteTfRecord.

transportabilityMat

(optional) A matrix with a column named key specifying keys to be used for generating treatment effect predictions for out-of-sample points in earth obser-

vation data settings.

long, lat (optional) Vectors specifying longitude and latitude coordinates for units. Used

only for describing highest and lowest probability neighbrhood units if specified.

conda\_env (default = "CausalImagesEnv") A conda environment where computational en-

vironment lives, usually created via causalimages::BuildBackend()

conda\_env\_required

(default = T) A Boolean stating whether use of the specified conda environment

is required.

figuresTag (default = "") A string specifying an identifier that is appended to all figure

names.

figuresPath (default = "./") A string specifying file path for saved figures made in the anal-

ysis.

plotBands (default = 1L) An integer or vector specifying which band position (from the

acquired image representation) should be plotted in the visual results. If a vector, plotBands should have 3 (and only 3) dimensions (corresponding to the 3

dimensions to be used in RGB plotting).

plotResults (default = T) Should analysis results be plotted?

optimizeImageRep

(default = TRUE)

nDepth\_Dense (default = 1L) Hidden depth of dense layers. Default of 0L means a single pro-

jection layer is performed after the convolutional layer (i.e., no hidden layers are

used).

strides (default = 2L) Integer specifying the strides used in the convolutional layers.=

yDensity (default = normal) Specifies the density for the outcome. Current options in-

clude normal and lognormal.

nMonte\_variational

(default = 5L) An integer specifying how many Monte Carlo iterations to use in the calculation of the expected likelihood in each training step.

nMonte\_predictive

(default = 20L) An integer specifying how many Monte Carlo iterations to use in the calculation of posterior means (e.g., mean cluster probabilities).

nMonte\_salience

(default = 100L) An integer specifying how many Monte Carlo iterations to use in the calculation of the salience maps (e.g., image gradients of expected cluster probabilities).

batchSize (default = 25L) Batch size used in SGD optimization.

kernelSize (default = 5L) Dimensions used in convolution kernels.

nSGD (default = 400L) Number of stochastic gradient descent (SGD) iterations.

channelNormalize

(default = T) Should channelwise image feature normalization be attempted? Default is T, as this improves training.

nWidth\_ImageRep

(default = 96L) Integer specifying the image/image sequence embedding dimension.

TfRecords\_BufferScaler

(default = 4L) The buffer size used in tfrecords mode is batchSize\*TfRecords\_BufferScaler. Lower TfRecords\_BufferScaler towards 1 if out-of-memory problems.

quiet (default = F) Should we suppress information about progress?

dilationRate (default = NULL) Dilation rate.

#### Value

Returns a list consiting of

- clusterTaus\_mean default
- clusterProbs\_mean. Estimated mean image effect cluster probabilities.
- clusterTaus\_sigma. Estimated cluster standard deviations.
- clusterProbs\_lowerConf. Estimated lower confidence for effect cluster probabilities.
- impliedATE. Implied ATE.
- individualTau\_est. Estimated individual-level image-based treatment effects.
- transportabilityMat. Transportability matrix withestimated cluster information.
- plottedCoordinates. List containing coordinates plotted in salience maps.
- whichNA\_dropped. A vector containing observations dropped due to missingness.

#### References

Connor T. Jerzak, Fredrik Johansson, Adel Daoud. Image-based Treatment Effect Heterogeneity. Forthcoming in *Proceedings of the Second Conference on Causal Learning and Reasoning (CLeaR)*, *Proceedings of Machine Learning Research (PMLR)*, 2023.

#### **Examples**

```
# For a tutorial, see
```

# github.com/cjerzak/causalimages-software/

 $Analyze Image Mediation \quad \textit{Analyze Image Mediation}$ 

### Description

Under construction.

#### Usage

AnalyzeImageMediation()

#### **Arguments**

obsW A numeric vector where 0's correspond to control units and 1's to treated units.

obsY A numeric vector containing observed outcomes.

X (optional) A numeric matrix containing tabular information used if orthogonalize

= T.

doConvLowerDimProj

(default = T) Should we project the nFilters convolutional feature dimensions down to nDimLowerDimConv to reduce the number of required parameters.

nDimLowerDimConv

 $(default=3L)\ If\ doConvLowerDimProj=T,\ then,\ in\ each\ convolutional\ layer,\ we\ project\ the\ nFilters\ feature\ dimensions\ down\ to\ nDimLowerDimConv\ to\ reduce$ 

the number of parameters needed.

nFilters (default = 32L) Integer specifying the number of convolutional filters used.

orthogonalize (default = F) A Boolean specifying whether to perform the image decomposition

after orthogonalizing with respect to tabular covariates specified in X.

