Representation Learning - A Case Study

Magdalena Wiercioch

Faculty of Physics, Astronomy and Applied Computer Science & Faculty of Mathematics and Computer Science
Jagiellonian University
e-mail: *magdalena.wiercioch@uj.edu.pl*

58. Cracow School of Theoretical Physics

Outline

Problem

"When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one."

Problem

"When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one."

- Discriminative models are better than generative models.
- It was intended to justify trunsduction - label the unlabeled point only, not a full model.
- It neglected the structure of the data.

Problem

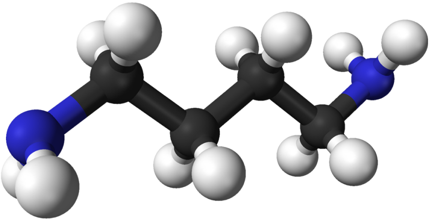
How do humans learn actually?

Problem

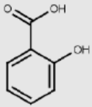
- Kids can recognize their parents.
- They cannot paint them yet.
- This model is far more useful than a perfect reproduction.

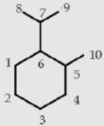
Attributes of representation

- Useful - It can do an inference on what is possible.
- Simple - it skips a lot of details.
- Accurate enough - it enables to recognize and identify novelty.



Neighbours matrix

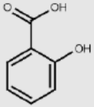


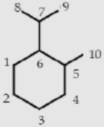


n	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	1	0	0	0	0
2		0	1	0	0	0	0	0	0	0
3			0	1	0	0	0	0	0	0
4				0	1	0	0	0	0	0
5					0	1	0	0	0	1
6						0	1	0	0	0
7							0	1	1	0
8								0	0	0
9									0	0
10										0

$$[A(G)]_{ij} = \begin{cases} 1; & i \neq j, \ e_{ij} \in E(G) \\ 0; & i = j, \ e_{ij} \notin E(G) \end{cases}$$

Distance matrix

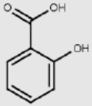


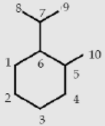


n	1	2	3	4	5	6	7	8	9	10
1	0	1	2	3	2	1	2	3	3	3
2		0	1	2	3	2	3	4	4	4
3			0	1	2	3	4	5	5	3
4				0	1	2	3	4	4	2
5					0	1	2	3	3	1
6						0	1	2	2	2
7							0	1	1	3
8								0	2	4
9									0	4
10										0

$$[D(G)]_{ij} = \begin{cases} d_{ij} \; ; \; i \neq j \\ 0 \; ; \; i = j \end{cases}$$

Connections matrix

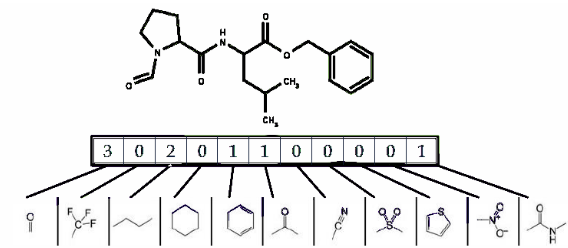




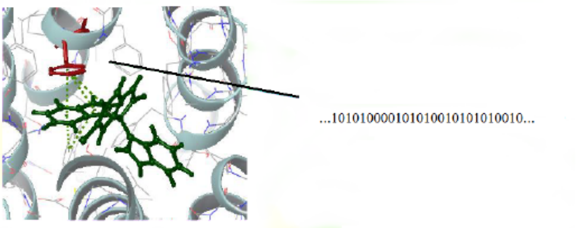
n	1	2	3	4	5	6	7	8	9	10
1	0	2				1				
2	2	0	1							
3		1	0	2						
4			2	0	1					
5				1	0	2				1
6	1				2	0	1			
7						1	0	1	2	
8							1	0	1	
9							2		0	
10					1					0

