Identification of interpretable components' importance

Arkadiusz Tomczyk, Bartłomiej Wójcik
arkadiusz.tomczyk@p.lodz.pl,bartlomiej.wojcik@dokt.p.lodz.pl

Institute of Information Technology
Lodz University of Technology
Poland

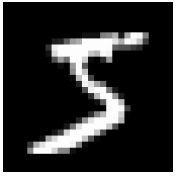
20-21.11.2024

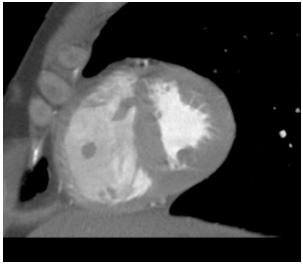
Background

- we are interested in data with internal structure

Background

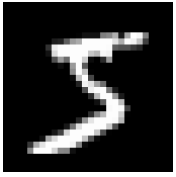
- we are interested in data with internal structure
  - images $\rightarrow$ *grids*

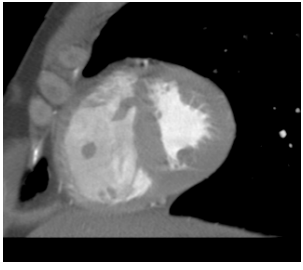




Background

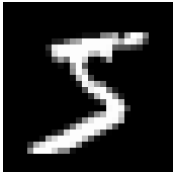
- we are interested in data with internal structure
  - images $\rightarrow$ *grids*
  - texts and series $\rightarrow$ *sequences*

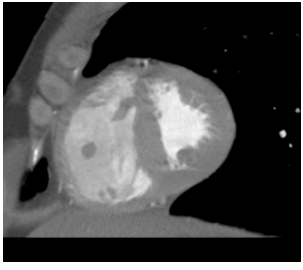




Background

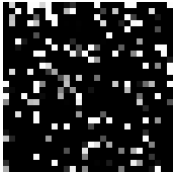
- we are interested in data with internal structure
  - images $\rightarrow$ *grids*
  - texts and series $\rightarrow$ *sequences*
  - chemical molecules, social networks, etc. $\rightarrow$ *graphs*

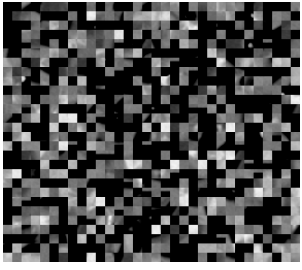




Background

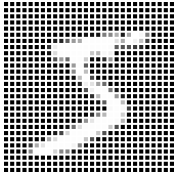
- we are interested in data with internal structure
  - images → *grids*
  - texts and series → *sequences*
  - chemical molecules, social networks, etc. → *graphs*
- properties of such data depend on both components and relations between them

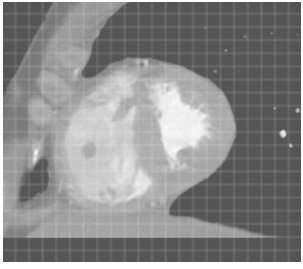




Background

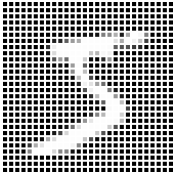
- we are interested in data with internal structure
  - images $\rightarrow$ *grids*
  - texts and series $\rightarrow$ *sequences*
  - chemical molecules, social networks, etc. $\rightarrow$ *graphs*
- properties of such data depend on both components and relations between them
- frequently we have to deal with too complex structures, which has several drawbacks:

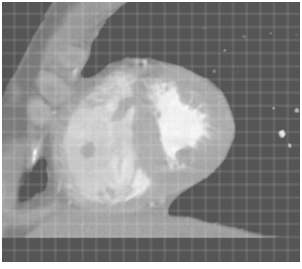




Background

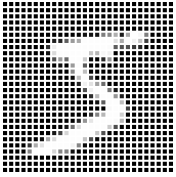
- we are interested in data with internal structure
  - images → *grids*
  - texts and series → *sequences*
  - chemical molecules, social networks, etc. → *graphs*
- properties of such data depend on both components and relations between them
- frequently we have to deal with too complex structures, which has several drawbacks:
  - design of classic algorithms is hard

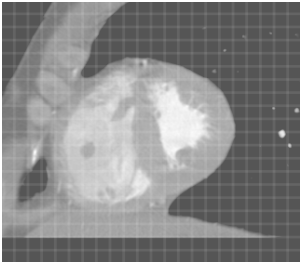




Background

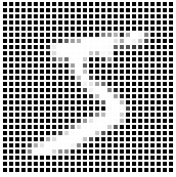
- we are interested in data with internal structure
  - images → *grids*
  - texts and series → *sequences*
  - chemical molecules, social networks, etc. → *graphs*
- properties of such data depend on both components and relations between them
- frequently we have to deal with too complex structures, which has several drawbacks:
  - design of classic algorithms is hard
  - neural models are very complex

