Classification of cancer pathology reports with Deep Learning methods

Stefano Martina
stefano.martina@unifi.it

Università degli Studi Firenze

10 December 2019

Overview

Cancer

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)

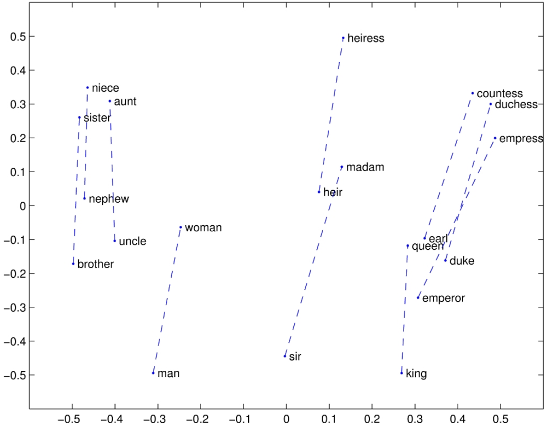
Bag-of-words

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

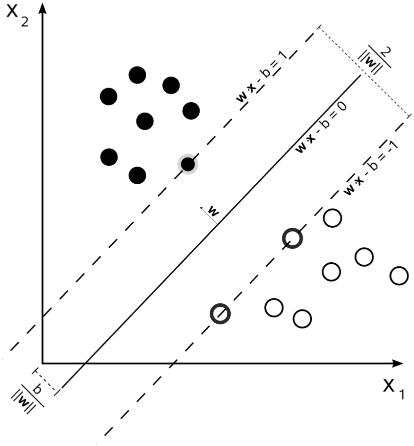
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Word vectors

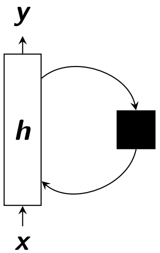
- ✔ Transforms words in vectors
- ✔ Unsupervised learning method
- ✔ Semantic relations encoded in vector space geometric relations



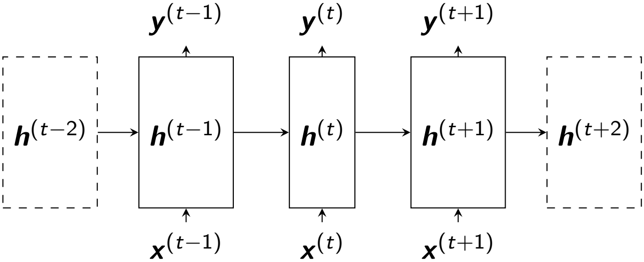
Support Vector Machine (SVM)



Recurrent Neural Network (RNN)



(a)

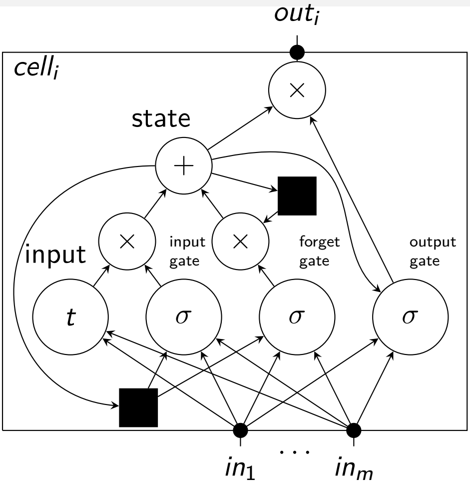


(b)

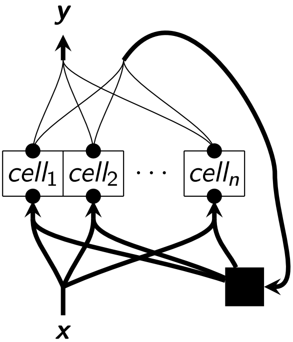
Figure: RNN, folded (a) and unfolded (b) models.

