Graph neural networks for information extraction

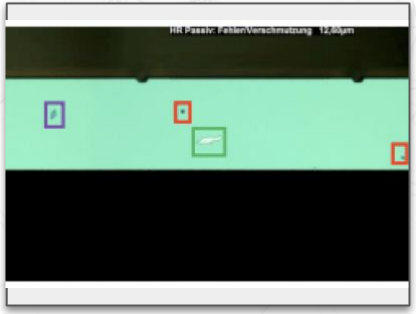
EuroPython 2021

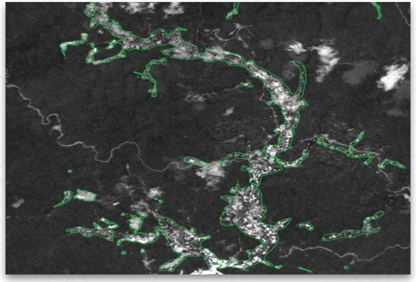
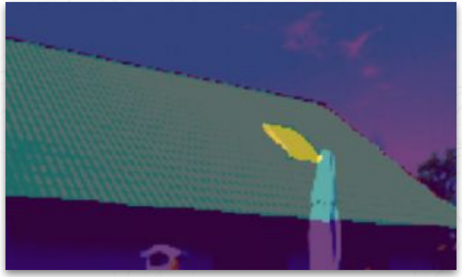
Augusto Stoffel

dida Machine Learning

dida

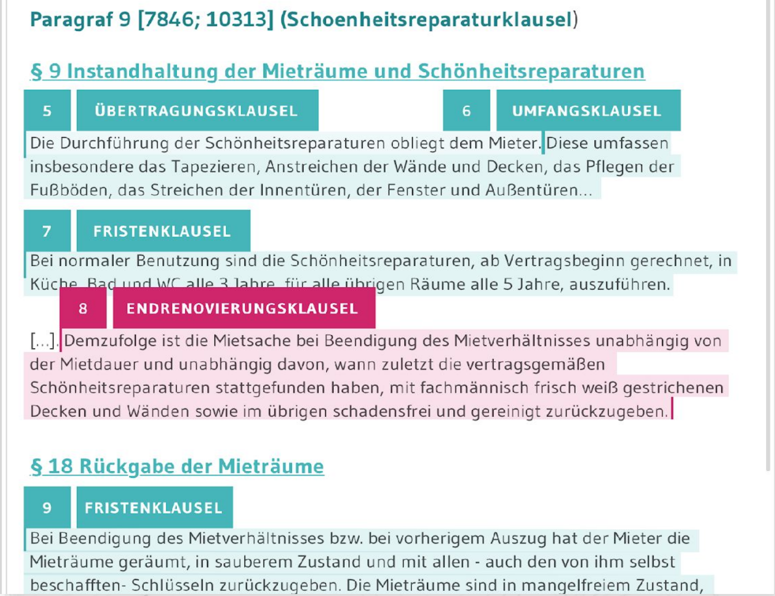
Deep learning landscape



Computer Vision

Example Project



Natural Language Processing (NLP)

Example Project

dida

Sample GNN use case

Dewey Hobbs GmbH

Dewey Hobbs – Hobbsstraße 21 – 12345 Hobbshausen

ABC Chemicals AG
Kochstraße 2
5678 Salzstadt

Dewey Hobbs GmbH
Deweystraße 21
12345 Hobbshausen

Tel.: 0211 12345 67
E-Mail: accounting@hobbs.de
Internet: www.deweyhobbs.de

Rechnung

Rechnung Nr. 2020-06-1001   Bestellung-Nr.: KA12345   Kunden-Nr.: 1001

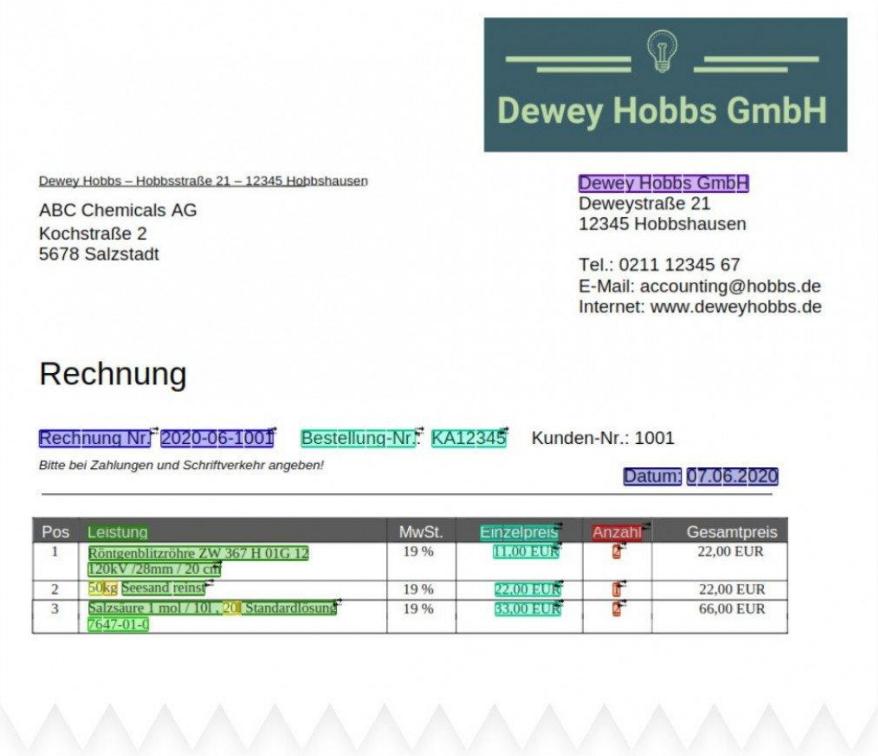
Bitte bei Zahlungen und Schriftverkehr angeben!