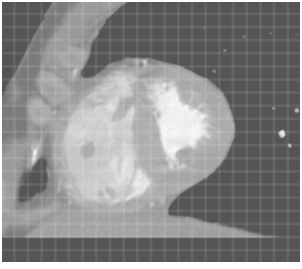




Background

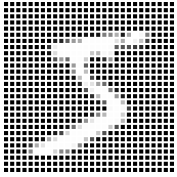
- we are interested in data with internal structure
  - images → *grids*
  - texts and series → *sequences*
  - chemical molecules, social networks, etc. → *graphs*
- properties of such data depend on both components and relations between them
- frequently we have to deal with too complex structures, which has several drawbacks:
  - design of classic algorithms is hard
  - neural models are very complex
  - it is hard to posses knowledge from domain experts

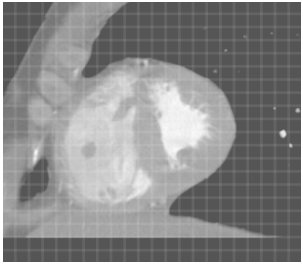




Background

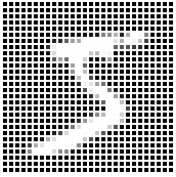
- we are interested in data with internal structure
  - images $\rightarrow$ *grids*
  - texts and series $\rightarrow$ *sequences*
  - chemical molecules, social networks, etc. $\rightarrow$ *graphs*
- properties of such data depend on both components and relations between them
- frequently we have to deal with too complex structures, which has several drawbacks:
  - design of classic algorithms is hard
  - neural models are very complex
  - it is hard to posses knowledge from domain experts
  - it is hard to explain results to domain experts

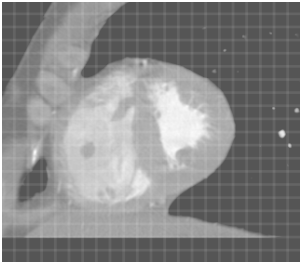




Background

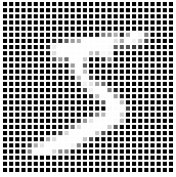
- we are interested in data with internal structure
  - images $\rightarrow$ *grids*
  - texts and series $\rightarrow$ *sequences*
  - chemical molecules, social networks, etc. $\rightarrow$ *graphs*
- properties of such data depend on both components and relations between them
- frequently we have to deal with too complex structures, which has several drawbacks:
  - design of classic algorithms is hard
  - neural models are very complex
  - it is hard to posses knowledge from domain experts
  - it is hard to explain results to domain experts
  - knowledge discovery is difficult

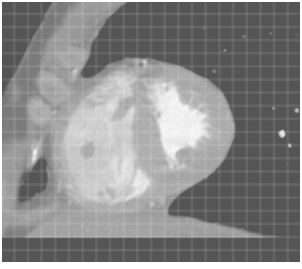




Background

- we are interested in data with internal structure
  - images → *grids*
  - texts and series → *sequences*
  - chemical molecules, social networks, etc. → *graphs*
- properties of such data depend on both components and relations between them
- frequently we have to deal with too complex structures, which has several drawbacks:
  - design of classic algorithms is hard
  - neural models are very complex
  - it is hard to posses knowledge from domain experts
  - it is hard to explain results to domain experts
  - knowledge discovery is difficult
- fortunately, we can change the internal representation of data



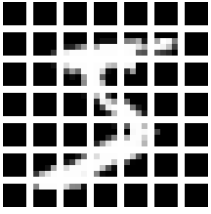


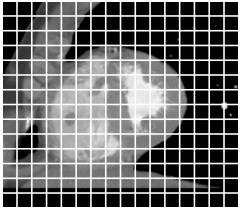
Background

- alternative representations are not entirely new:

Background

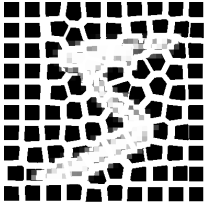
- alternative representations are not entirely new:
  - patches $\rightarrow$ *regular structure*, *ViT* [1]

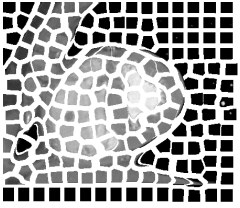




Background

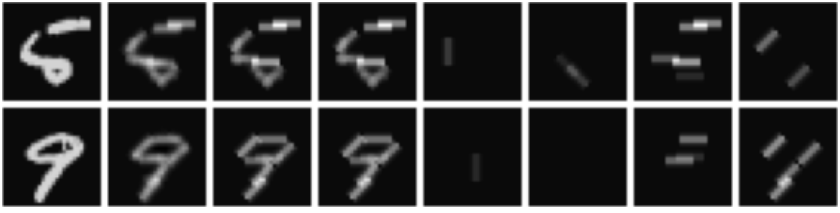
- alternative representations are not entirely new:
  - patches $\rightarrow$ *regular structure*, *ViT* [1]
  - segments $\rightarrow$ *irregular structure*, *LIME* [2]





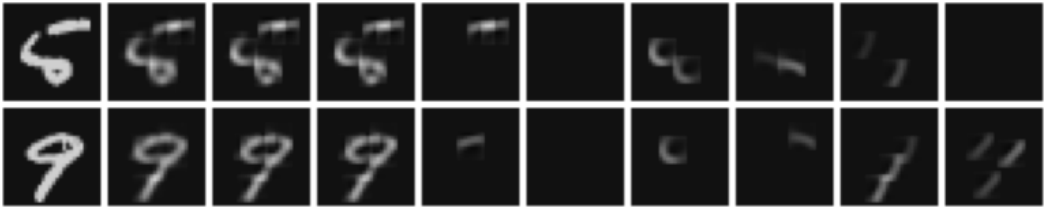
Background

- alternative representations are not entirely new:
  - patches → *regular structure*, *ViT* [1]
  - segments → *irregular structure*, *LIME* [2]
  - sparse coding → *irregular structure*, *trainable* [3]