imageKeysOfUnits

(default = 1:length(obsY)) A vector of length length(obsY) specifying the unique image ID associated with each unit. Samples of imageKeysOfUnits are fed into acquireImageFxn to call images into memory.

acquire Image Rep Fxn

A function specifying how to load images representations associated with imageKeysOfUnits into memory. For example, if observation 3 has a value of "a34f" in imageKeysOfUnits, acquireImageFxn should extract the image associated with the unique key "a34f". First argument should be image key values and second argument have be training (in case behavior in training/)

acquireImageFxn

(default = acquireImageRepFxn) Similar to acquireImageRepFxn; this is a function specifying how to load images associated with imageKeysOfUnits into memory.

transportabilityMat

(optional) A matrix with a column named keys specifying keys to be used by acquireImageRepFxn for generating treatment effect predictions for outof-sample points.

long, lat (optional) Vectors specifying longitude and latitude coordinates for units. Used only for describing highest and lowest probability neighborhood units if specified.

conda\_env (default = NULL) A string specifying a conda environment wherein tensorflow,

tensorflow\_probability, and gc are installed.

figuresKey (default = "") A string specifying an identifier that is appended to all figure

names.

figuresPath (default = "./") A string specifying file path for saved figures made in the anal-

ysis.

simMode (default = F) Should the analysis be performed in comparison with ground truth

from simulation?

plotResults (default = T) Should analysis results be plotted?

nDepthHidden\_conv

(default = 3L) Hidden depth of convolutional layer.

nDepthHidden\_dense

(default = 0L) Hidden depth of dense layers. Default of 0L means a single projection layer is performed after the convolutional layer (i.e., no hidden layers are

used).

maxPoolSize (default = 2L) Integer specifying the max pooling size used in the convolutional

layers.

strides (default = 2L) Integer specifying the strides used in the convolutional layers.=

yDensity (default = normal) Specifies the density for the outcome. Current options in-

clude normal and lognormal.

nMonte\_variational

(default = 5L) An integer specifying how many Monte Carlo iterations to use in

the calculation of the expected likelihood in each training step.

nMonte\_predictive

(default = 20L) An integer specifying how many Monte Carlo iterations to use

in the calculation of posterior means (e.g., mean cluster probabilities).

nMonte\_salience

(default = 100L) An integer specifying how many Monte Carlo iterations to use in the calculation of the salience maps (e.g., image gradients of expected cluster

probabilities).

batchSize (default = 25L) Batch size used in SGD optimization.

kernelSize (default = 5L) Dimensions used in convolution kernels.

nSGD (default = 400L) Number of stochastic gradient descent (SGD) iterations.

nDenseWidth (default = 32L) Width of dense projection layers post-convolutions.

reparameterizationType

(default = "Flipout") Either "Flipout", or "Reparameterization". Specifies the estimator used in the Bayesian neural components. With "Flipout", convolutions are performed via CPU; with "Reparameterization", they are per-

formed by GPU if available.

channelNormalize

(default = T) Should channelwise image feature normalization be attempted?

Default is T, as this improves training.

quiet (default = F) Should we suppress information about progress?

## Value

A list consiting of

- ATE\_est ATE estimate.
- ATE\_se Standard error estimate for the ATE.

#### References

• TBA.

## **Examples**

```
# For a tutorial, see
# github.com/cjerzak/causalimages-software/
```

BuildBackend

Perform causal estimation under image confounding

## **Description**

Perform causal estimation under image confounding

## Usage

BuildBackend(conda\_env, conda)

## **Arguments**

conda\_env (default = "CausalImagesEnv") Name of the conda environment in which to

place the backends.

conda (default = auto) The path to a conda executable. Using "auto" allows reticulate

to attempt to automatically find an appropriate conda binary.

## Value

Builds the computational environment for causalimages. This function requires an Internet connection. You may find out a list of conda Python paths via: system("which python")

## **Examples**

```
# For a tutorial, see
# github.com/cjerzak/causalimages-software/
```

GetAndSaveGeolocatedImages

Getting and saving geo-located images from a pool of .tif's

## **Description**

A function that finds the image slice associated with the long and lat values, saves images by band (if save\_as = "csv") in save\_folder.

### Usage

```
GetAndSaveGeolocatedImages(long, lat, keys, tif_pool, save_folder,
image_pixel_width, save_as, lyrs)
```

#### **Arguments**

long Vector of numeric longitudes.lat Vector of numeric latitudes.

keys The image keys associated with the long/lat coordinates.

tif\_pool A character vector specifying the fully qualified path to a corpus of .tif files.

image\_pixel\_width

An even integer specifying the pixel width (and height) of the saved images.

save\_folder (default = ".") What folder should be used to save the output? Example: "~/Downloads"

save\_as (default = ".csv") What format should the output be saved as? Only one option

currently (.csv)

lyrs (default = NULL) Integer (vector) specifying the layers to be extracted. Default

is for all layers to be extracted.