Datum: 07.06.2020

Pos	Leistung	MwSt.	Einzelpreis	Anzahl	Gesamtpreis
1	Röntgenblitzröhre ZW 367 H 01G 12 120kV /28mm / 20 cm	19 %	11,00 EUR	2	22,00 EUR
2	50kg Seesand reinst	19 %	22,00 EUR	1	22,00 EUR
3	Salzsäure 1 mol / 10l , 20l Standardlösung 7647-01-0	19 %	33,00 EUR	2	66,00 EUR

dida

Sample GNN use case


Dewey Hobbs GmbH

Dewey Hobbs – Hobbsstraße 21 – 12345 Hobbshausen

ABC Chemicals AG
Kochstraße 2
5678 Salzstadt

Dewey Hobbs GmbH
Deweystraße 21
12345 Hobbshausen

Tel.: 0211 12345 67
E-Mail: accounting@hobbs.de
Internet: www.deweyhobbs.de

Rechnung

Rechnung Nr.: 2020-06-1001 Bestellung-Nr.: KA12345 Kunden-Nr.: 1001

Bitte bei Zahlungen und Schriftverkehr angeben!

Datum: 07.06.2020

Pos	Leistung	MwSt.	Einzelpreis	Anzahl	Gesamtpreis
1	Röntgenblitzröhre ZW 367 H 01G 12 120kV /28mm / 20 cm	19 %	11,00 EUR	2	22,00 EUR
2	50kg Seesand reinst	19 %	22,00 EUR	1	22,00 EUR
3	Salzsäure 1 mol / 10l , 20l Standardlösung 7647-01-0	19 %	33,00 EUR	2	66,00 EUR

dida

A typical information extraction pipeline

Where do graphs enter in this picture?

- Encode information in a graph
- Use graph algorithms to process the information

A typical information extraction pipeline

Where do graphs enter in this picture?

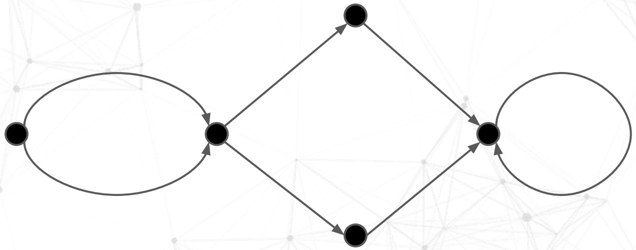
- Encode information in a graph
- Use graph algorithms to process the information

Agenda:

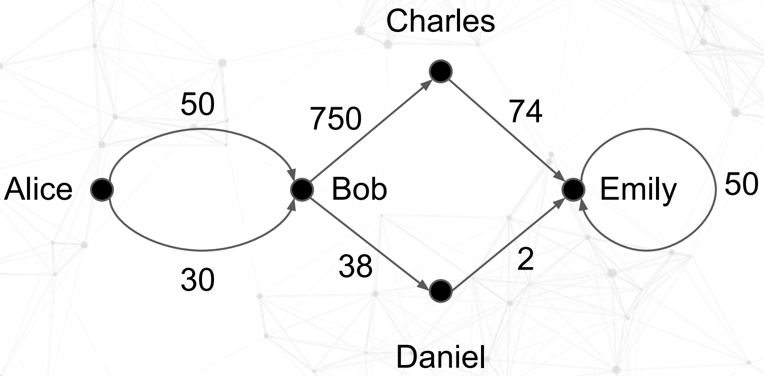
- Graphs and GNNs
- Comparison with convolutional networks
- Use case

dida

What is a graph?



What is a graph?

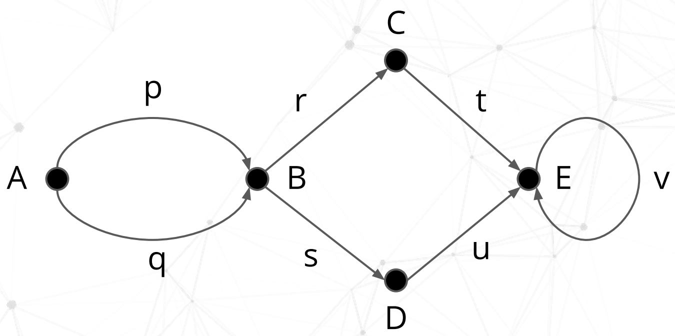


Variations: simple graph, multigraph, pseudograph; directed or not

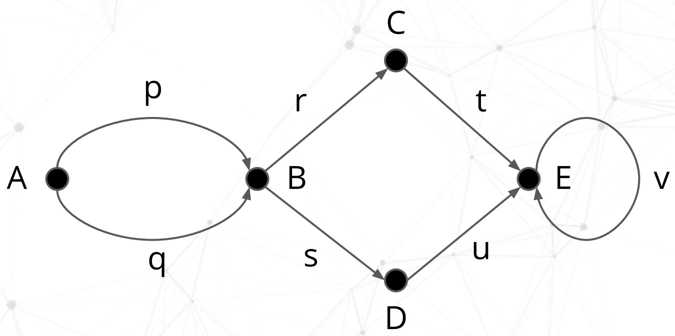
What is a graph?

Definition: A *directed pseudograph*, or simply *graph*, is

- a set $\mathcal{V}$ of vertices
- a set $\mathcal{E}$ of edges
- functions source, target: $\mathcal{E} \rightarrow \mathcal{V}$



What is a graph?



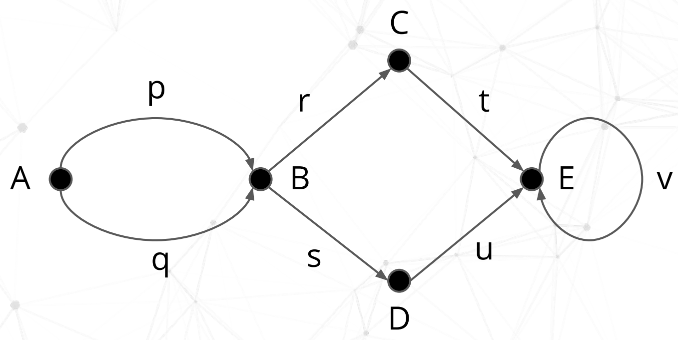
Definition: A *directed pseudograph*, or simply *graph*, is

- a set $\mathcal{V}$ of vertices
- a set $\mathcal{E}$ of edges
- functions source, target: $\mathcal{E} \to \mathcal{V}$

Example:

- $\mathcal{V}$ = {A, B, C, ..., E}
- $\mathcal{E}$ = {p, q, r, ..., v}
- source(p) = A
- target(p) = B
- source(v) = E
- target(v) = E

What is a graph?