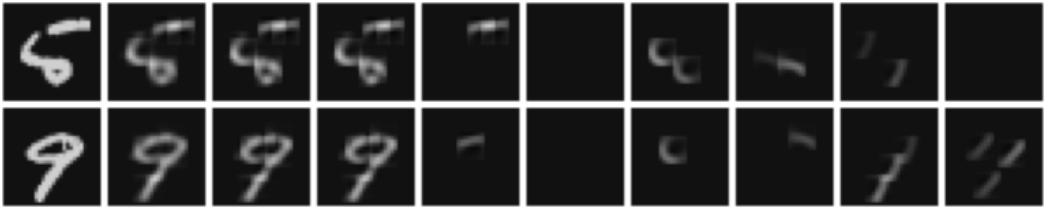
Background

- alternative representations are not entirely new:
  - patches → *regular structure*, *ViT* [1]
  - segments → *irregular structure*, *LIME* [2]
  - sparse coding → *irregular structure*, *trainable* [3]



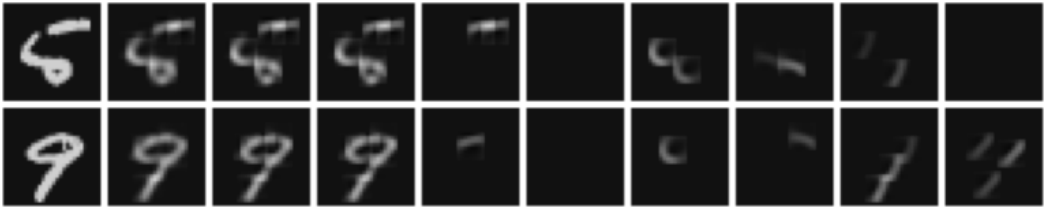
Background

- alternative representations are not entirely new:
  - patches → *regular structure*, *ViT* [1]
  - segments → *irregular structure*, *LIME* [2]
  - sparse coding → *irregular structure*, *trainable* [3]
- alternative representations are not only related to images:



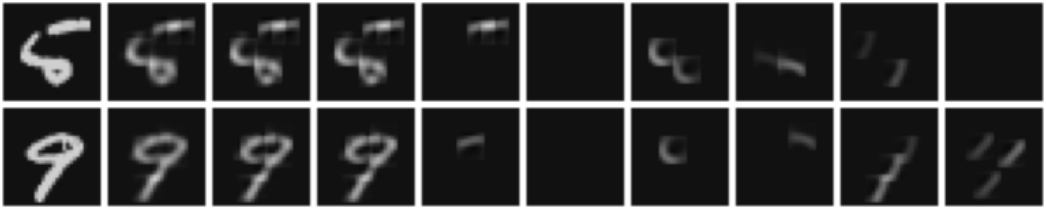
Background

- alternative representations are not entirely new:
  - patches → *regular structure*, *ViT* [1]
  - segments → *irregular structure*, *LIME* [2]
  - sparse coding → *irregular structure*, *trainable* [3]
- alternative representations are not only related to images:
  - sentences → *tokens*



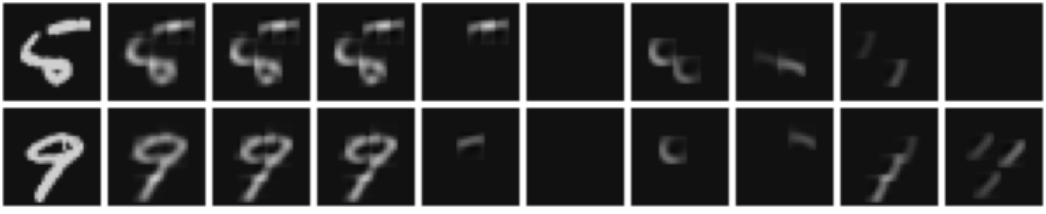
Background

- alternative representations are not entirely new:
  - patches → *regular structure*, *ViT* [1]
  - segments → *irregular structure*, *LIME* [2]
  - sparse coding → *irregular structure*, *trainable* [3]
- alternative representations are not only related to images:
  - sentences → *tokens*
  - proteins → *secondary structures*



Background

- alternative representations are not entirely new:
  - patches → *regular structure*, *ViT* [1]
  - segments → *irregular structure*, *LIME* [2]
  - sparse coding → *irregular structure*, *trainable* [3]
- alternative representations are not only related to images:
  - sentences → *tokens*
  - proteins → *secondary structures*



- it would be optimal if the components were interpretable

Motivation

- depending on component and structure type, we can use different neural models to process them $\rightarrow$ *MLP, CNN, LSTM, Transformer, GNN* [4, 5, 6, 7]

Motivation

- depending on component and structure type, we can use different neural models to process them $\rightarrow$ *MLP, CNN, LSTM, Transformer, GNN* [4, 5, 6, 7]
- those models are usually black boxes, which working principles are hard to understand for humans $\rightarrow$ *quality of post-hoc explainability methods is still questionable*

