

Classification of cancer pathology reports with Deep Learning methods

Stefano MARTINA
stefano.martina@unifi.it



UNIVERSITÀ
DEGLI STUDI
FIRENZE

6 November 2019



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).

Overview

- 1 Introduction
 - Cancer registries
 - ICD-O
- 2 Machine Learning
 - Representations
 - Classic models
 - RNN
 - Attention Models
- 3 Scientific questions
- 4 Materials and Methods
 - Datasets
 - Models
- 5 Experiments
 - Bag-of-words VS word vectors, SVM VS deep learning
 - Preliminary attention VS max
 - Attention VS max, hierarchical VS plain
- 6 Conclusions



Cancer registries

- ✓ **Collect** administrative and clinical data of a specific region
- ✓ **Quantify** the impact of the disease
- ✓ Provide **analytic** data to healthcare operators and decision makers
- ✓ **Manual** classification of reports



International Classification of Diseases for Oncology (ICD-O-3)

Topographical

C _ _ . _

- ✓ first two digits **site**
- ✓ third digit **subsite**

E.g. C50.2 upper-inner quadrant (2) of breast (50)

Morphological

_ _ _ _ / _

- ✓ first four digits **cell type**
- ✓ fifth digit **behaviour**

E.g. 8140/3 is an adenocarcinoma (adeno 8140; carcinoma 3)

the dog is on the table

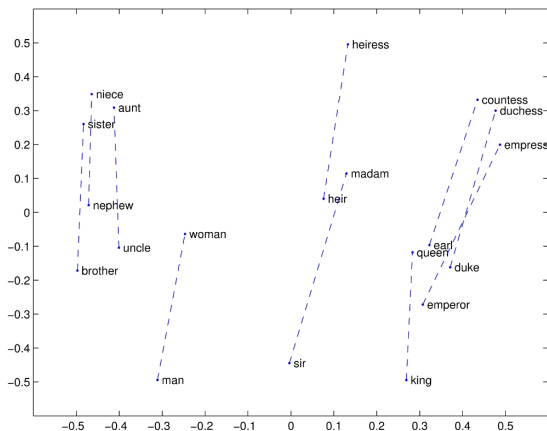
0	0	1	1	0	1	1	1
are	cat	dog	is	now	on	table	the

Term-Frequency Inverse-Document-Frequency (TF-IDF)

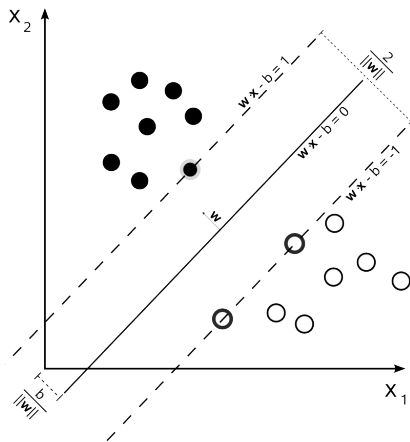
$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

Word vectors

- ✓ Transforms words in vectors
- ✓ Unsupervised learning method
- ✓ Semantic relations encoded in vector space geometric relations



Support Vector Machine (SVM)



Recurrent Neural Network (RNN)

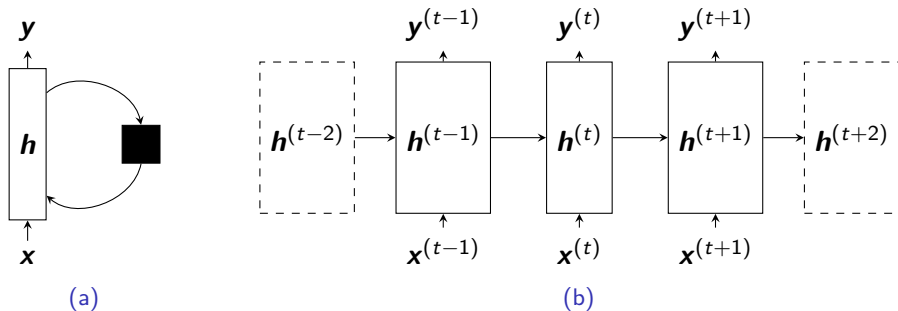
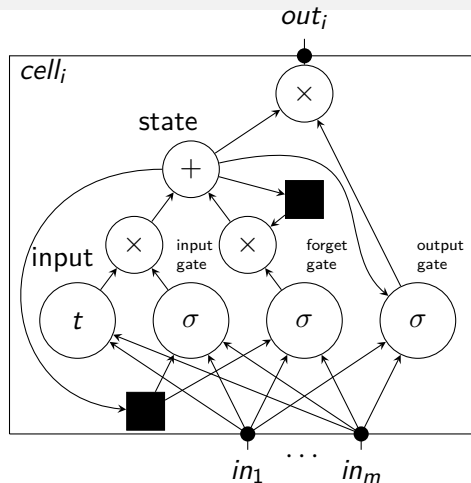


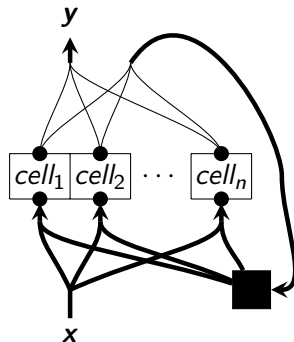
Figure: RNN, folded (a) and unfolded (b) models.