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

Long Short-Term Memory (LSTM)



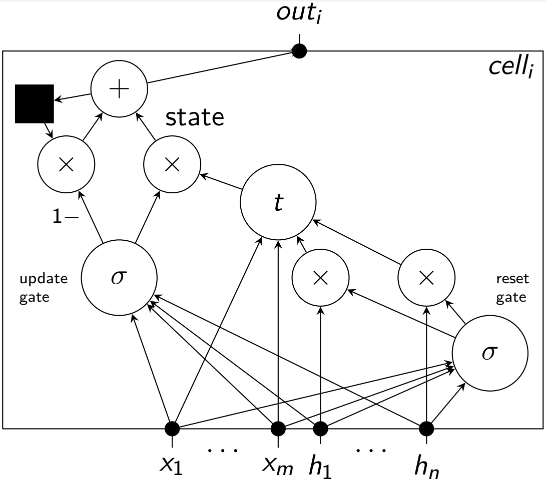
(a)



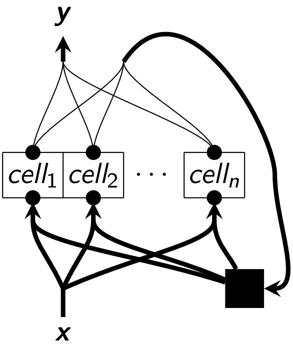
(b)

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

Gated Recurrent Unit (GRU)



(a)

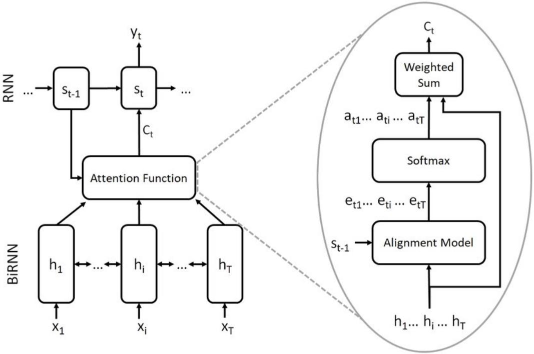


(b)

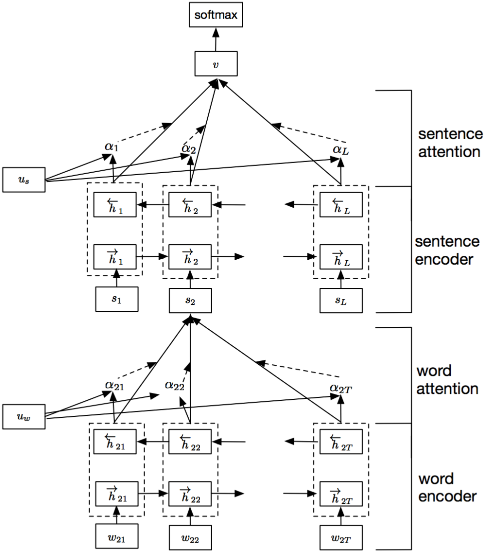
Figure: memory cell (a), and general scheme (b). The black box is a delay

Attention models

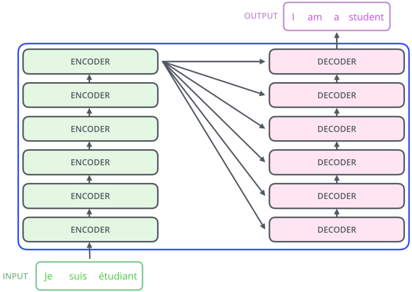
- ✔ Developed for seq-to-seq task
- ✔ State of the art in machine translation



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

Existing works (linear classifiers)

V. Jouhet, G. Defossez, A. Burgun, P. Le Beux, P. Levillain, P. Ingrand, and V. Claveau.

Automated Classification of Free-text Pathology Reports for Registration of Incident Cases of Cancer:.
Methods of Information in Medicine, 51(3):242–251, July 2011

- ✔ SVM and Naive Bayes classifiers
- ✔ 5 121 French pathology reports, 26 topographic classes and 18 morphological classes
- ✔ accuracy of 72.6% on topography and 86.4% on morphology

R. Kavuluru, I. Hands, E. B. Durbin, and L. Witt. Automatic extraction of ICD-O-3 primary sites from cancer pathology reports.
In *Clinical Research Informatics AMIA symposium (forthcoming)*, 2013

- ✔ SVM, Naive Bayes, and logistic regression
- ✔ 56 426 English reports, 14, 42, and 57 topography classes
- ✔ Micro-averaged F1 measure of 90%

Existing works (Deep Learning)

J. X. Qiu, H.-J. Yoon, P. A. Fearn, and G. D. Tourassi. Deep Learning for Automated Extraction of Primary Sites From Cancer Pathology Reports.
IEEE Journal of Biomedical and Health Informatics, 22(1):244–251, Jan. 2018

- ✔ CNN, word vectors pretrained on PubMed
- ✔ 942 breast and lung cancer English reports, 12 topography classes
- ✔ Micro-averaged F1 score of 72.2% on minimally populated, and 81.1% on well populated classes