Motivation

- depending on component and structure type, we can use different neural models to process them → *MLP, CNN, LSTM, Transformer, GNN* [4, 5, 6, 7]
- those models are usually black boxes, which working principles are hard to understand for humans → *quality of post-hoc explainability methods is still questionable*
- the reason for this is that used architectures are not intuitive for humans and are usually overdesigned → *this leads to completely unexpected solutions*

Motivation

- depending on component and structure type, we can use different neural models to process them → *MLP, CNN, LSTM, Transformer, GNN* [4, 5, 6, 7]
- those models are usually black boxes, which working principles are hard to understand for humans → *quality of post-hoc explainability methods is still questionable*
- the reason for this is that used architectures are not intuitive for humans and are usually overdesigned → *this leads to completely unexpected solutions*

Design architecture that allows to discover important structure components.

Motivation

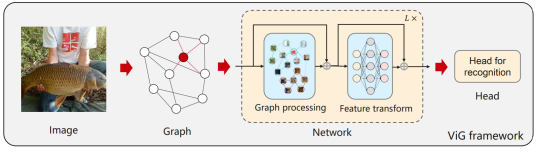
- depending on component and structure type, we can use different neural models to process them → *MLP, CNN, LSTM, Transformer, GNN* [4, 5, 6, 7]
- those models are usually black boxes, which working principles are hard to understand for humans → *quality of post-hoc explainability methods is still questionable*
- the reason for this is that used architectures are not intuitive for humans and are usually overdesigned → *this leads to completely unexpected solutions*

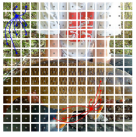
Design architecture that allows to discover important structure components.

Design architectures with interpretable working principles.

Method

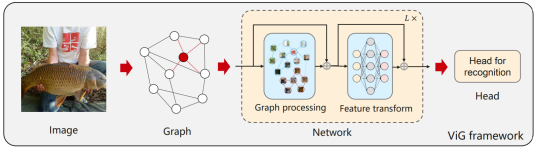
- we started with ViG [8]

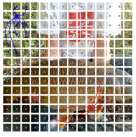




Method

- we started with ViG [8]

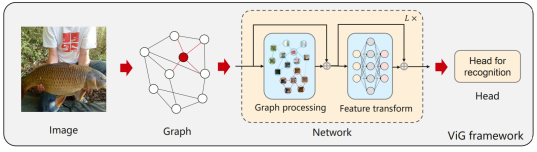


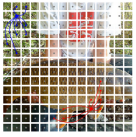


- but this method had a few drawbacks:

Method

- we started with ViG [8]

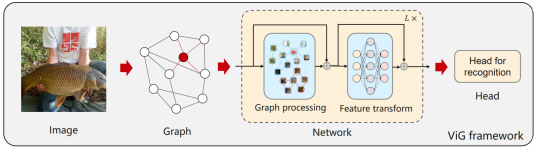


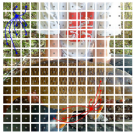


- but this method had a few drawbacks:
  - it used k-NN to determine neighbour patches in the graph

Method

- we started with ViG [8]

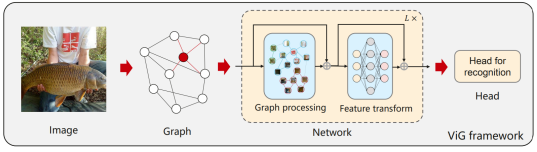


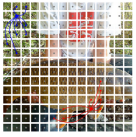


- but this method had a few drawbacks:
  - it used k-NN to determine neighbour patches in the graph
  - the receptive field of each patch was much larger than in visualizations

Method

- we started with ViG [8]





- but this method had a few drawbacks:
  - it used k-NN to determine neighbour patches in the graph
  - the receptive field of each patch was much larger than in visualizations
  - the method allowed finding similar patches, but did not explain which patches were important for classification

Improvments

- we noticed that adding trainable weights for edges improves classification and gives us more interpretability

Improvments

- we noticed that adding trainable weights for edges improves classification and gives us more interpretability
- to further improve the possibility of explaining network we started simplifying it by:

Improvments

- we noticed that adding trainable weights for edges improves classification and gives us more interpretability
- to further improve the possibility of explaining network we started simplifying it by:
  - adjusting receptive field of network to avoid overlapping of patches

Improvments

- we noticed that adding trainable weights for edges improves classification and gives us more interpretability
- to further improve the possibility of explaining network we started simplifying it by:
  - adjusting receptive field of network to avoid overlapping of patches
  - instead of computing edges using k-NN we used single weight for each patch

Improvments

- we noticed that adding trainable weights for edges improves classification and gives us more interpretability
- to further improve the possibility of explaining network we started simplifying it by:
  - adjusting receptive field of network to avoid overlapping of patches
  - instead of computing edges using k-NN we used single weight for each patch
  - replacing graph convolution layer with simple mean aggregation