$$[B(G)]_{ij} = \begin{cases} b_{ij}\ ;\ i \neq j \\ 0\ ;\ i = j \end{cases}$$

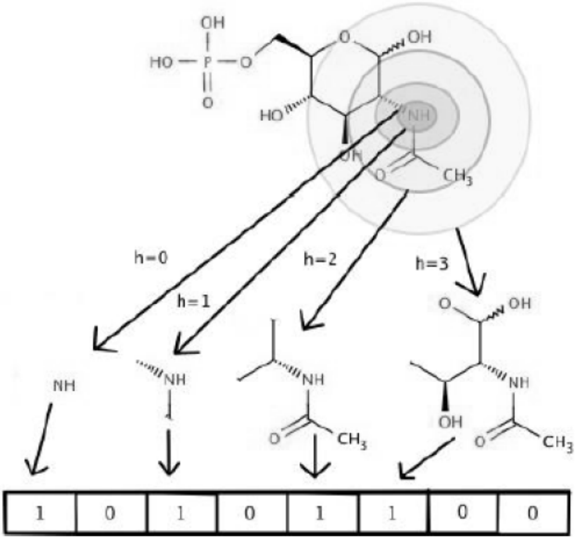
Substructural fingerprint



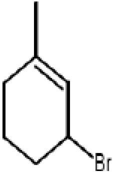
Structural interaction fingerprint



Hashed fingerprint

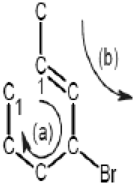


SMILES



(a) CC1=CC(Br)CCC1

(b) CC1=CC(CCC1)Br



SDF

```
Mrv0541 01151319442D

 10 10  0  0  0  0            999 V2000
   -1.6721   -0.1120    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.9576   -0.5245    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.9576   -1.3495    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.6721   -1.7620    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.3866   -1.3495    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.3866   -0.5245    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.6721    0.7130    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.3866    1.1255    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.9576    1.1255    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.2432   -0.1120    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  6  2  0  0  0  0
  2  3  2  0  0  0  0
  3  4  1  0  0  0  0
  4  5  2  0  0  0  0
  5  6  1  0  0  0  0
  1  7  1  0  0  0  0
  7  8  2  0  0  0  0
  7  9  1  0  0  0  0
  2 10  1  0  0  0  0
M  END
>  <CdId>
5

>  <Mol Weight>
138.1207

>  <Formula>
C7H6O3

$$$$
```

header

atomic coordinates

connection table

properties

Moments

- Moments are scalar quantities used to characterize a function and to capture its significant features.

Moments

Definition

General moment $M_{pq}^{(f)}$ of an image f(x, y), gdzie p, q re non-negative integers and $r = p + q$ is called the order of the moment, defined as

$$M_{pq}^{(f)} = \iint p_{pq}(x, y) f(x, y) \mathrm{d}x \mathrm{d}y,$$

where $p_{00}(x, y), p_{10}(x, y), \ldots, p_{kj}(x, y), \ldots$ are polynomial basis functions defined on D.

Moments

Definition

General moment $M_{pq}^{(f)}$ of an image f(x, y), gdzie p, q re non-negative integers and $r = p + q$ is called the order of the moment, defined as

$$M_{pq}^{(f)} = \iint p_{pq}(x, y) f(x, y) \mathrm{d}x \mathrm{d}y,$$

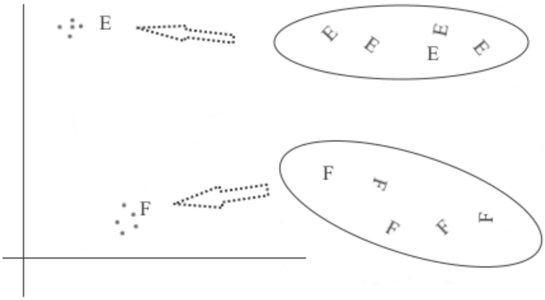
where $p_{00}(x, y), p_{10}(x, y), \ldots, p_{kj}(x, y), \ldots$ are polynomial basis functions defined on D.
Depending on the polynomial basis used, we recognize various systems of moments.