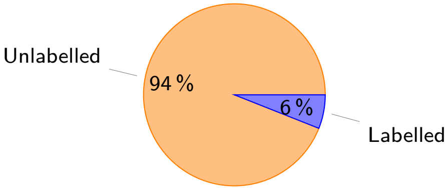
S. Gao, M. T. Young, J. X. Qiu, H.-J. Yoon, J. B. Christian, P. A. Fearn, G. D. Tourassi, and A. Ramanthan. Hierarchical attention networks for information extraction from cancer pathology reports.
Journal of the American Medical Informatics Association, 25(3):321–330, Mar. 2018

- ✔ RNN with hierarchical attention
- ✔ Same dataset
- ✔ Micro-averaged F1 score of 80%

Scientific questions

- Q1 Implement large scale study on machine 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

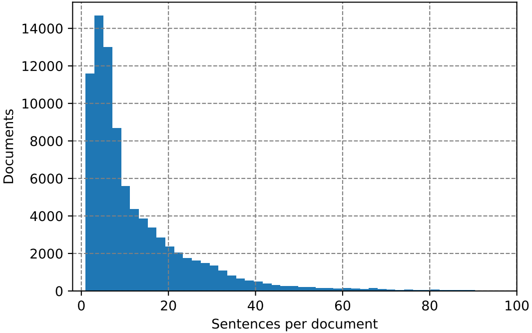
Structure

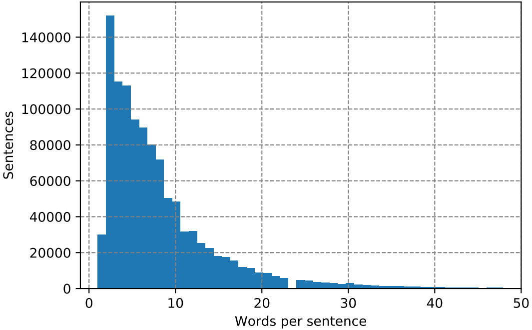
- ✔ 3 text fields: macroscopy, diagnosis, anamnesis
- ✔ field length from 0 to 1368 (quartiles 34, 62, 134)

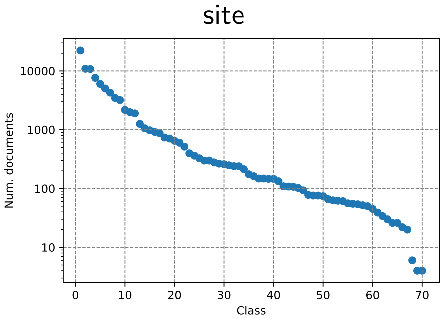
Preparation

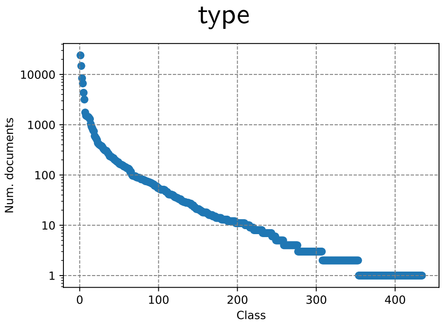
- Data comes in two tables to merge:
  1. neoplasm table, containing administrative and clinical variables
  2. histology table, containing the text fields

  - 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


Documents
Sentences per document
0
2000
4000
6000
8000
10000
12000
14000
0
20
40
60
80
100


Sentences
Words per sentence
0
20000
40000
60000
80000
100000
120000
140000
0
10
20
30
40
50


site
Num. documents
10000
1000
100
10
0
10
20
30
40
50
60
70
Class


type
Num. documents
10000
1000
100
10
1
0
100
200
300
400
Class

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-ATTh*** 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-MAXh*** hierarchical GRU with max pooling trained on GloVe
- ***G-MAXi*** GRU with max pooling, in interpretable setting, trained on GloVe
- ***G-ATTi*** GRU with attention, in interpretable setting, trained on GloVe

B-LSTM, ***G-CRNN***, ***G-LSTM***

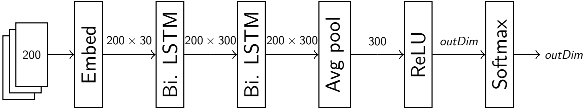


Figure: Scheme for ***B-LSTM*** model.

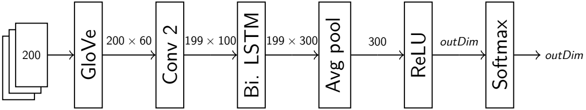


Figure: Scheme for ***G-CRNN*** model.