Improvments

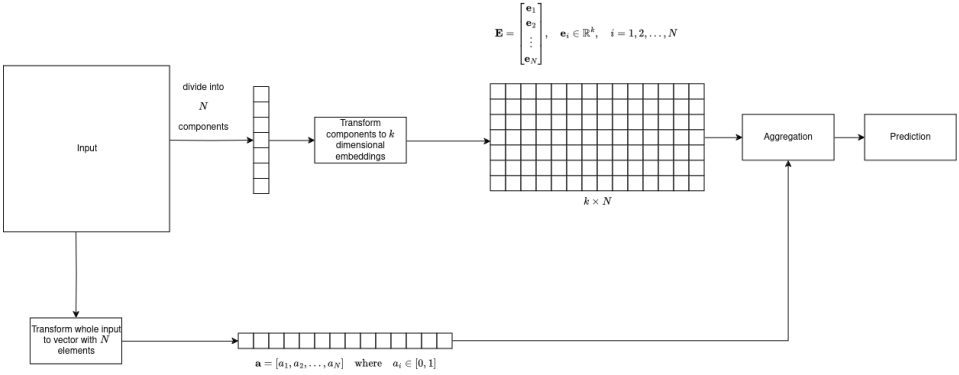
- we noticed that adding trainable weights for edges improves classification and gives us more interpretability
- to further improve the possibility of explaining network we started simplifying it by:
  - adjusting receptive field of network to avoid overlapping of patches
  - instead of computing edges using k-NN we used single weight for each patch
  - replacing graph convolution layer with simple mean aggregation
  - replacing CNN based head with simple linear layers

Architecture

- these experiments allowed us to create general framework for building networks which can discover meaningful structure components in different domains and problems.

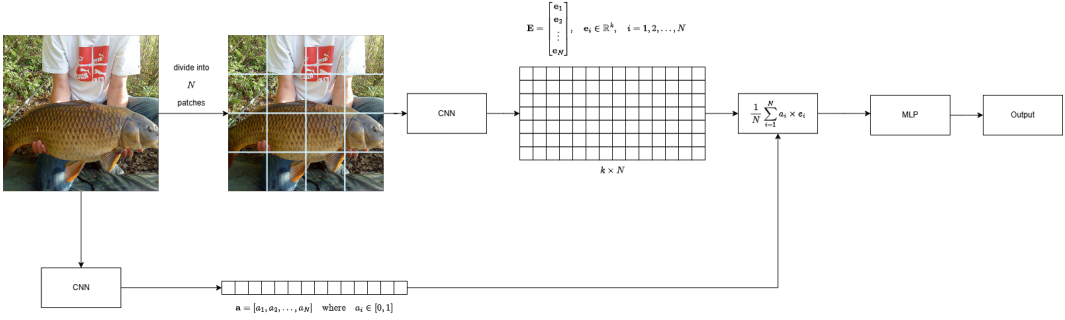
Architecture

- these experiments allowed us to create general framework for building networks which can discover meaningful structure components in different domains and problems.



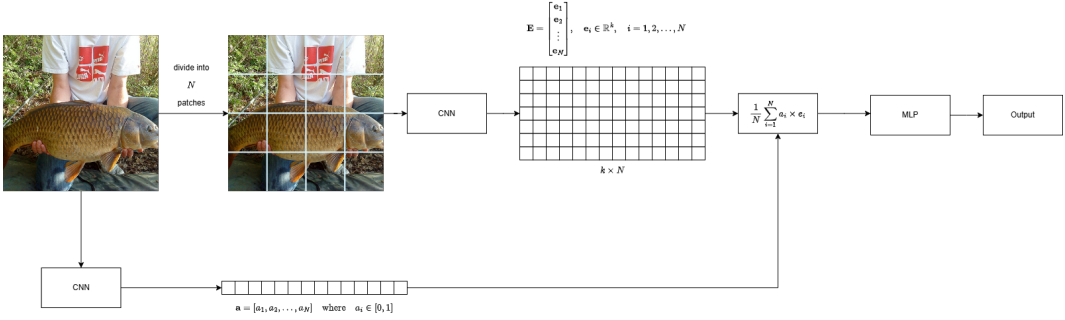
Approach

- this general framework can be applied for images split into components (patches)

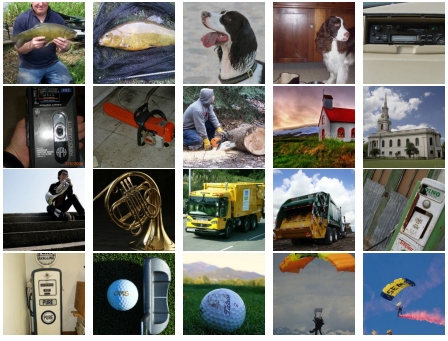


Approach

- this general framework can be applied for images split into components (patches)
- there is one CNN network creating N component embeddings of size k and second CNN network with N outputs that represent components importance