Geometric moments

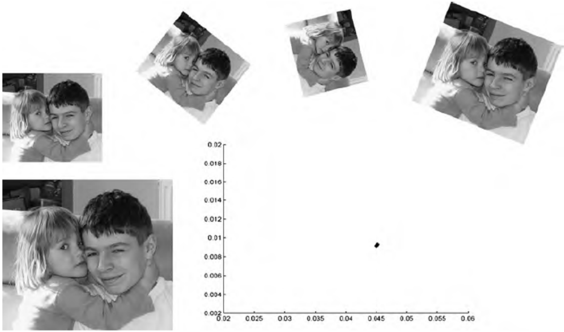
$p_{kj}(x, y) = x^k y^j$ leads to geometric moments.

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x, y) \mathrm{d}x \mathrm{d}y.$$

Vector of invariants (Hu, 1962)



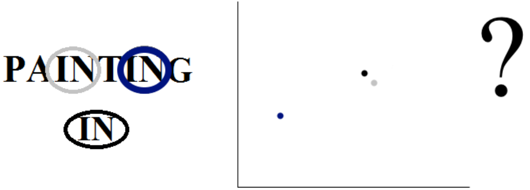
Vector of invariants



Hu invariants - drawbacks

- mutual dependence;
- restriction to second- and third-order moments only.

Research motivation



- Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov: *Enriching Word Vectors with Subword Information*, arXiv, 2016.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean: *Efficient estimation of word representations in vector space*, arXiv, 2013.

Background

- Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman: *Indexing by latent semantic analysis*. Journal of the American Society for Information Science, 1990.
- neural networks
- Hinrich Schütze: *Dimensions of meaning*. Proceedings of the 1992 ACM/IEEE Conference on Supercomputing, 1992.
- N. Sakamoto, K. Yamamoto, and S. Nakagawa: *Combination of syllable based n-gram search and word search for spoken term detection through spoken queries and iv/oov classification*. Automatic Speech Recognition and Understanding (ASRU), 2015.

Model

- We extended the method introduced by Bojanowski et al.
- The model demonstrated by Bojanowski is derived from continuous Skip-gram (SG) model proposed by Mikolov et al.

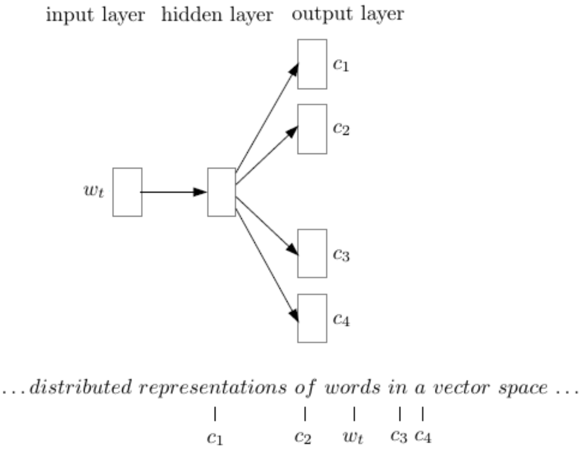
Skip-gram model

- The goal of Skip-gram model is to find word representation that is useful for predicting the surrounding words in a corpus.
- Let us denote the sequence of training words - vocabulary, $W = \{w_1, w_2, \ldots, w_S\}$, where S is the size of vocabulary.
- Skip-gram model maximizes the average log probability

$$l(W) = \sum_{t=1}^{S} \sum_{c \in C_t} \log p(w_c | w_t),$$

where C_t is the context.

Skip-gram model



Skip-gram model

- The probability of observing a context word w_c given w_t is parametrized using the word vectors.
- Given a scoring function s, which maps pairs of (word, context) to value in $\mathbb{R}$, a possible choice to define the probability of a context word is the softmax.

$$p(Context|Word) = y_c = \frac{e^{w_c^\top w_t}}{\sum_{j=1}^{S} e^{w_j^\top w_t}},$$

  where w_c, w_t, w_j are vector representations of words and y_c is the output of the c-th neuron of the output layer.
- The parametrization for the scoring function is done by taking the scalar product between word and context embeddings: $s(Word, Context) = w_t^\top w_c$.

Subword model by Bojanowski et al.

- The Skip-gram model ignores the internal structure of words.
- They introduced a different scoring function s

$$s(w, c) = \sum_{g \in G_w} z_g^\top v_c,$$

where $G_w = \{1, \ldots, G\}$ is the set of letter n-grams which appear in w.