#### Value

Finds the image slice associated with the long and lat values, saves images by band (if save\_as = "csv") in save\_folder. The save format is: sprintf("%s/Key%s\_BAND%s.csv", save\_folder, keys[i], band\_)

### **Examples**

GetElementFromTfRecordAtIndices

 $\it Reads \ unique \ key \ indices \ from \ a$  . tfrecord  $\it file.$ 

## **Description**

Reads unique key indices from a .tfrecord file saved via a call to causalimages::WriteTfRecord.

## Usage

## **Arguments**

uniqueKeyIndices

(integer vector) Observation indices to be retrieved from a .tfrecord

conda\_env

(Default = NULL) A conda environment where tensorflow v2 lives. Used only if

a version of tensorflow is not already active.

conda\_env\_required

(default = F) A Boolean stating whether use of the specified conda environment

is required.

file

(character string) A character string stating the path to a .tfrecord

#### Value

Returns content from a .tfrecord associated with uniqueKeyIndices

#### **Examples**

```
# Example usage (not run):
#GetElementFromTfRecordAtIndices(
  #uniqueKeyIndices = 1:10,
  #file = "./NigeriaConfoundApp.tfrecord")
```

#### GetImageRepresentations

Generates image and video representations useful in earth observation tasks for casual inference, following the approach in Rolf, Esther, et al. (2021).

## **Description**

Generates image and video representations useful in earth observation tasks for casual inference, following the approach in Rolf, Esther, et al. (2021).

## **Usage**

```
GetImageRepresentations(imageKeysOfUnits, file, nWidth_ImageRep,
 conda_env, conda_env_required, kernelSize,
 temporalKernelSize, nWidth_ImageRep,
 strides, InitImageProcess, batchSize,
 TfRecords_BufferScaler, seed, dataType)
```

#### **Arguments**

imageKeysOfUnits

A vector of length length(imageKeysOfUnits) specifying the unique image ID associated with each unit. Samples of imageKeysOfUnits are fed into the

package to call images into memory.

file (default = NULL) Path to a tfrecord file generated by WriteTfRecord.

(default = "CausalImagesEnv") A conda environment where computational enconda\_env

vironment lives, usually created via causalimages::BuildBackend()

conda\_env\_required

(default = T) A Boolean stating whether use of the specified conda environment is required.

InitImageProcess

(default = NULL)

nWidth\_ImageRep

(default = 128L) Number of embedding features output.

batchSize (default = 16L) Integer specifying batch size in obtaining representations.

strides (default = 2L) Integer specifying the strides used in the convolutional layers.

temporalKernelSize

(default = 2L) Dimensions used in the temporal part of the convolution kernels

if using image sequences.

kernelSize (default = 5L) Dimensif(ions used in the convolution kernels.

TfRecords\_BufferScaler

(default = 10L) The buffer size used in tfrecords mode is batchSize\*TfRecords\_BufferScaler.

Lower TfRecords\_BufferScaler towards 1 if out-of-memory problems.

dataType (default = "image") String specifying whether to assume "image" or "video"

data types.

seed (default = NULL) Integer specifying the seed for pseudo random number genera-

tion.

#### Value

A list containing two items:

- Representations (matrix) A matrix containing image/video representations, with rows corresponding to observations.
- ImageRepArm\_OneObs, ImageRepArm\_batch\_R, ImageRepArm\_batch (functions) Image modeling functions.
- ImageModel\_And\_State\_And\_MPPolicy\_List List containing image model parameters fed into functions.

#### References

• Rolf, Esther, et al. "A generalizable and accessible approach to machine learning with global satellite imagery." *Nature Communications* 12.1 (2021): 4392.

#### **Examples**

```
# For a tutorial, see
```

# github.com/cjerzak/causalimages-software/

12 image2

image2	Visualizing matrices as heatmaps with correct north-south-east-west
	orientation

## **Description**

A function for generating a heatmap representation of a matrix with correct spatial orientation.

## Usage

```
image2( x
xaxt, yaxt, xlab, ylab,
main, cex.main, col.lab,
col.main, cex.lab, box)
```

## **Arguments**

```
(required) The numeric matrix to be visualized.
Χ
                   (default = "") The x-axis tick labels.
xaxt
                   (default = "") The y-axis tick labels.
yaxt
                   (default = "") The x-axis labels.
xlab
ylab
                   (default = "") The y-axis labels.
                   (default = "") The main figure label.
main
                   (default = 1.) The main figure label sizing factor.
cex.main
col.lab
                   (default = "black") Axis label color.
col.main
                   (default = "black") Main label color.
                   (default = 1.5) Cex for the labels.
cex.lab
                   (default = FALSE) Draw a box around the image?
box
```

### Value

Returns a heatmap representation of the matrix, x, with correct north/south/east/west orientation.