Definition: A *directed pseudograph*, or simply *graph*, is

- a set $\mathcal{V}$ of vertices
- a set $\mathcal{E}$ of edges
- functions source, target: $\mathcal{E} \rightarrow \mathcal{V}$

Example:

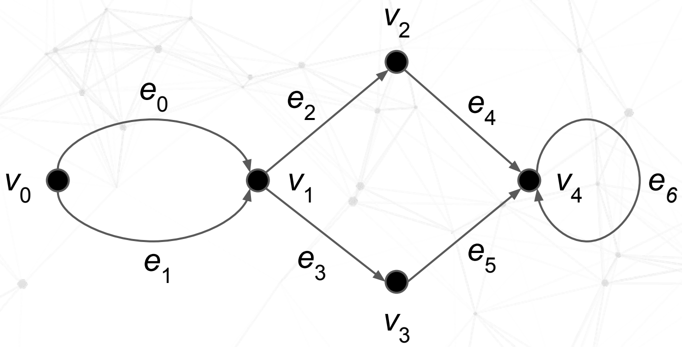
- $\mathcal{V}$ = {A, B, C, ..., E}
- $\mathcal{E}$ = {p, q, r, ..., v}
- source(p) = A
- target(p) = B
- source(v) = E
- target(v) = E

Note: Often there are labels associated to vertices and edges.

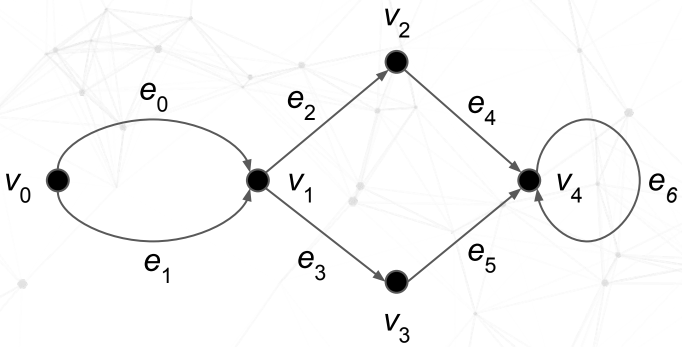
- L_0: $\mathcal{V} \rightarrow$ {vertice labels}
- L_1: $\mathcal{E} \rightarrow$ {edge labels}

dida

Representing graphs in Python

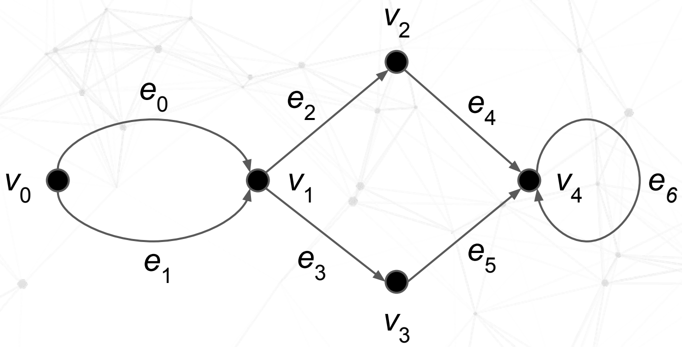


Representing graphs in Python




```
adjacencies = [
    (0, 1),  # e_0
    (0, 1),  # e_1
    (1, 2),  # e_2
    ...,
    (4, 4),  # e_6
]
```

Representing graphs in Python




```
adjacencies = [
    (0, 1),  # e_0
    (0, 1),  # e_1
    (1, 2),  # e_2
    ...,
    (4, 4),  # e_6
]
```

A graph with nodes labeled by NodeT and edges labeled by EdgeT can be modeled as follows:

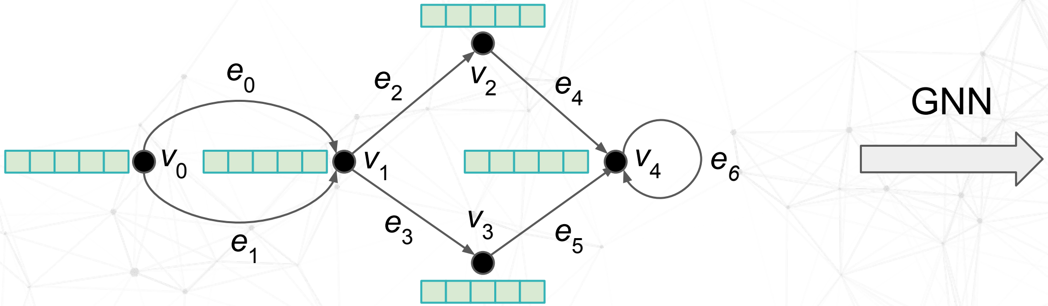
```
class Graph(Generic[NodeT, EdgeT]):

    nodes: List[NodeT]
    edges: List[EdgeT]
    adjacencies: List[Tuple[int, int]]

    def source(self, k: int) -> int:
        """Source of the kth edge"""
        return self.adjacencies[k][0]

    def target(self, k: int) -> int:
        """Target of the kth edge"""
        return self.adjacencies[k][1]
```

dida

Graph neural networks

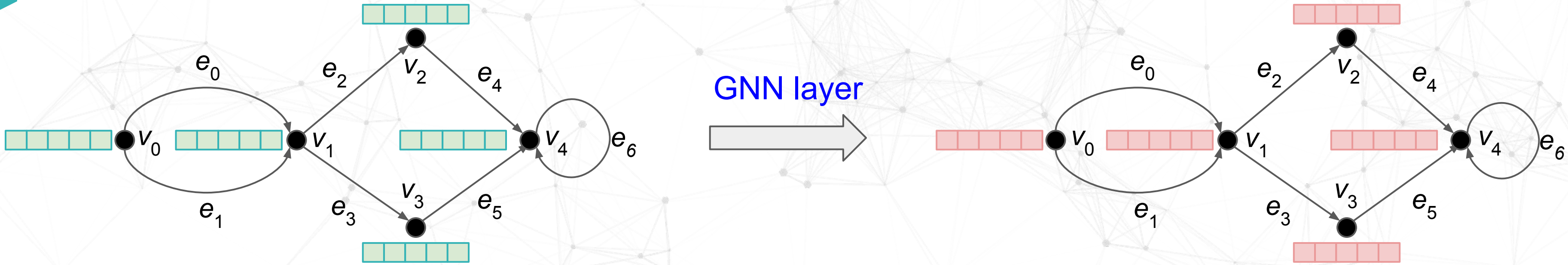
Let's consider the *node classification* problem:

- Input: a graph G of type `Graph[Tensor, EnumT]`
- Output: a labeling of its vertices, $L \colon \mathcal{V} \to \{0, 1\}$




```
labels = [
    0,  # v_0
    1,  # v_1
    0,  # v_2
    0,  # v_3
    1   # v_4
]
```

Graph neural networks



For instance: $y_v = \sigma\left(\sum_{w \xrightarrow{e} v} \frac{1}{\deg^{+}_{t(e)}(v)} W_{t(e)} x_w\right)$ where

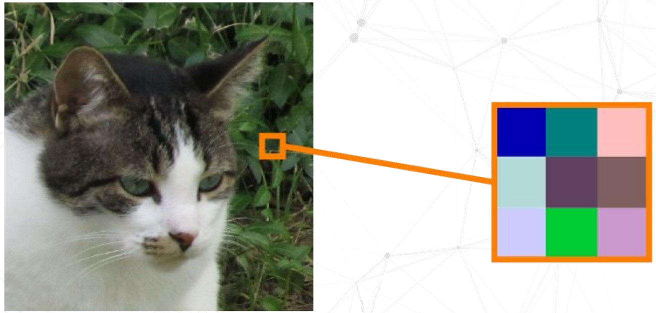
- y_v is the output feature vector of node v
- x_w is the input feature vector of node w
- $t(e)$ is the type of the edge $e: v \rightarrow w$
- W_t is a learned weight matrix corresponding to edge type t
- $\deg_t^{+}(v)$ is the number of incoming edges of type t for node v
- σ is an activation function

Schlichtkrull et at, “Modeling Relational Data with Graph Convolutional Networks”, arxiv:1703.06103

dida

Comparison with convolutional networks

Graph (convolutional) networks generalize CNNs from computer vision



Original image   A 3×3 patch

$$y_{i,j} = \sigma\left(\sum_{\substack{k=-1,0,1 \\ l=-1,0,1}} W_{k,l}\, x_{i+k,j+l} \right)$$

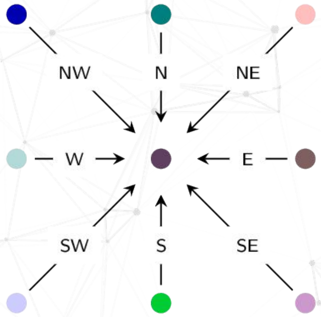
where

- $y_{i,j}$ is the output feature vector of pixel (i, j)
- $x_{i,j}$ is the input feature vector of pixel (i, j)
- $W_{k,l}$ are learned weight matrices
- σ is an activation function

dida

Comparison with convolutional networks

Graph (convolutional) networks generalize CNNs from computer vision



Original image

A graph representing the 3 × 3 patch

$$y_{i,j} = \sigma\left(\sum_{\substack{k=-1,0,1\\ l=-1,0,1}} W_{k,l}\, x_{i+k,j+l}\right)$$

where

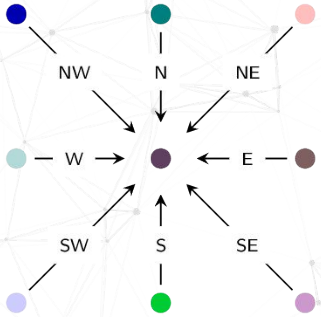
- $y_{i,j}$ is the output feature vector of pixel (i, j)
- $x_{i,j}$ is the input feature vector of pixel (i, j)
- $W_{k,l}$ are learned weight matrices
- σ is an activation function

dida

Comparison with convolutional networks

Graph (convolutional) networks generalize CNNs from computer vision

Original image



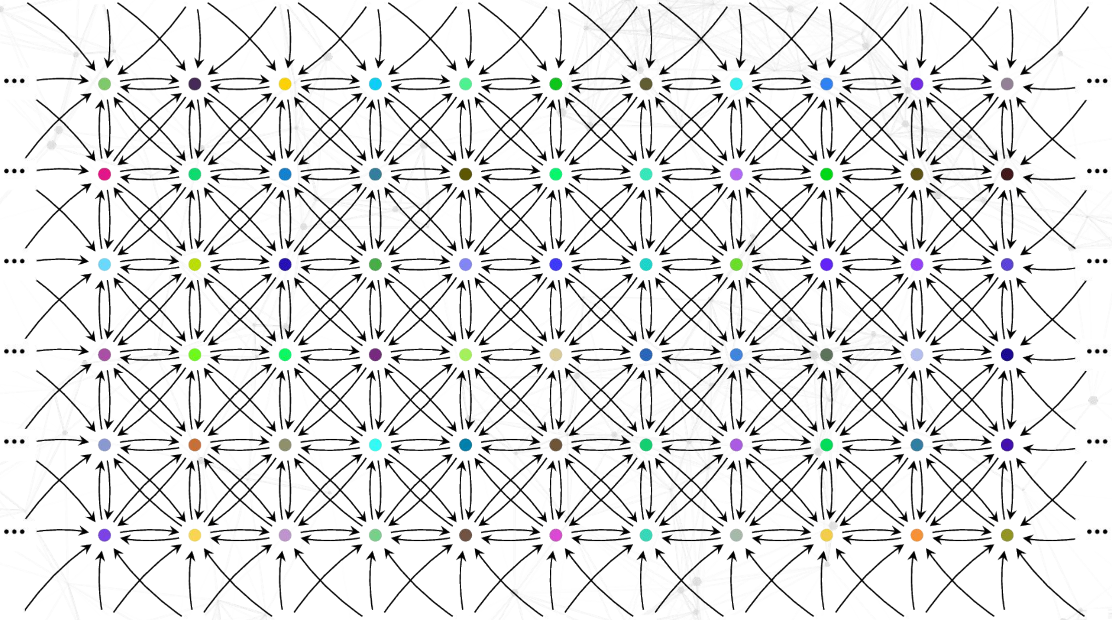
A graph representing the 3 × 3 patch

$$y_p = \sigma\left(W_0 x_p + \sum_{q \xrightarrow{t} p} W_t x_q \right)$$

where

- x_p is the input feature vector of pixel p
- y_p is the output feature vector of pixel p
- $t \in \{N, NW, W, \ldots\}$ is a cardinal direction
- W_t are learned weight matrices
- σ is an activation function