Results



(a) input images

(b) patch importances

Results

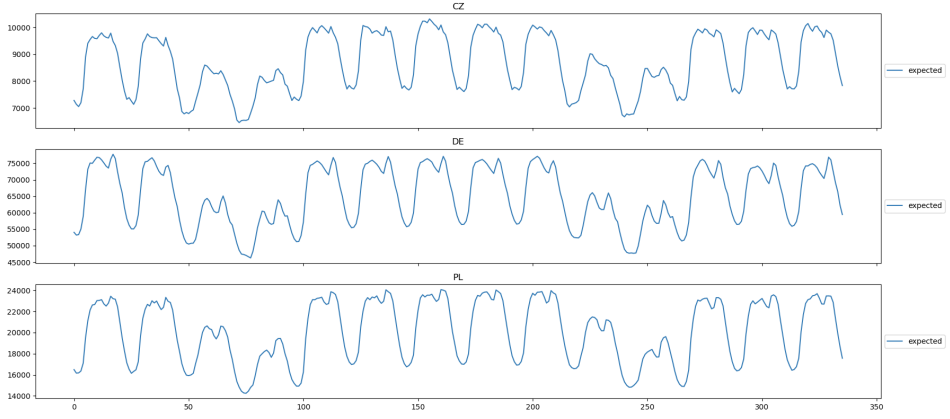
method	without importance		with importance	
	mean [%]	standard deviation [%]	mean [%]	standard deviation [%]
ours ($k = 2048$)	67.03	0.88	**68.81**	0.63
ours ($k = 1024$)	66.28	0.40	67.69	0.82
ours ($k = 512$)	65.33	1.28	66.30	1.49
ours ($k = 256$)	64.70	0.54	64.59	1.50
ours ($k = 128$)	62.63	1.01	62.56	1.01
ours ($k = 64$)	60.97	0.24	61.87	1.78
ours ($k = 8$)	54.25	0.59	54.01	1.31

Table: Test set classification evaluation (accuracy) for different embedding sizes k (statistics from 3 runs).

method	mean [%]	standard deviation [%]
ResNet-18	73.49	0.24

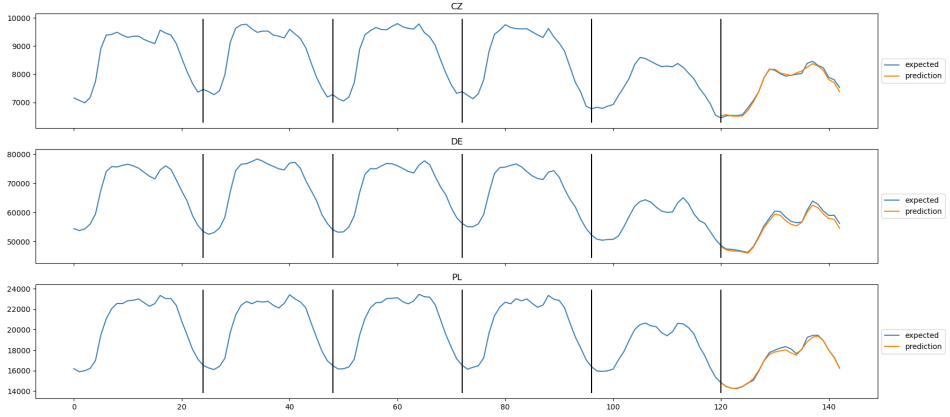
Data

- energy consumption in selected European countries measured once per hour



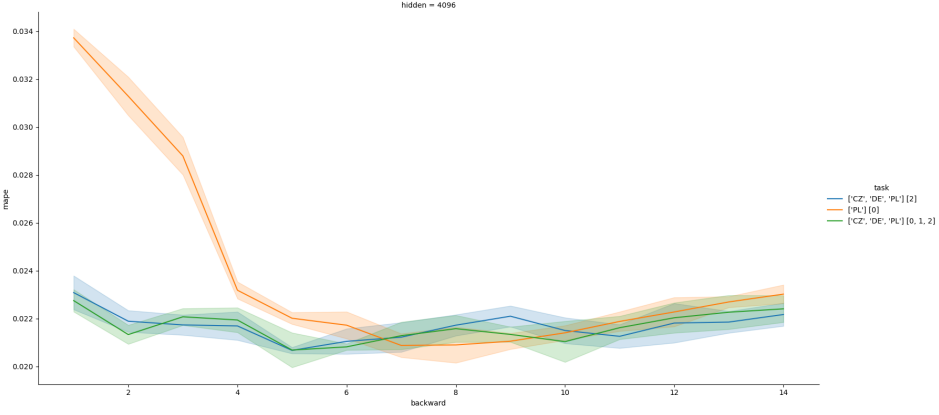
Task

- predict energy consumption for the next day basing on some number of backward days N