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}; \theta)$$

Long Short-Term Memory (LSTM)



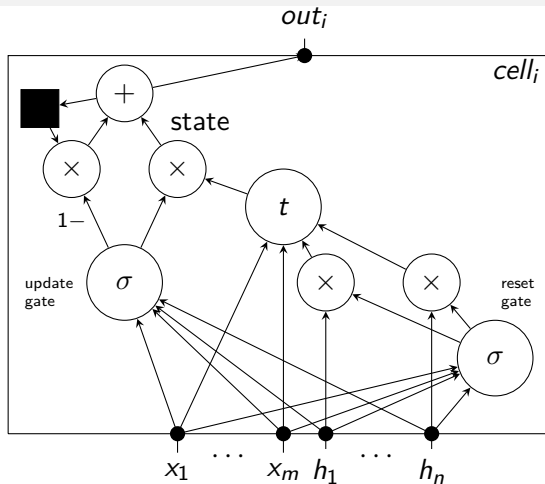
(a)



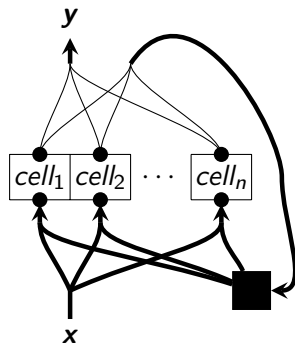
(b)

Figure: memory cell (a), and general scheme (b). The black box is a delay

Gated Recurrent Unit (GRU)



(a)

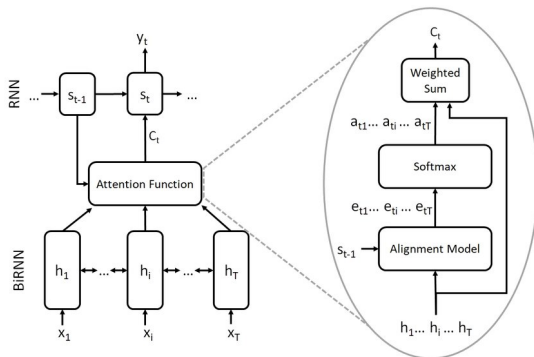


(b)

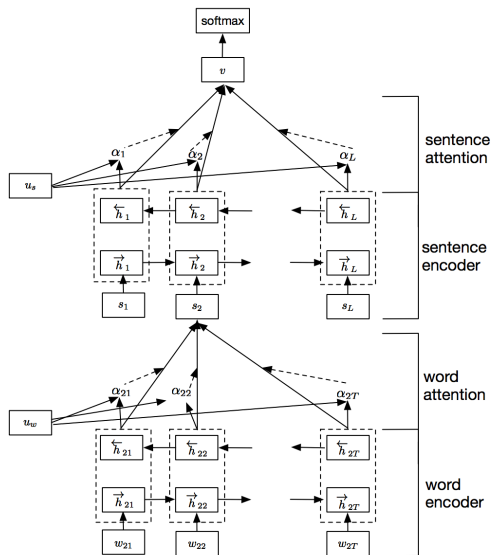
Figure: memory cell (a), and general scheme (b). The black box is a delay

Attention models

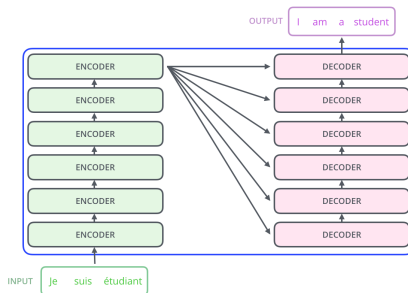
- ✓ Developed for seq-to-seq task
- ✓ State of the art in machine translation



Hierarchical Attention Network (HAN)



Bidirectional Encoder Representations from Transformers (BERT)

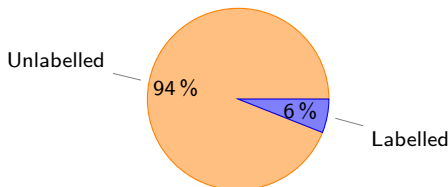


- ✓ State of the art in many NLP tasks
- ✓ Attention based
- ✓ Learn Context dependent word representations
- ✓ Pretrained on unlabeled data
- ✓ Fine tuned to specific task

Scientific questions

- Q1 Implement **large scale** study on deep learning applied to pathology reports, existing works are on
 - ✓ **small** datasets or
 - ✓ **few** classes
- Q2 Apply **novel** deep learning techniques, like **attention** models and **BERT**
- Q3 **Compare** classical **bag-of-words** techniques with newer deep learning techniques in this domain
- Q4 **Compare** novel **attention**-based and **hierarchical** techniques with simpler models
- Q5 **Investigate** the contribution and applicability of **unsupervised** learning techniques on uncommon text corpora
- Q6 **Investigate** the possibility to give **interpretation** to deep learning models