Comparison with convolutional networks

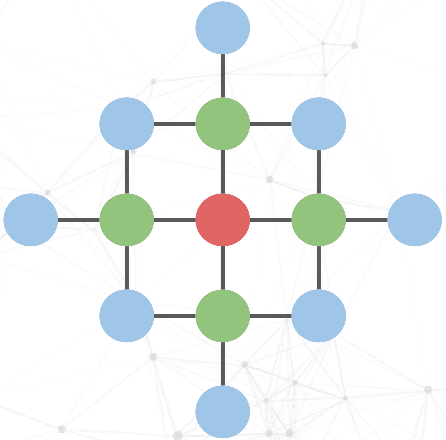


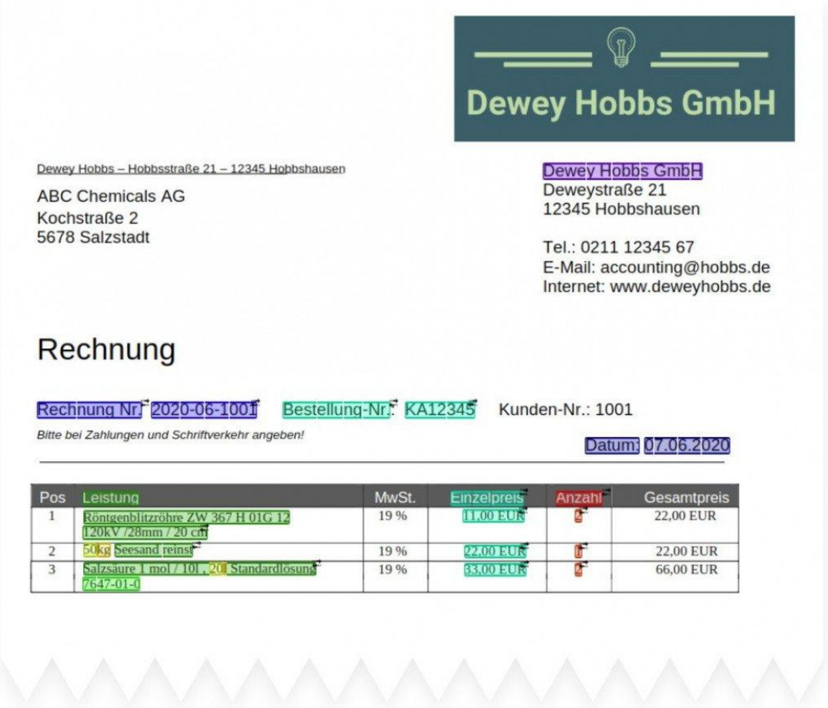
A graph representing the entire image

Use case: information extraction from tables

Idea: Model the document structure as a graph

- each word is a node in the graph
- neighbouring words are connected by edges





Dewey Hobbs GmbH

Dewey Hobbs – Hobbsstraße 21 – 12345 Hobbshausen

ABC Chemicals AG
Kochstraße 2
5678 Salzstadt

Dewey Hobbs GmbH
Deweystraße 21
12345 Hobbshausen

Tel.: 0211 12345 67
E-Mail: accounting@hobbs.de
Internet: www.deweyhobbs.de

Rechnung

Rechnung Nr.: 2020-06-1001 Bestellung-Nr.: KA12345 Kunden-Nr.: 1001

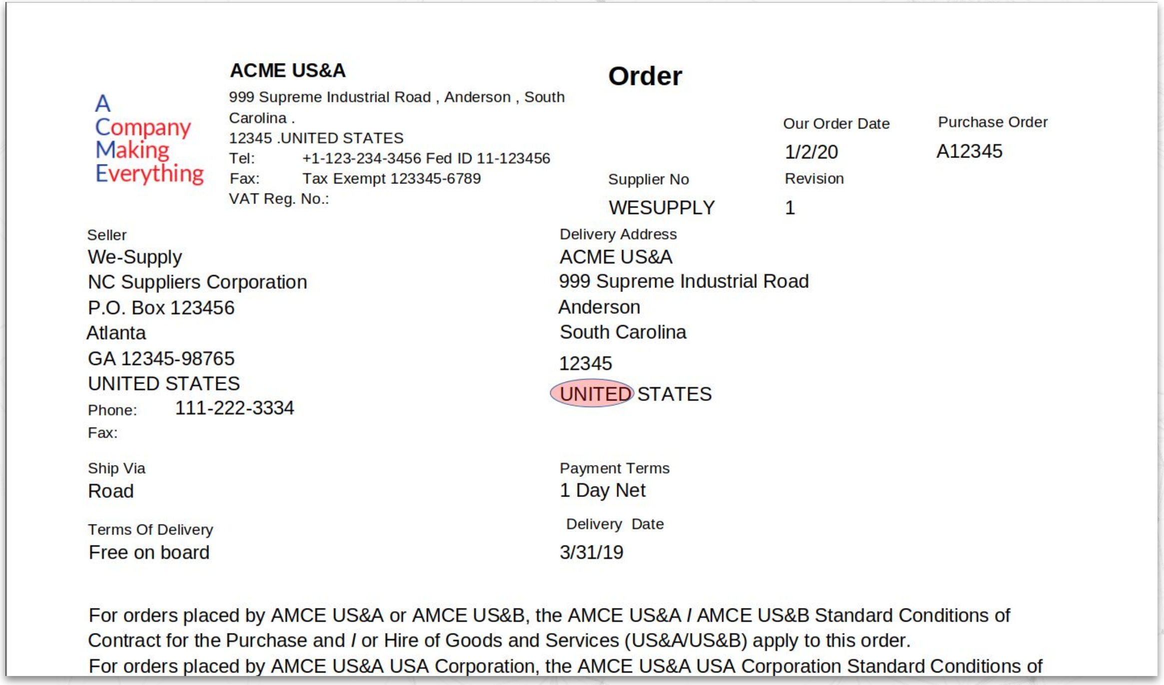
Bitte bei Zahlungen und Schriftverkehr angeben!