Subword model by Bojanowski et al.

- The Skip-gram model ignores the internal structure of words.
- They introduced a different scoring function s

  $$s(w, c) = \sum_{g \in G_w} z_g^\top v_c,$$

  where $G_w = \{1, \ldots, G\}$ is the set of letter n-grams which appear in w.
- Limitations:
  - n-grams with a length greater or equal than 3 and smaller or equal than 6 were considered.
  - We claim it may be insufficient for short and rare words.

Fragmentation model

- Let us denote by $G_w = \{1, \ldots, G\}$ the set of letter n-grams which appear in w and $H_w = \{1, \ldots, H\}$ to be the set of syllable n-grams which appear in w.

Fragmentation model

- Let us denote by $G_w = \{1, \dots, G\}$ the set of letter n-grams which appear in w and $H_w = \{1, \dots, H\}$ to be the set of syllable n-grams which appear in w.
- We associate a vector representation z_g to each letter n-gram g and a vector representation z_h to each syllable n-gram h.

Fragmentation model

- Let us denote by $G_w = \{1, \ldots, G\}$ the set of letter n-grams which appear in w and $H_w = \{1, \ldots, H\}$ to be the set of syllable n-grams which appear in w.
- We associate a vector representation z_g to each letter n-gram g and a vector representation z_h to each syllable n-gram h.
- The new word representation is considered as the direct concatenation of the two vector representations of its n-grams (letter and syllables)

$$z_{new} = [z_g, z_h].$$

Fragmentation model

- Let us denote by $G_w = \{1, \ldots, G\}$ the set of letter n-grams which appear in w and $H_w = \{1, \ldots, H\}$ to be the set of syllable n-grams which appear in w.
- We associate a vector representation z_g to each letter n-gram g and a vector representation z_h to each syllable n-gram h.
- The new word representation is considered as the direct concatenation of the two vector representations of its n-grams (letter and syllables)

$$z_{new} = [z_g, z_h].$$

- The scoring function is

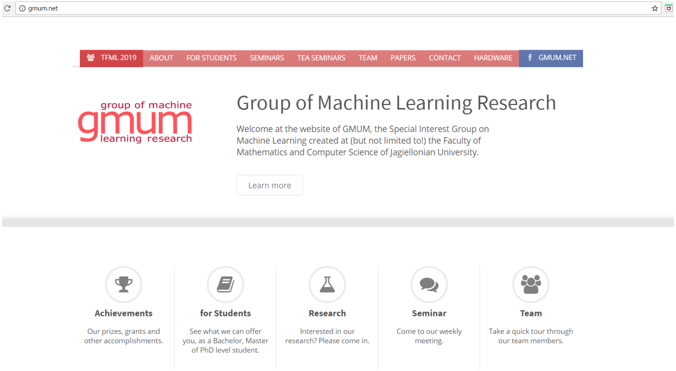
$$s(w, c) = \sum_{new \in G_w \cup H_w} z_{new}^{\top} v_c.$$

- The upgraded model makes use of n-grams of varied length n.

Conclusions

- Invariance based methods are being tested in order to be applied to objects described by many features.
- The Skip-Gram based method outperforms state-of-the-art approaches on dense languages when tasks such as word similarity ranking or syntactic and semantic analogies are taken into consideration.

...


gmum.net
TFML 2019
ABOUT
FOR STUDENTS
SEMINARS
TEA SEMINARS
TEAM
PAPERS
CONTACT
HARDWARE
GMUM.NET
group of machine
gmum
learning research
Group of Machine Learning Research
Welcome at the website of GMUM, the Special Interest Group on Machine Learning created at (but not limited to!) the Faculty of Mathematics and Computer Science of Jagiellonian University.
Learn more
Achievements
Our prizes, grants and other accomplishments.
for Students
See what we can offer you, as a Bachelor, Master of PhD level student.
Research
Interested in our research? Please come in.
Seminar
Come to our weekly meeting.
Team
Take a quick tour through our team members.

Thank you.