Dataset



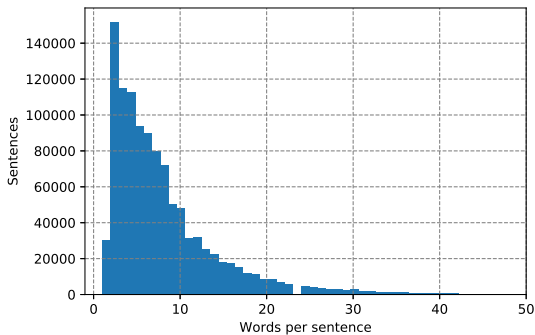
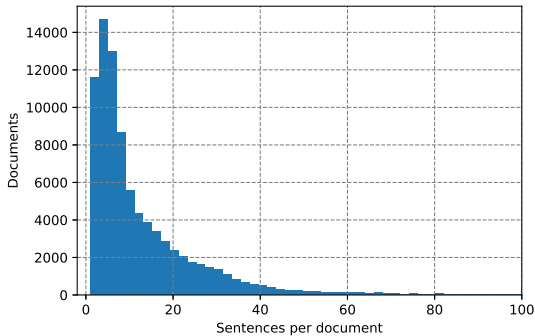
- ✓ 1 592 385 anatomopathological exam results
 - ▶ From **Tuscany** cancer registry
 - ▶ In period **2004-2013**
- ✓ 94 524 (6%) labeled

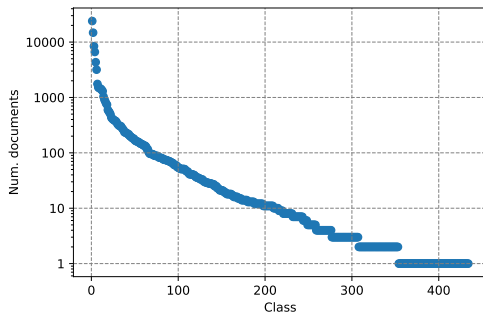
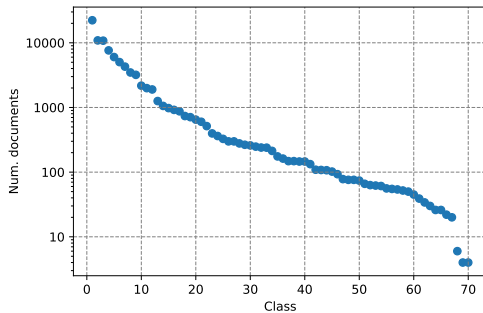
Structure

- ✓ 3 text **fields**: macroscopy, diagnosis, anamnesis
- ✓ field **length** from 0 to 1368 (quartiles 34, 62, 134)

Preparation

- ✓ Data comes in **two** tables to merge:
 1. neoplasm table, containing administrative and clinical variables
 2. histology table, containing the text fields
 - ▶ there are neoplasms without histology associated
 - ★ (register have access to more data)
 - ▶ there are histologies without neoplasm associated
 - ★ (not tumor biopsies)
- ✓ The 3 text fields are **merged**





Models

U-SVM SVM trained on TF-IDF representations using unigrams

B-SVM SVM trained on TF-IDF using unigrams and bigrams

B-XGB XGBoost trained on TF-IDF using unigrams and bigrams

B-LSTM LSTM trained on TF-IDF using bigrams

G-CRNN mixed convolutional and LSTM trained on GloVe

G-LSTM LSTM trained on GloVe

G-GRU GRU trained on GloVe

G-ATT GRU with attention trained on GloVe

G-ATT_h hierarchical GRU with attention trained on GloVe

BERT pretrained on unlabeled data and fine tuned with labeled data

G-MAX GRU with max pooling trained on GloVe

G-MAX_h hierarchical GRU with max pooling trained on GloVe

G-MAX_i GRU with max pooling, in interpretable setting, trained on GloVe

G-ATT_i GRU with attention, in interpretable setting, trained on GloVe

B-LSTM, G-CRNN, G-LSTM

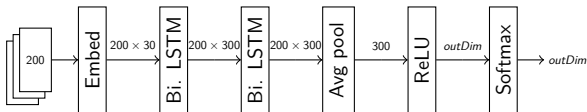


Figure: Scheme for **B-LSTM** model.

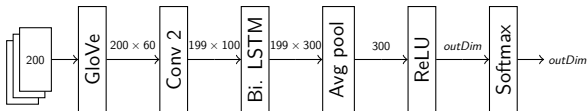


Figure: Scheme for **G-CRNN** model.

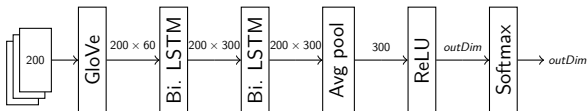


Figure: Scheme for **G-LSTM** model.