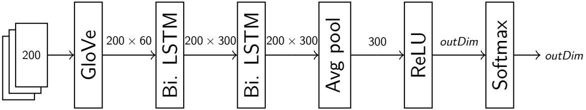


Figure: Scheme for ***G-LSTM*** model.

G-GRU, G-ATT, G-MAX

Plain model

$$
\begin{aligned}
e_t &= E(x_t; \theta^e) \\
h_t^f &= F(e_t, h_{t-1}^f; \theta^f) \\
h_t^r &= R(e_t, h_{t+1}^r; \theta^r) \\
u_t &= G(h_t; \theta^h) \\
\phi &= A(\boldsymbol{u}; \theta^a) \\
f(\boldsymbol{x}) &= g(\phi; \theta^c)
\end{aligned}
$$

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

G-ATTi, G-MAXi

Interpretable model

$$
\begin{aligned}
e_t &= E(x_t; \theta^e) \\
h_t^f &= F(e_t, h_{t-1}^f; \theta^f) \\
h_t^r &= R(e_t, h_{t+1}^r; \theta^r) \\
u_t &= G(h_t; \theta^h) \\
f(\boldsymbol{x}) &= A(\boldsymbol{u}; \theta^a)
\end{aligned}
$$

- ✔ $\phi = \sum_t a_t(\boldsymbol{u}; \theta^a) u_t, \; a_t(\boldsymbol{u}; \theta^a) = \frac{e^{\langle c, c_t \rangle}}{\sum_i e^{\langle c, c_i \rangle}}, \; c_t = C(\boldsymbol{u}; \theta^a)$
- ✔ $\phi_j = \max_t u_{j,t}$

G-ATTh, G-MAXh

Hierarchical model

$$
\begin{aligned}
e_{s,t} &= E(x_{s,t}; \theta^e) \\
h^f_{s,t} &= F(e_{s,t}, h^f_{s,t-1}; \theta^f) \\
h^r_{s,t} &= R(e_{s,t}, h^r_{s,t+1}; \theta^r) \\
u_{s,t} &= G(h_{s,t}; \theta^h) \\
\phi_s &= A(\boldsymbol{u}_s; \theta^a) \\
\bar{h}^f_s &= \bar{F}(\phi_s, \bar{h}^f_{s-1}; \bar{\theta}^f) \\
\bar{h}^r_s &= \bar{R}(\phi_s, \bar{h}^r_{s+1}; \bar{\theta}^r) \\
\bar{\phi} &= \bar{A}(\bar{\boldsymbol{h}}; \bar{\theta}^a) \\
f(\boldsymbol{x}) &= g(\bar{\phi}; \theta^c)
\end{aligned}
$$

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

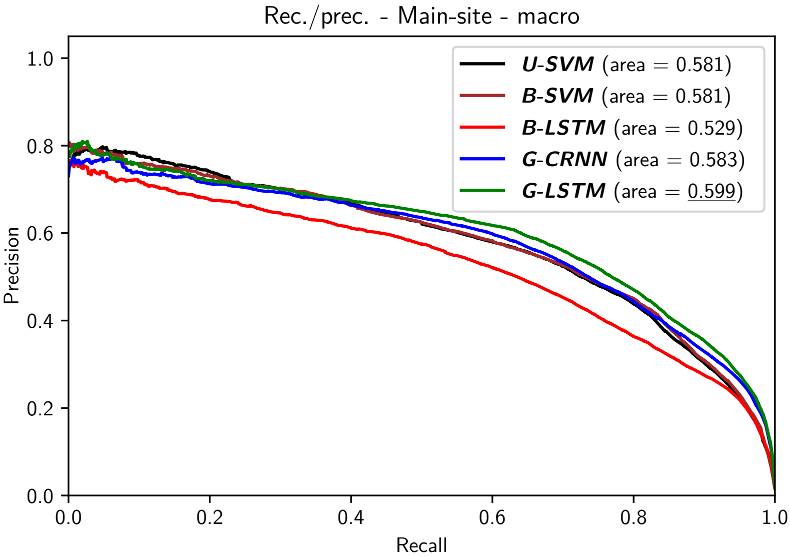
- ✔ 10-fold cross validation
- ✔ All tasks (main site, site+subsite, type, behavior)

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

Table: Results for Main-site task.