Datum: 07.06.2020

Pos	Leistung	MwSt.	Einzelpreis	Anzahl	Gesamtpreis
1	Röntgenblitzröhre ZW 367 H 01G 12 120kV /28mm / 20 cm	19 %	11,00 EUR	2	22,00 EUR
2	50kg Seesand reinst	19 %	22,00 EUR	1	22,00 EUR
3	Salzsäure 1 mol / 10l , 20l Standardlösung 7647-01-0	19 %	33,00 EUR	2	66,00 EUR

Prediction Result (example)

Use case: information extraction from tables



A Company Making Everything

ACME US&A

999 Supreme Industrial Road , Anderson , South Carolina .
12345 .UNITED STATES
Tel: +1-123-234-3456 Fed ID 11-123456
Fax: Tax Exempt 123345-6789
VAT Reg. No.:

Order

Our Order Date
1/2/20

Purchase Order
A12345

Supplier No
WESUPPLY

Revision
1

Seller
We-Supply
NC Suppliers Corporation
P.O. Box 123456
Atlanta
GA 12345-98765
UNITED STATES
Phone: 111-222-3334
Fax:

Delivery Address
ACME US&A
999 Supreme Industrial Road
Anderson
South Carolina
12345
UNITED STATES

Ship Via
Road

Payment Terms
1 Day Net

Terms Of Delivery
Free on board

Delivery Date
3/31/19

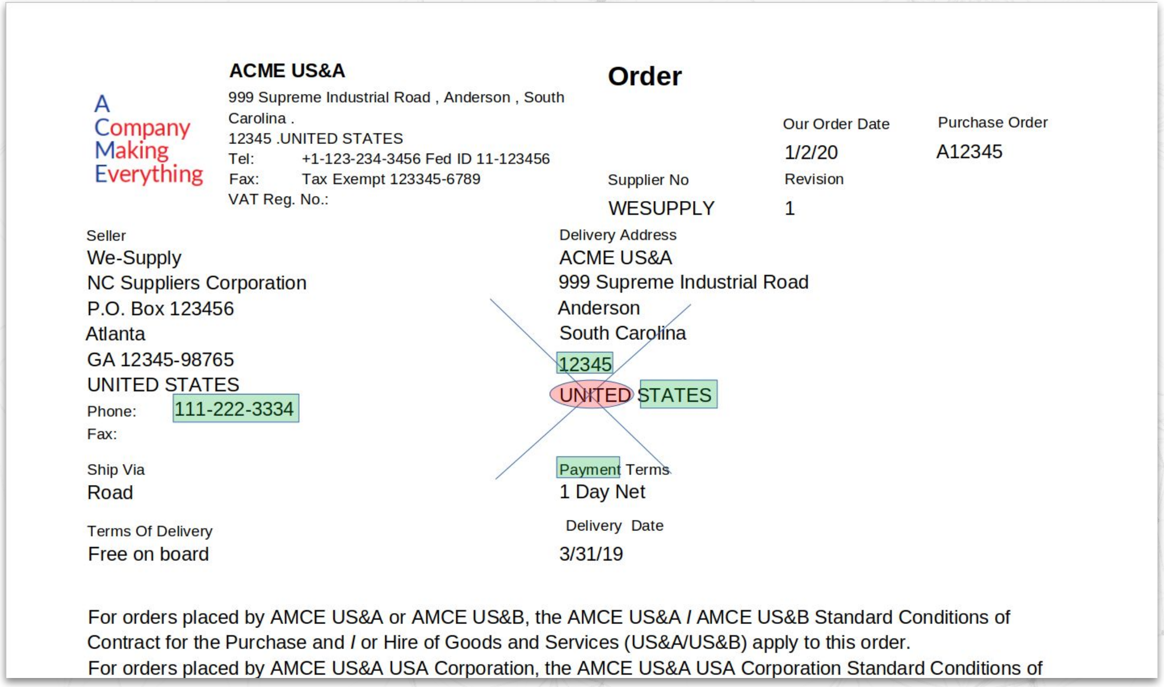
For orders placed by AMCE US&A or AMCE US&B, the AMCE US&A / AMCE US&B Standard Conditions of Contract for the Purchase and / or Hire of Goods and Services (US&A/US&B) apply to this order.
For orders placed by AMCE US&A USA Corporation, the AMCE US&A USA Corporation Standard Conditions of

Nodes:

- Each word is a node in the graph

dida

Use case: information extraction from tables



A Company Making Everything

ACME US&A

999 Supreme Industrial Road , Anderson , South Carolina .
12345 .UNITED STATES
Tel: +1-123-234-3456 Fed ID 11-123456
Fax: Tax Exempt 123345-6789
VAT Reg. No.:

Order

Our Order Date: 1/2/20
Purchase Order: A12345
Supplier No: WESUPPLY
Revision: 1

Seller
We-Supply
NC Suppliers Corporation
P.O. Box 123456
Atlanta
GA 12345-98765
UNITED STATES
Phone: 111-222-3334
Fax:

Delivery Address
ACME US&A
999 Supreme Industrial Road
Anderson
South Carolina
12345
UNITED STATES

Ship Via
Road

Payment Terms
1 Day Net

Terms Of Delivery
Free on board

Delivery Date
3/31/19

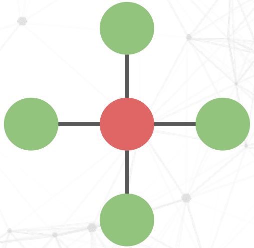
For orders placed by AMCE US&A or AMCE US&B, the AMCE US&A / AMCE US&B Standard Conditions of Contract for the Purchase and / or Hire of Goods and Services (US&A/US&B) apply to this order.
For orders placed by AMCE US&A USA Corporation, the AMCE US&A USA Corporation Standard Conditions of

Nodes:

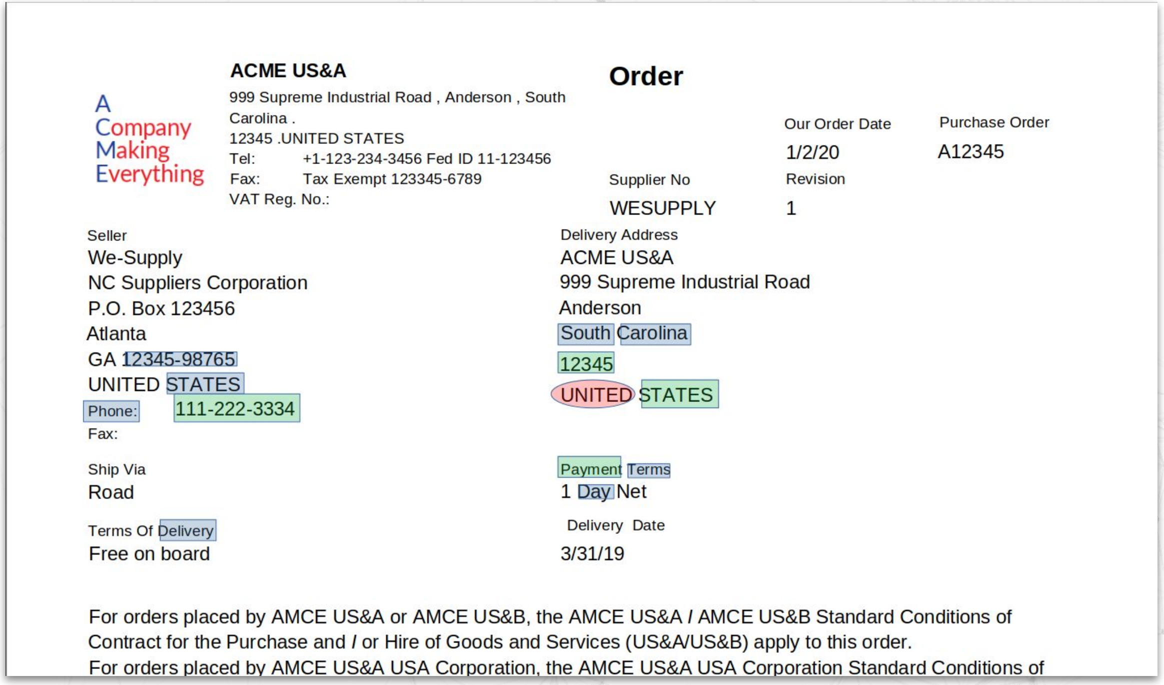
- Each word is a node in the graph

Edges:

- Neighboring words are connected



Use case: information extraction from tables



ACME US&A

Order

A Company Making Everything

999 Supreme Industrial Road , Anderson , South Carolina .
12345 .UNITED STATES
Tel: +1-123-234-3456 Fed ID 11-123456
Fax: Tax Exempt 123345-6789
VAT Reg. No.:

Our Order Date
1/2/20

Purchase Order
A12345

Supplier No
WESUPPLY

Revision
1

Seller
We-Supply
NC Suppliers Corporation
P.O. Box 123456
Atlanta
GA 12345-98765
UNITED STATES
Phone: 111-222-3334
Fax:

Delivery Address
ACME US&A
999 Supreme Industrial Road
Anderson
South Carolina
12345
UNITED STATES

Ship Via
Road

Payment Terms
1 Day Net

Terms Of Delivery
Free on board

Delivery Date
3/31/19

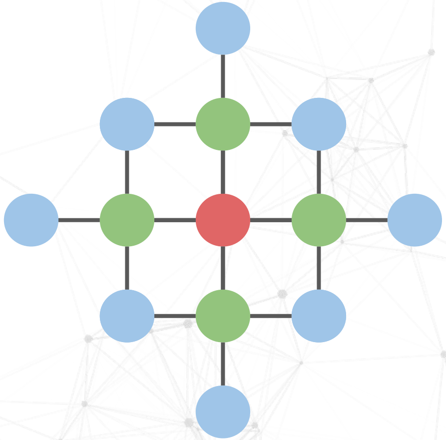
For orders placed by AMCE US&A or AMCE US&B, the AMCE US&A / AMCE US&B Standard Conditions of Contract for the Purchase and / or Hire of Goods and Services (US&A/US&B) apply to this order.
For orders placed by AMCE US&A USA Corporation, the AMCE US&A USA Corporation Standard Conditions of

Nodes:

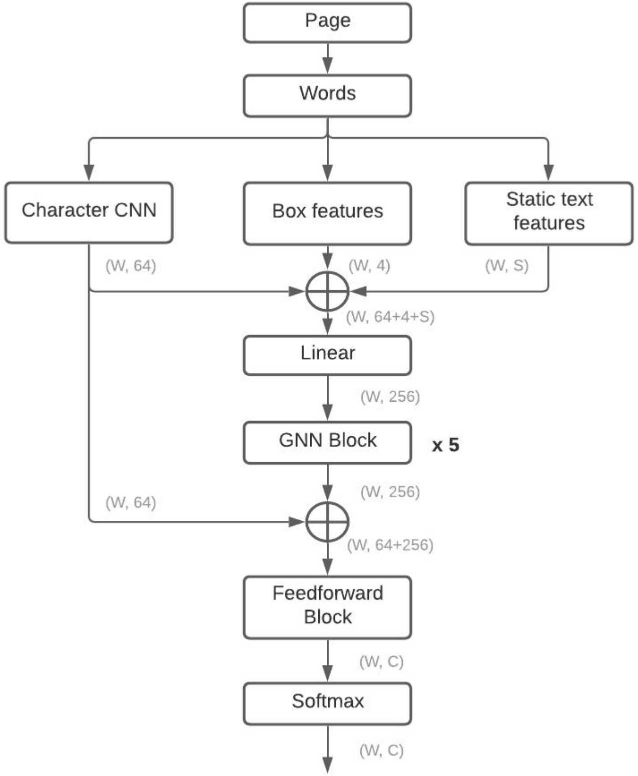
- Each word is a node in the graph

Edges:

- Neighboring words are connected



Use case: a possible model architecture



Legend

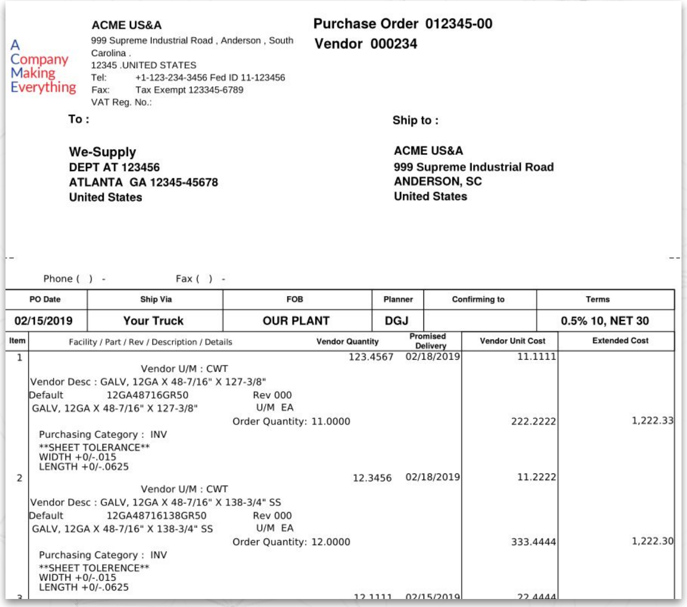
W: Number of words

S: Number of static text features

C: Number of classes

⊕: Concatenate tensors

Use case: results


A Company Making Everything

ACME US&A
999 Supreme Industrial Road , Anderson , South Carolina .
12345 .UNITED STATES
Tel: +1-123-234-3456 Fed ID 11-123456
Fax: Tax Exempt 123345-6789
VAT Reg. No.:

Purchase Order 012345-00
Vendor 000234

To :

We-Supply
DEPT AT 123456
ATLANTA GA 12345-45678
United States

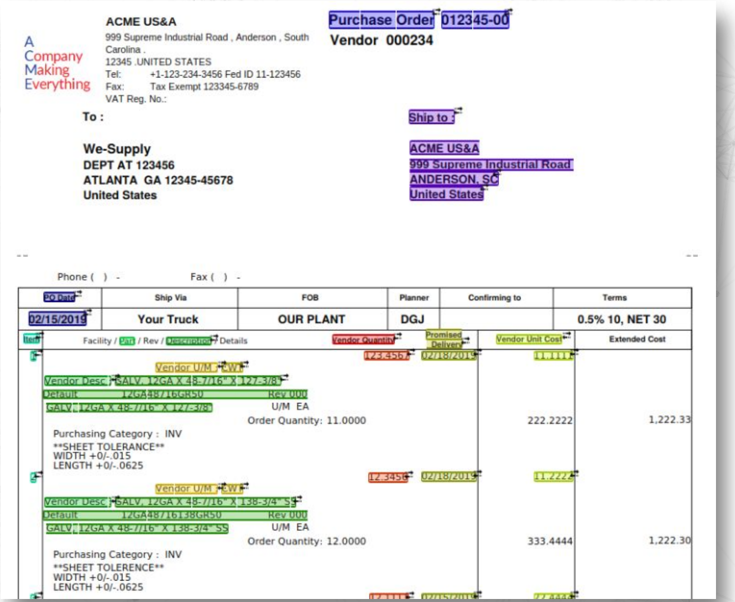
Ship to :

ACME US&A
999 Supreme Industrial Road
ANDERSON, SC
United States

Phone ( ) - Fax ( ) -

PO Date	Ship Via	FOB	Planner	Confirming to	Terms
02/15/2019	Your Truck	OUR PLANT	DGJ		0.5% 10, NET 30

Item	Facility / Part / Rev / Description / Details	Vendor Quantity	Promised Delivery	Vendor Unit Cost	Extended Cost
1		123.4567	02/18/2019	11.1111	
	Vendor U/M : CWT Vendor Desc : GALV, 12GA X 48-7/16" X 127-3/8" Default 12GA48716GR50 Rev 000 GALV, 12GA X 48-7/16" X 127-3/8" U/M EA Order Quantity: 11.0000			222.2222	1,222.33
	Purchasing Category : INV **SHEET TOLERANCE** WIDTH +0/-.015 LENGTH +0/-.0625				
2		12.3456	02/18/2019	11.2222	
	Vendor U/M : CWT Vendor Desc : GALV, 12GA X 48-7/16" X 138-3/4" SS Default 12GA48716138GR50 Rev 000 GALV, 12GA X 48-7/16" X 138-3/4" SS U/M EA Order Quantity: 12.0000			333.4444	1,222.30
	Purchasing Category : INV **SHEET TOLERENCE** WIDTH +0/-.015 LENGTH +0/-.0625				
3		12.1111	02/15/2019	22.4444	


A Company Making Everything

ACME US&A
999 Supreme Industrial Road , Anderson , South Carolina .
12345 .UNITED STATES
Tel: +1-123-234-3456 Fed ID 11-123456
Fax: Tax Exempt 123345-6789
VAT Reg. No.:

Purchase Order 012345-00
Vendor 000234

To :

We-Supply
DEPT AT 123456
ATLANTA GA 12345-45678
United States

Ship to :

ACME US&A
999 Supreme Industrial Road
ANDERSON, SC
United States

Phone ( ) - Fax ( ) -

PO Date	Ship Via	FOB	Planner	Confirming to	Terms
02/15/2019	Your Truck	OUR PLANT	DGJ		0.5% 10, NET 30

Item	Facility / Part / Rev / Description / Details	Vendor Quantity	Promised Delivery	Vendor Unit Cost	Extended Cost
1		123.4567	02/18/2019	11.1111	
	Vendor U/M : CWT Vendor Desc : GALV, 12GA X 48-7/16" X 127-3/8" Default 12GA48716GR50 Rev 000 GALV, 12GA X 48-7/16" X 127-3/8" U/M EA Order Quantity: 11.0000			222.2222	1,222.33
	Purchasing Category : INV **SHEET TOLERANCE** WIDTH +0/-.015 LENGTH +0/-.0625				
2		12.3456	02/18/2019	11.2222	
	Vendor U/M : CWT Vendor Desc : GALV, 12GA X 48-7/16" X 138-3/4" SS Default 12GA48716138GR50 Rev 000 GALV, 12GA X 48-7/16" X 138-3/4" SS U/M EA Order Quantity: 12.0000			333.4444	1,222.30
	Purchasing Category : INV **SHEET TOLERENCE** WIDTH +0/-.015 LENGTH +0/-.0625				

About 1000 training documents

Macro F1 score: 88%

dida

Implementations, literature

- PyTorch Geometric
    - "GNN Cheatsheet" in the docs contains an interesting list of papers
- Spektral for TensorFlow
- Deep Graph Library (framework agnostic)
- https://en.wikipedia.org/wiki/Graph_neural_network exists since 5 July 2021

dida