Plain model

$$e_t = E(x_t; \theta^e)$$

$$h_t^f = F(e_t, h_{t-1}^f; \theta^f)$$

$$h_t^r = R(e_t, h_{t+1}^r; \theta^r)$$

$$u_t = G(h_t; \theta^h)$$

$$\phi = A(\mathbf{u}; \theta^a)$$

$$f(\mathbf{x}) = g(\phi; \theta^c)$$

- ✓ $\phi = (h_T^f, h_1^r)$ (in this case G is the identity function)
- ✓ $\phi = \sum_t a_t(\mathbf{u}; \theta^a) u_t$, $a_t(\mathbf{u}; \theta^a) = \frac{e^{\langle c, c_t \rangle}}{\sum_i e^{\langle c, c_i \rangle}}$, $c_t = C(\mathbf{u}; \theta^a)$
- ✓ $\phi_j = \max_t u_{j,t}$

Interpretable model

$$e_t = E(x_t; \theta^e)$$

$$h_t^f = F(e_t, h_{t-1}^f; \theta^f)$$

$$h_t^r = R(e_t, h_{t+1}^r; \theta^r)$$

$$u_t = G(h_t; \theta^h)$$

$$f(\mathbf{x}) = A(\mathbf{u}; \theta^a)$$

$$\checkmark \quad \phi = \sum_t a_t(\mathbf{u}; \theta^a) u_t, \quad a_t(\mathbf{u}; \theta^a) = \frac{e^{\langle c, c_t \rangle}}{\sum_i e^{\langle c, c_i \rangle}}, \quad c_t = C(\mathbf{u}; \theta^a)$$

$$\checkmark \quad \phi_j = \max_t u_{j,t}$$

Hierarchical model

$$e_{s,t} = E(x_{s,t}; \theta^e)$$

$$h_{s,t}^f = F(e_{s,t}, h_{s,t-1}^f; \theta^f)$$

$$h_{s,t}^r = R(e_{s,t}, h_{s,t+1}^r; \theta^r)$$

$$u_{s,t} = G(h_{s,t}; \theta^h)$$

$$\phi_s = A(\mathbf{u}_s; \theta^a)$$

$$\bar{h}_s^f = \bar{F}(\phi_s, \bar{h}_{s-1}^f; \bar{\theta}^f)$$

$$\bar{h}_s^r = \bar{R}(\phi_s, \bar{h}_{s+1}^r; \bar{\theta}^r)$$

$$\bar{\phi} = \bar{A}(\bar{\mathbf{h}}; \bar{\theta}^a)$$

$$f(\mathbf{x}) = g(\bar{\phi}; \theta^c)$$

Bag-of-words VS word vectors, SVM VS deep learning

Answered questions

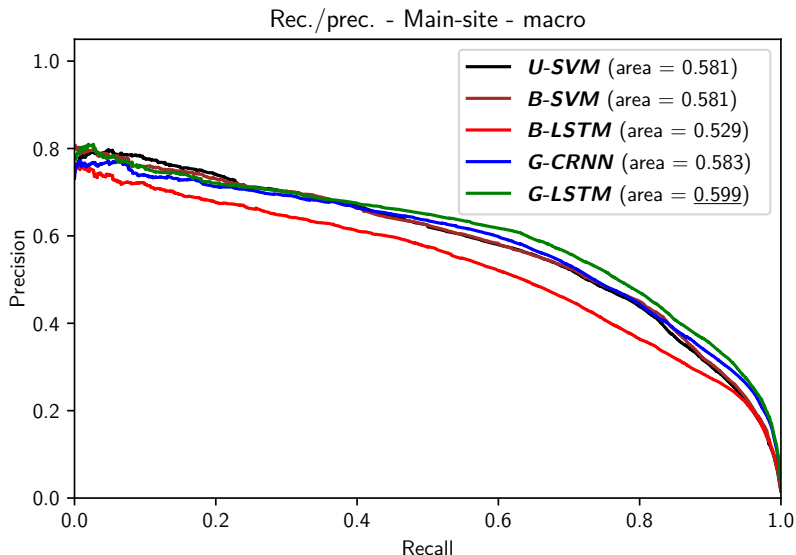
- Q1 Implement **large scale** study on deep learning applied to pathology reports
 - Q3 **Compare** classical **bag-of-words** techniques with newer deep learning techniques in this domain
 - Q5 **Investigate** the contribution and applicability of **unsupervised** learning techniques on uncommon text corpora
-
- ✓ **10-fold** cross validation
 - ✓ All tasks (**main** site, **subsite**, **type**, **behavior**)

Bag-of-words VS word vectors, SVM VS deep learning

Table: Results for Main-site task.