		U-SVM	*B-SVM*	*B-LSTM*	*G-CRNN*	*G-LSTM*
accuracy		89.8 ± 2.0	89.8 ± 2.0	88.6 ± 2.0	90.0 ± 1.6	$\underline{90.5} \pm 1.6$
kappa		88.5 ± 2.2	88.6 ± 2.3	87.2 ± 2.3	88.9 ± 1.8	$\underline{89.3} \pm 1.8$
MAPs		93.0 ± 1.5	93.0 ± 1.5	92.2 ± 1.5	93.5 ± 1.2	$\underline{93.8} \pm 1.1$
MAPc		61.6 ± 3.9	61.3 ± 4.0	55.7 ± 3.7	62.7 ± 3.5	$\underline{64.1} \pm 4.1$
pre.	**ma.**	$\underline{65.5} \pm 4.8$	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	$\underline{89.5} \pm 1.7$
rec.	**ma.**	55.7 ± 4.1	54.7 ± 3.8	51.6 ± 3.2	56.5 ± 3.0	$\underline{58.1} \pm 3.5$
	we.	89.8 ± 2.0	89.8 ± 2.0	88.6 ± 2.0	90.0 ± 1.6	$\underline{90.5} \pm 1.6$
f1s.	**ma.**	$\underline{58.4} \pm 4.1$	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	$\underline{89.7} \pm 1.7$

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



Preliminary attention VS max

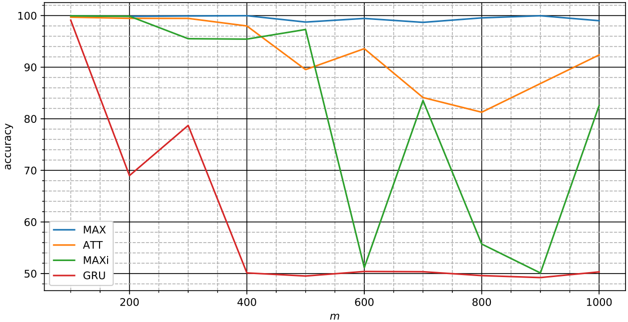
Answered questions

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 0 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
U-SVM	89.7	95.9	96.8	60.0	82.4	94.0	95.6	53.7
B-XGB	89.1	95.8	97.2	58.0	84.1	94.4	96.5	59.6
G-GRU	89.9	96.5	97.7	58.3	83.3	94.6	96.6	55.2
BERT	89.9	96.3	97.8	56.6	84.3	93.2	94.9	51.1
G-MAXi	88.0	95.4	96.2	46.1	73.4	91.0	93.6	31.3
G-MAXh	89.9	96.2	97.8	58.8	83.7	94.4	96.4	54.5
G-ATTh	89.9	96.3	97.7	58.0	83.7	94.4	96.2	57.5
G-MAX	**90.3**	**96.6**	**98.1**	**61.9**	84.6	**95.0**	**96.9**	59.2
G-ATT	90.1	96.2	97.6	60.0	**84.8**	94.9	**96.9**	**61.3**

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

Interpretability

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

Interpretability

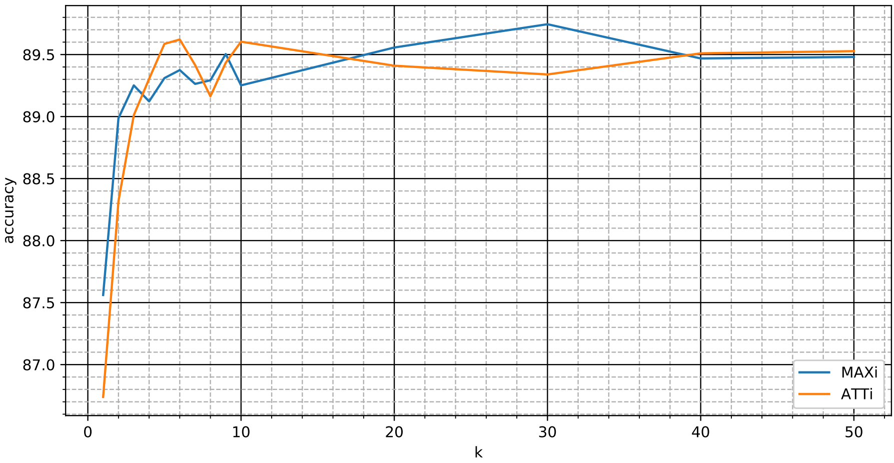


Figure: Training of a plain GRU model on a dataset created using ***G-MAXi*** and ***G-ATTi*** to keep the first k words