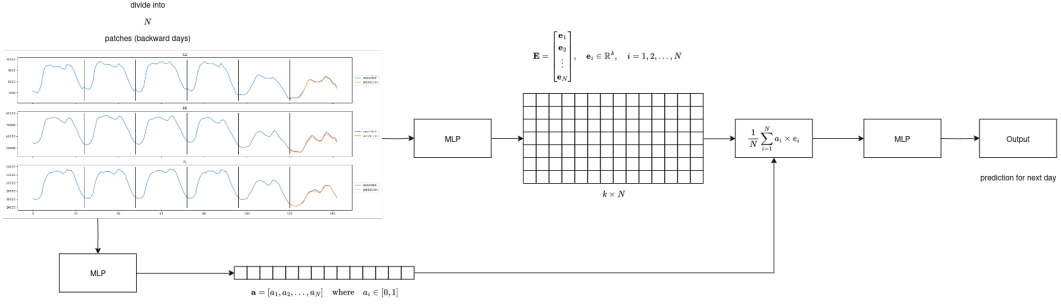
Reference

- we can do that using MLP network with $N \times 24$ inputs and 24 outputs



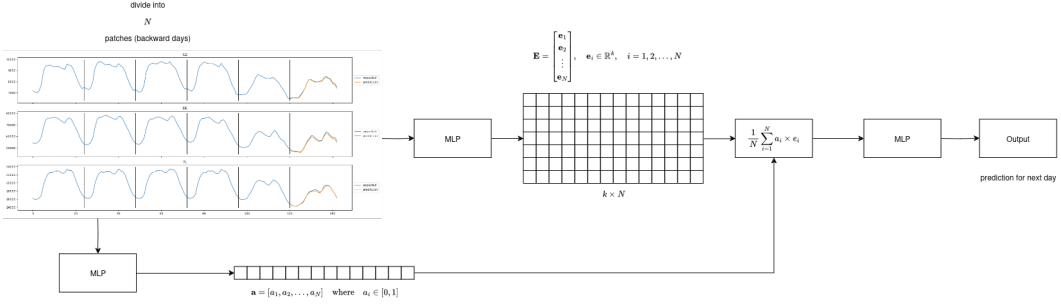
Approach

- in our approach we treat every day is a separate component (patch)



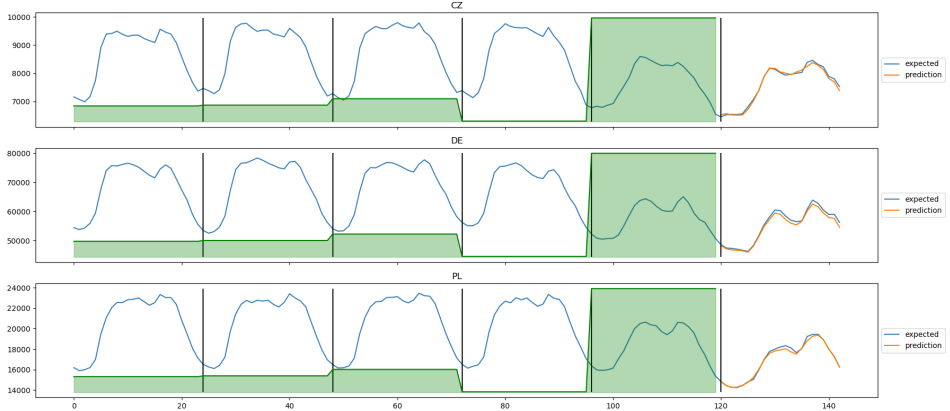
Approach

- in our approach we treat every day is a separate component (patch)
- there is one MLP network with 24 inputs and k outputs creating component embeddings and second MLP network with $N \times 24$ inputs and N outputs that represent components importance

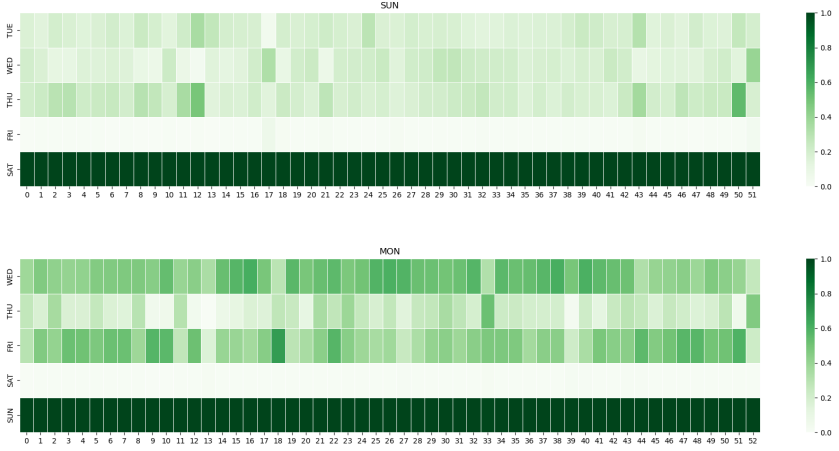


Results

- the proposed approach allows to discover, which backward days can have the highest influence on currently predicted day

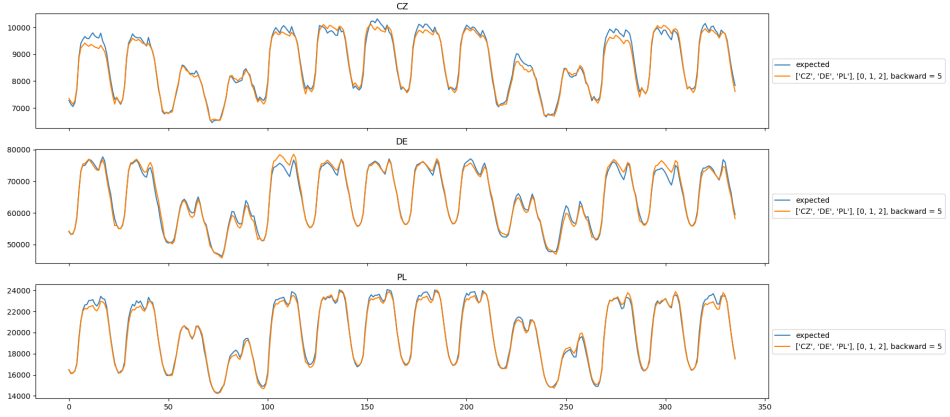


Results



- model on its own was able to discover specific patterns for every day of the week