		<i>U-SVM</i>	<i>B-SVM</i>	<i>B-LSTM</i>	<i>G-CRNN</i>	<i>G-LSTM</i>
accuracy		89.8 ± 2.0	89.8 ± 2.0	88.6 ± 2.0	90.0 ± 1.6	<u>90.5</u> ± 1.6
kappa		88.5 ± 2.2	88.6 ± 2.3	87.2 ± 2.3	88.9 ± 1.8	<u>89.3</u> ± 1.8
MAPs		93.0 ± 1.5	93.0 ± 1.5	92.2 ± 1.5	93.5 ± 1.2	<u>93.8</u> ± 1.1
MAPc		61.6 ± 3.9	61.3 ± 4.0	55.7 ± 3.7	62.7 ± 3.5	<u>64.1</u> ± 4.1
pre.	ma.	<u>65.5</u> ± 4.8	64.7 ± 3.2	55.0 ± 2.8	61.5 ± 3.4	61.8 ± 3.7
	we.	88.7 ± 2.0	88.8 ± 2.0	87.8 ± 1.8	89.2 ± 1.6	<u>89.5</u> ± 1.7
rec.	ma.	55.7 ± 4.1	54.7 ± 3.8	51.6 ± 3.2	56.5 ± 3.0	<u>58.1</u> ± 3.5
	we.	89.8 ± 2.0	89.8 ± 2.0	88.6 ± 2.0	90.0 ± 1.6	<u>90.5</u> ± 1.6
f1s.	ma.	<u>58.4</u> ± 4.1	57.5 ± 3.6	52.1 ± 3.1	57.0 ± 2.7	58.2 ± 3.3
	we.	88.9 ± 2.0	89.0 ± 2.1	88.0 ± 2.0	89.3 ± 1.6	<u>89.7</u> ± 1.7

Bag-of-words VS word vectors, SVM VS deep learning



Preliminary attention VS max

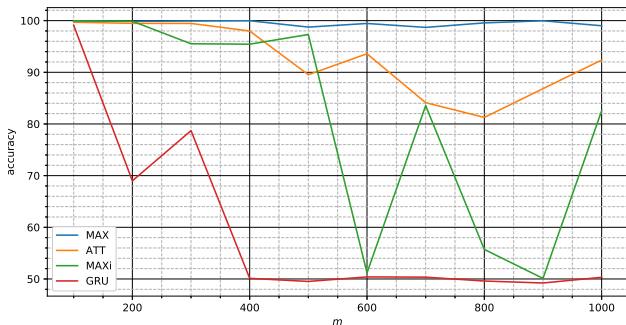
Answered questions

Q4 Compare novel attention model with simpler max pooling

Q6 Investigate the possibility to give interpretation to deep learning models

- ✓ Artificial dataset
- ✓ Same-size models

Preliminary attention VS max



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

Attention VS max, hierarchical VS plain

Answered questions

- Q1 Implement **large scale** study on deep learning applied to pathology reports
- Q2 Apply **novel** deep learning techniques, like **attention** models and **BERT**
- Q3 **Compare** classical **bag-of-words** techniques with newer deep learning techniques in this domain
- Q4 **Compare** novel **attention**-based and **hierarchical** techniques with simpler models
- Q6 **Investigate** the possibility to give **interpretation** to deep learning models

- ✓ **Temporal** setting
- ✓ On **main site** and **type** tasks
- ✓ Different **difficulty** classes

Attention VS max, hierarchical VS plain

	Topography				Morphology			
	Acc.	Top 3	Top 5	MacroF1	Acc.	Top 3	Top 5	Macro F1
<i>U-SVM</i>	89.7	95.9	96.8	60.0	82.4	94.0	95.6	53.7
<i>B-XGB</i>	89.1	95.8	97.2	58.0	84.1	94.4	96.5	59.6
<i>G-GRU</i>	89.9	96.5	97.7	58.3	83.3	94.6	96.6	55.2
<i>BERT</i>	89.9	96.3	97.8	56.6	84.3	93.2	94.9	51.1
<i>G-MAXi</i>	88.0	95.4	96.2	46.1	73.4	91.0	93.6	31.3
<i>G-MAXh</i>	89.9	96.2	97.8	58.8	83.7	94.4	96.4	54.5
<i>G-ATT_h</i>	89.9	96.3	97.7	58.0	83.7	94.4	96.2	57.5
<i>G-MAX</i>	90.3	96.6	98.1	61.9	84.6	95.0	96.9	59.2
<i>G-ATT</i>	90.1	96.2	97.6	60.0	84.8	94.9	96.9	61.3

	Topography			Morphology		
	easy (1000 < s) (4 cls)	avg. (100 < s ≤ 1000) (18 cls)	hard (s ≤ 100) (39 cls)	easy (1000 < s) (5 cls)	avg. (100 < s ≤ 1000) (18 cls)	hard (s ≤ 100) (111 cls)
<i>U-SVM</i>	95.7	86.9	50.9	90.5	68.6	48.4
<i>B-XGB</i>	95.6	86.4	48.2	92.0	72.4	54.8
<i>G-GRU</i>	96.1	72.2	48.0	91.4	71.6	49.7
<i>BERT</i>	95.7	73.2	44.9	92.9	74.4	43.9
<i>G-MAXi</i>	95.0	66.6	31.4	87.1	41.9	25.1
<i>G-MAXh</i>	95.8	72.4	48.8	92.7	71.8	48.8
<i>G-ATT_h</i>	96.0	73.1	47.1	91.9	72.3	52.6
<i>G-MAX</i>	96.0	73.3	53.1	92.7	72.3	53.8
<i>G-ATT</i>	96.0	73.1	50.3	92.8	72.3	56.7