## **Examples**

```
#set seed
set.seed(1)

#Geneate data
x <- matrix(rnorm(50*50), ncol = 50)
diag(x) <- 3

# create plot
image2(x, main = "Example Text", cex.main = 2)</pre>
```

LongLat2CRS 13

ı	anal	at 2CRS
	angi	AI/(R)

Get the spatial point of long/lat coordinates

## **Description**

A function converts long/lat coordinates into a spatial points object defined by a coordinate reference system (CRS).

## Usage

```
LongLat2CRS(long, lat, CRS_ref)
```

#### **Arguments**

long Vector of numeric longitudes.lat Vector of numeric latitudes.

CRS\_ref A CRS into which the long-lat point should be projected.

#### Value

Returns the long/lat location as a spatial point in the new CRS defined by CRS\_ref

## **Examples**

SimulateImageSystem

Simulate causal systems involving images

## **Description**

This function generates simulated causal structures using images. It is currently under construction.

## Usage

```
SimulateImageSystem(...)
```

## **Arguments**

dag

(character string) An input DAG specifying causal structure. This input should be of the form 'i->t,i->y,t->y,....' Currently, only one node in a DAG can be an image (this should be labeled "i"). The non-image nodes can have arbitrary string labels. The image can be a confounder, effect moderator, effect mediator. If the image is to be used as a moderator, use the notation, t-i>y.

14 WriteTfRecord

(optional) In estimation mode, users input the data matrices associated with the . . . non-image nodes of DAG and image node i. For example, if x is a DAG node, users must, in estimation mode, supply data to x in a form that can be coerced to a tensor. (character string, optional) In estimation mode, users specify the treatment varitreatment able here. If treatment is specified, users must provide other data inputs to the DAG (see ...). (character string, optional) The path to where analysis specific images are loimage\_pool cated. This can be specified both in simulation and estimation mode. If not specified, the simulation uses a pool of Landsat images from Nigeria. (character string, default is 'scene') Defines the unit of analysis used in the analysis\_level simulation framework. This is ignored in estimation mode, where the unit of analysis is inferred from the data dimensions.

Value

A list:

control

• In *simulation mode*, the function returns a list with as many elements as unique nodes in DAG. Each element represents the simulated data.

(list) A list containing control parameters in the data generating process.

• In estimation mode, the function returns an estimated treatment effect with 95\

#### References

Connor T. Jerzak, Fredrik Johansson, Adel Daoud. Image-based Treatment Effect Heterogeneity. Forthcoming in *Proceedings of the Second Conference on Causal Learning and Reasoning (CLeaR)*, *Proceedings of Machine Learning Research (PMLR)*, 2023.

## **Examples**

```
#set seed
set.seed(1)

# Simulation mode
#simulatedData <- causalimage('r->i, i->t, t->y, r->y')
#print(names(simulatedData))

# Estimation mode
#estimatedResults <- causalimage('r->i, i->t, t->y, r->y', y=y, r=r, y=y', treatment='t')
#print( estimatedResults )
```

WriteTfRecord

Write an image corpus as a .tfrecord file

### **Description**

Writes an image corpus to a .tfrecord file for rapid reading of images into memory for fast ML training.

WriteTfRecord 15

## Usage

```
WriteTfRecord(file, imageKeysOfUnits, acquireImageFxn,
conda_env, conda_env_required, writeVideo)
```

## **Arguments**

file A character string naming a file for writing.

uniqueImageKeys

A vector specifying the unique image keys of the corpus. A key grabs an im-

age/video array via acquireImageFxn(key)

acquire Image Fxn

A function whose input is an observation index and whose output is an image.

writeVideo (default = FALSE) Should we assume we're writing image sequence data of form

batch by time by height by width by channels?

conda\_env (default = "Causal Images Env") A conda environment where computational en-

vironment lives, usually created via causalimages::BuildBackend()

conda\_env\_required

(default = T) A Boolean stating whether use of the specified conda environment

is required.

#### Value

Writes a unique key-referenced .tfrecord from an image/video corpus for use in image-based causal inference training.

## **Examples**

```
# Example usage (not run):
#WriteTfRecord(
# file = "./NigeriaConfoundApp.tfrecord",
# uniqueImageKeys = 1:n,
# acquireImageFxn = acquireImageFxn)
```

# **Index**

```
AnalyzeImageConfounding, 2
AnalyzeImageHeterogeneity, 3
AnalyzeImageMediation, 6
BuildBackend, 8
GetAndSaveGeolocatedImages, 8
GetElementFromTfRecordAtIndices, 9
GetImageRepresentations, 10
image2, 12
LongLat2CRS, 13
SimulateImageSystem, 13
WriteTfRecord, 14
```