Results



- although some components appeared to be less important prediction results were still satisfactory

Results

method	mean [%]	standard deviation [%]	min [%]
MLP ($N = 5$)	**2.07**	0.07	1.97
MLP ($N = 6$)	2.08	0.01	2.07
MLP ($N = 7$)	2.13	0.06	2.07
ours ($N = 5$, $k = 256$)	2.10	0.16	**1.95**
ours ($N = 5$, $k = 64$)	2.15	0.15	**1.89**
ours ($N = 6$, $k = 64$)	2.32	0.13	2.18
ours ($N = 7$, $k = 64$)	2.33	0.15	2.17
ours ($N = 6$, $k = 256$)	2.41	0.14	2.28
ours ($N = 7$, $k = 256$)	2.42	0.26	2.01

Table: Test set prediction evaluation (MAPE) for different architectures, component embedding sizes k, and backward days N (statistics from 5 runs).

- work on time series analysis is carried out in cooperation with prof. Grzegorz Dudek

Summary

- proposed models, while processing data, allow to assign importances to internal components of the structure

Summary

- proposed models, while processing data, allow to assign importances to internal components of the structure
- used architectures do not decrease models learning quality

Summary

- proposed models, while processing data, allow to assign importances to internal components of the structure
- used architectures do not decrease models learning quality
- analysis of the observed importances can lead to knowledge discovery

Summary

- proposed models, while processing data, allow to assign importances to internal components of the structure
- used architectures do not decrease models learning quality
- analysis of the observed importances can lead to knowledge discovery
- future research directions:

Summary

- proposed models, while processing data, allow to assign importances to internal components of the structure
- used architectures do not decrease models learning quality
- analysis of the observed importances can lead to knowledge discovery
- future research directions:
  - more insight into current results

Summary

- proposed models, while processing data, allow to assign importances to internal components of the structure
- used architectures do not decrease models learning quality
- analysis of the observed importances can lead to knowledge discovery
- future research directions:
  - more insight into current results
  - more challenging image datasets

Summary

- proposed models, while processing data, allow to assign importances to internal components of the structure
- used architectures do not decrease models learning quality
- analysis of the observed importances can lead to knowledge discovery
- future research directions:
  - more insight into current results
  - more challenging image datasets
  - sparsity extortion

Summary

- proposed models, while processing data, allow to assign importances to internal components of the structure
- used architectures do not decrease models learning quality
- analysis of the observed importances can lead to knowledge discovery
- future research directions:
  - more insight into current results
  - more challenging image datasets
  - sparsity extortion
  - spatial relationships $\rightarrow$ *proper model, positional encoding*

Summary

- proposed models, while processing data, allow to assign importances to internal components of the structure
- used architectures do not decrease models learning quality
- analysis of the observed importances can lead to knowledge discovery
- future research directions:
  - more insight into current results
  - more challenging image datasets
  - sparsity extortion
  - spatial relationships $\rightarrow$ *proper model, positional encoding*
  - application for irregular structures $\rightarrow$ *graphs*

Summary

- proposed models, while processing data, allow to assign importances to internal components of the structure
- used architectures do not decrease models learning quality
- analysis of the observed importances can lead to knowledge discovery
- future research directions:
  - more insight into current results
  - more challenging image datasets
  - sparsity extortion
  - spatial relationships $\rightarrow$ *proper model, positional encoding*
  - application for irregular structures $\rightarrow$ *graphs*
  - other problems and domains $\rightarrow$ *structure prediction, chemistry*

Summary

- proposed models, while processing data, allow to assign importances to internal components of the structure
- used architectures do not decrease models learning quality
- analysis of the observed importances can lead to knowledge discovery
- future research directions:
  - more insight into current results
  - more challenging image datasets
  - sparsity extortion
  - spatial relationships $\rightarrow$ *proper model, positional encoding*
  - application for irregular structures $\rightarrow$ *graphs*
  - other problems and domains $\rightarrow$ *structure prediction, chemistry*

Questions?

References I

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby.
An image is worth 16x16 words: Transformers for image recognition at scale.
CoRR, abs/2010.11929, 2020.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin.
"Why should i trust you?": Explaining the predictions of any classifier.
In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[3] Brendt Wohlberg.
SPORCO: A Python package for standard and convolutional sparse representations.
In *Proceedings of the 15th Python in Science Conference*, pages 1–8, Austin, TX, USA, July 2017.

References II

[4] Thomas N. Kipf and Max Welling.
Semi-supervised classification with graph convolutional networks.
CoRR, abs/1609.02907, 2016.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby.
An image is worth 16x16 words: Transformers for image recognition at scale.
CoRR, abs/2010.11929, 2020.

[6] Shaked Brody, Uri Alon, and Eran Yahav.
How attentive are graph attention networks?
CoRR, abs/2105.14491, 2021.

[7] Yunsheng Shi, Zhengjie Huang, Wenjin Wang, Hui Zhong, Shikun Feng, and Yu Sun.
Masked label prediction: Unified massage passing model for semi-supervised classification.
CoRR, abs/2009.03509, 2020.

References III

[8] Kai Han, Yunhe Wang, Jianyuan Guo, Yehui Tang, and Enhua Wu.
Vision gnn: An image is worth graph of nodes.
Advances in neural information processing systems, 35:8291–8303, 2022.