Interpretability

y_i	Relevant $h_{i,j}$	$x_{i,j}$; relevant $h_{i,j}$
61	<u>61</u> (PROSTATE GLAND)	DISOMOGENICITA' DIFFUSE . <u>PSA NON PERVENUTO</u> . ADENOCARCINOMA PROSTATICO A GRADO DI DIFFERENZIAZIONE MEDIO - BASSO (<u>GLEASON 3 + 4</u>) NEI PRELIEVI DI CUI AI NN . 2 E 3 . AGOBIOPSIA DELLA <u>PROSTATA : 1</u>) 1 PRELIEVO LL DX . 2) 2 PRELIEVI ML DX . 3) 2 PRELIEVI M DX . 4) 1 PRELIEVO M SX . 5) 2 PRELIEVI ML SX . 6) 1 PRELIEVO LL SX . 7) 1 PRELIEVO <u>TRANSIZIONALE SX . 8</u>) 1 PRELIEVO <u>TRANSIZIONALE DX</u> .
20	<u>18</u> (COLON) <u>20</u> (RECTUM) <u>21</u> (ANUS AND ANAL CANAL)	ISOLATI FRAMMENTI RIFERIBILI AD ADENOMA <u>TUBULARE INTESTINALE</u> DI ALTO GRADO . FRAMMENTI (NR . 2) DI <u>POLIPO PEDUNCOLATO</u> A 20 CM <u>DALL'</u> ORIFIZIO ANALE . (ESEGUITA COLORAZIONE EMATOSSILINA - EOSINA) . <u>-</u>
34	<u>34</u> (BRONCHUS AND LUNG) <u>56</u> (OVARY) <u>67</u> (BLADDER) <u>80</u> (UNKNOWN PRIMARY SITE)	VERSAMENTO PLEURICO SX DI N . D . D . E ADDENSAMENTI POLMONARI DI N . D . D . , NODULI PARETE ADDOMINALE . INFILTRAZIONE <u>CANCERIGNA</u> DEGLI STROMI CONNETTIVO - ADIPOSI . IMMUNOISTOCHEMICA : <u>CK7 +</u> , <u>CK20 -</u> , <u>TTF - 1 -</u> , <u>PROTEINA S - 100 -</u> . LESIONE DI CM 2 , 0 X 1 , 3 X 0 , 7 . 1 - 2) <u>SEZIONI SERIATE</u> . <u>-</u>

Interpretability

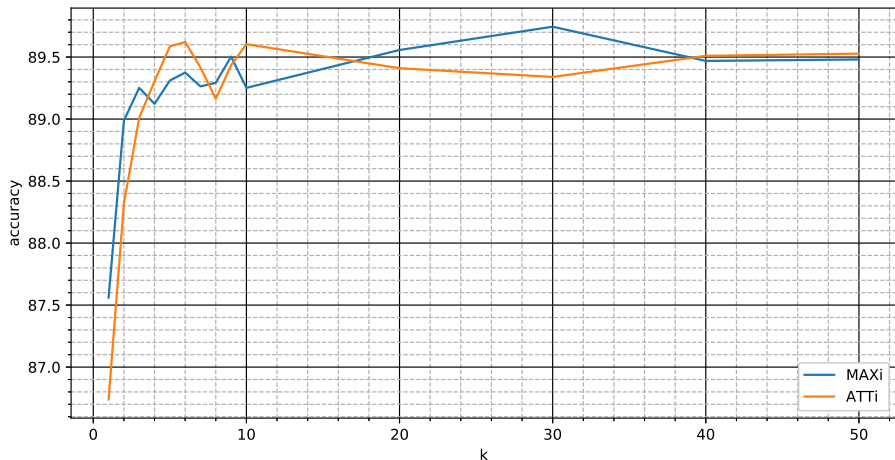


Figure: Training of a plain GRU model on a dataset created using ***G-MAXi*** and ***G-ATTi*** to keep the first k words

Conclusions

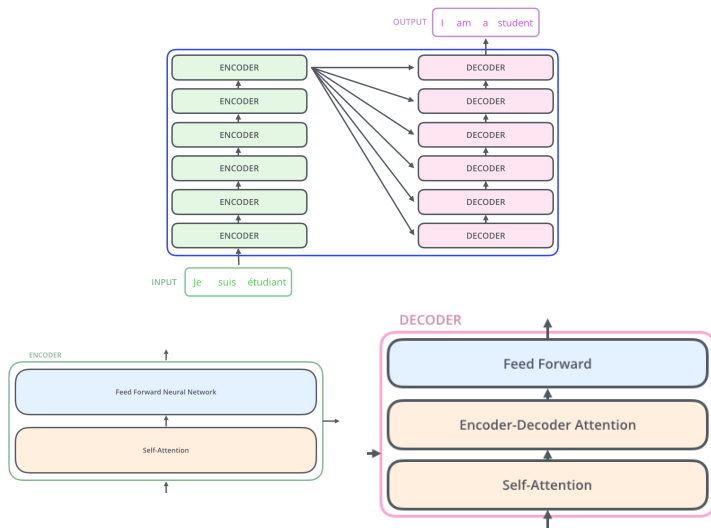
- ✓ We effectively implemented a **large scale** study on classical machine learning and novel deep learning methods applied to pathology reports
- ✓ In this context, **bag-of-words** techniques are not considerably worst than **deep learning**
- ✓ **Hierarchical** model are not beneficial
- ✓ **Attention** models are almost equivalent to a simpler element-wise **max** pooling model
- ✓ **Word vectors** can be effectively employed
- ✓ We can implement **interpretable** models without catastrophic loss

The End.



Questions? Thank you!

Bidirectional Encoder Representations from Transformers (BERT)

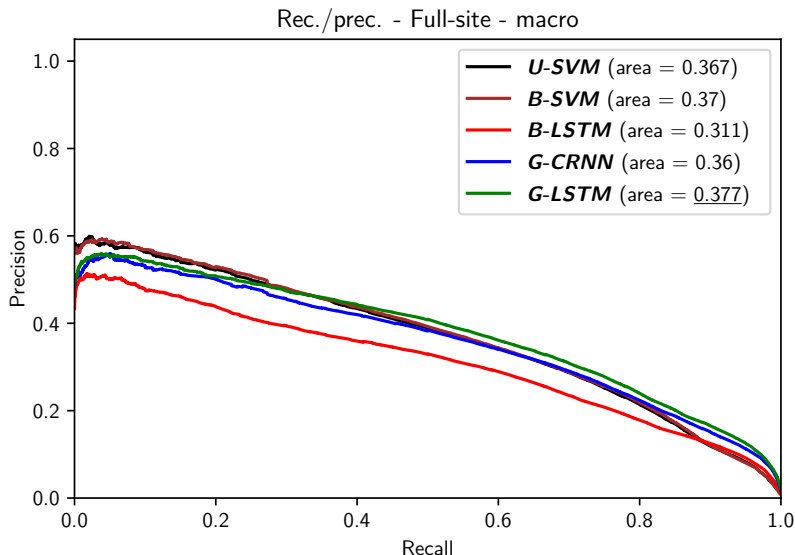


Bag-of-words VS word vectors, SVM VS deep learning

Table: Results for Full-site task.

		<i>U-SVM</i>	<i>B-SVM</i>	<i>B-LSTM</i>	<i>G-CRNN</i>	<i>G-LSTM</i>
accuracy		68.4 ± 2.3	68.7 ± 2.0	67.4 ± 1.7	70.1 ± 2.1	<u>70.9 ± 2.0</u>
kappa		66.5 ± 2.4	66.8 ± 2.1	65.6 ± 1.7	68.4 ± 2.2	<u>69.3 ± 2.1</u>
MAPs		78.4 ± 1.9	78.4 ± 1.7	78.5 ± 1.3	80.6 ± 1.4	<u>81.3 ± 1.4</u>
MAPc		43.1 ± 2.2	43.4 ± 2.2	36.8 ± 2.3	42.9 ± 2.6	<u>45.0 ± 2.0</u>
pre.	ma.	41.4 ± 1.6	<u>41.6 ± 1.5</u>	33.0 ± 2.8	38.7 ± 3.1	39.8 ± 2.3
	we.	66.3 ± 1.9	67.1 ± 1.7	66.1 ± 1.3	68.8 ± 1.9	<u>69.5 ± 1.5</u>
rec.	ma.	35.7 ± 1.9	35.1 ± 2.1	32.0 ± 2.5	36.6 ± 3.0	<u>38.0 ± 2.2</u>
	we.	68.4 ± 2.3	68.7 ± 2.0	67.4 ± 1.7	70.1 ± 2.1	<u>70.9 ± 2.0</u>
f1s.	ma.	36.6 ± 1.5	36.4 ± 1.7	31.2 ± 2.3	35.9 ± 2.9	<u>37.3 ± 2.1</u>
	we.	66.2 ± 2.1	66.8 ± 1.8	66.0 ± 1.3	68.5 ± 2.0	<u>69.5 ± 1.8</u>

Bag-of-words VS word vectors, SVM VS deep learning

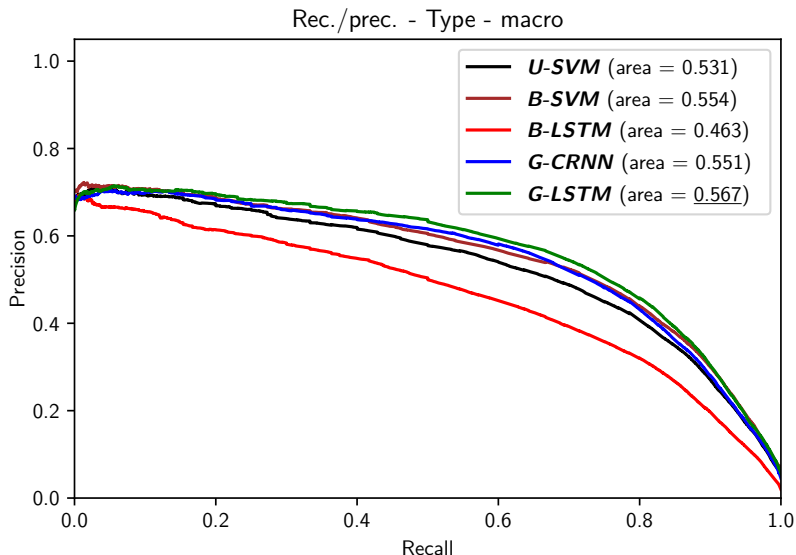


Bag-of-words VS word vectors, SVM VS deep learning

Table: Results for Type task.

		<i>U-SVM</i>	<i>B-SVM</i>	<i>B-LSTM</i>	<i>G-CRNN</i>	<i>G-LSTM</i>
accuracy		81.9 ± 1.9	82.9 ± 2.0	82.8 ± 1.4	84.6 ± 1.4	<u>84.9 ± 1.5</u>
kappa		79.5 ± 2.2	80.7 ± 2.3	80.6 ± 1.6	82.7 ± 1.6	<u>83.0 ± 1.7</u>
MAPs		87.8 ± 1.3	88.6 ± 1.4	88.7 ± 1.0	90.3 ± 0.9	<u>90.6 ± 1.0</u>
MAPc		62.4 ± 1.6	64.4 ± 1.8	55.1 ± 3.1	64.2 ± 1.9	<u>65.9 ± 1.9</u>
pre.	ma.	56.1 ± 2.4	<u>58.3 ± 1.9</u>	47.0 ± 3.3	56.5 ± 1.8	57.0 ± 2.6
	we.	80.3 ± 1.8	81.8 ± 1.9	82.0 ± 1.3	84.1 ± 1.3	<u>84.3 ± 1.5</u>
rec.	ma.	51.1 ± 2.6	52.2 ± 2.2	47.0 ± 2.6	56.8 ± 2.2	<u>58.6 ± 2.0</u>
	we.	81.9 ± 1.9	82.9 ± 2.0	82.8 ± 1.4	84.6 ± 1.4	<u>84.9 ± 1.5</u>
f1s.	ma.	51.4 ± 2.5	52.9 ± 1.9	45.0 ± 2.9	54.6 ± 1.9	<u>55.5 ± 2.3</u>
	we.	80.4 ± 2.0	81.7 ± 2.0	81.9 ± 1.3	83.8 ± 1.4	<u>84.0 ± 1.5</u>

Bag-of-words VS word vectors, SVM VS deep learning



Bag-of-words VS word vectors, SVM VS deep learning

Table: Results for Behavior task.

		<i>U-SVM</i>	<i>B-SVM</i>	<i>B-LSTM</i>	<i>G-CRNN</i>	<i>G-LSTM</i>
accuracy		95.9 ± 1.0	96.0 ± 1.1	94.1 ± 3.0	94.4 ± 4.2	<u>96.5 ± 0.8</u>
kappa		82.3 ± 4.6	82.8 ± 5.0	70.4 ± 25.5	67.6 ± 35.9	<u>85.6 ± 3.4</u>
MAPs		97.7 ± 0.6	97.8 ± 0.6	96.6 ± 1.8	96.8 ± 2.5	<u>98.1 ± 0.5</u>
MAPc		85.4 ± 5.9	85.9 ± 5.7	71.4 ± 18.4	75.5 ± 26.4	<u>89.5 ± 4.2</u>
pre.	ma.	87.0 ± 5.0	<u>87.9 ± 4.8</u>	69.9 ± 19.9	72.7 ± 27.1	85.5 ± 4.0
	we.	95.8 ± 1.1	95.9 ± 1.2	92.6 ± 6.4	92.1 ± 9.0	<u>96.6 ± 0.8</u>
rec.	ma.	78.6 ± 7.3	78.6 ± 7.4	67.6 ± 17.4	72.1 ± 25.4	<u>85.9 ± 4.9</u>
	we.	95.9 ± 1.0	96.0 ± 1.1	94.1 ± 3.0	94.4 ± 4.2	<u>96.5 ± 0.8</u>
f1s.	ma.	81.7 ± 6.3	82.0 ± 6.3	68.0 ± 18.5	72.1 ± 26.0	<u>85.5 ± 4.2</u>
	we.	95.8 ± 1.1	95.9 ± 1.2	93.2 ± 4.8	93.1 ± 6.7	<u>96.5 ± 0.8</u>

Bag-of-words VS word vectors, SVM VS deep learning