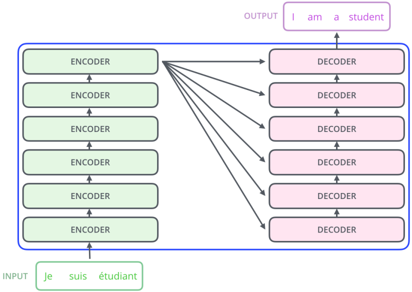
Conclusions

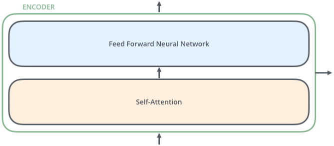
- We effectively implemented a large scale study on classical machine learning and novel deep learning methods applied to pathology reports
- In this context, bag-of-words techniques are not considerably worst than deep learning
- Hierarchical model are not beneficial
- Attention models are almost equivalent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

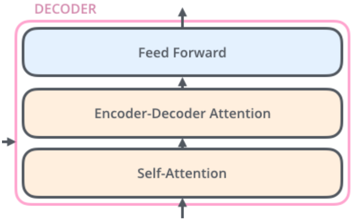
The End.

Questions? Thank you!

Bidirectional Encoder Representations from Transformers (BERT)





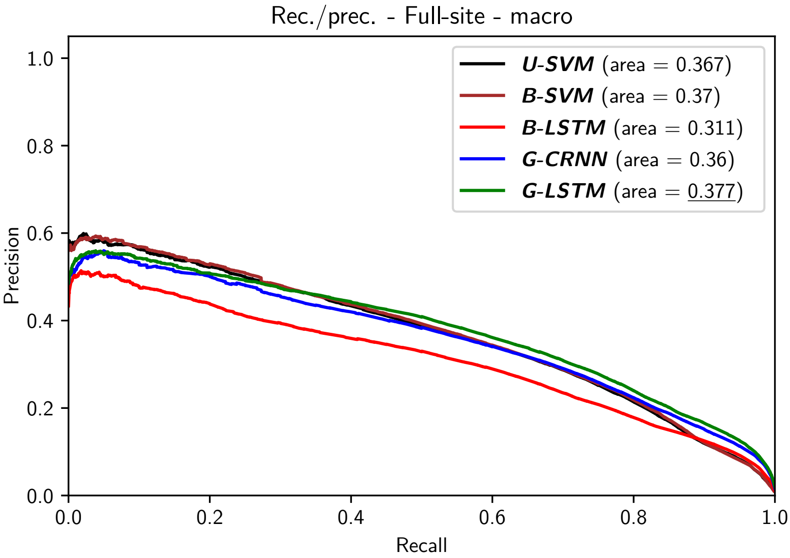


Bag-of-words VS word vectors, SVM VS deep learning

Table: Results for Full-site task.

		U-SVM	*B-SVM*	*B-LSTM*	*G-CRNN*	*G-LSTM*
accuracy		68.4 ± 2.3	68.7 ± 2.0	67.4 ± 1.7	70.1 ± 2.1	$\underline{70.9} \pm 2.0$
kappa		66.5 ± 2.4	66.8 ± 2.1	65.6 ± 1.7	68.4 ± 2.2	$\underline{69.3} \pm 2.1$
MAPs		78.4 ± 1.9	78.4 ± 1.7	78.5 ± 1.3	80.6 ± 1.4	$\underline{81.3} \pm 1.4$
MAPc		43.1 ± 2.2	43.4 ± 2.2	36.8 ± 2.3	42.9 ± 2.6	$\underline{45.0} \pm 2.0$
pre.	**ma.**	41.4 ± 1.6	$\underline{41.6} \pm 1.5$	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	$\underline{69.5} \pm 1.5$
rec.	**ma.**	35.7 ± 1.9	35.1 ± 2.1	32.0 ± 2.5	36.6 ± 3.0	$\underline{38.0} \pm 2.2$
	we.	68.4 ± 2.3	68.7 ± 2.0	67.4 ± 1.7	70.1 ± 2.1	$\underline{70.9} \pm 2.0$
f1s.	**ma.**	36.6 ± 1.5	36.4 ± 1.7	31.2 ± 2.3	35.9 ± 2.9	$\underline{37.3} \pm 2.1$
	we.	66.2 ± 2.1	66.8 ± 1.8	66.0 ± 1.3	68.5 ± 2.0	$\underline{69.5} \pm 1.8$

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

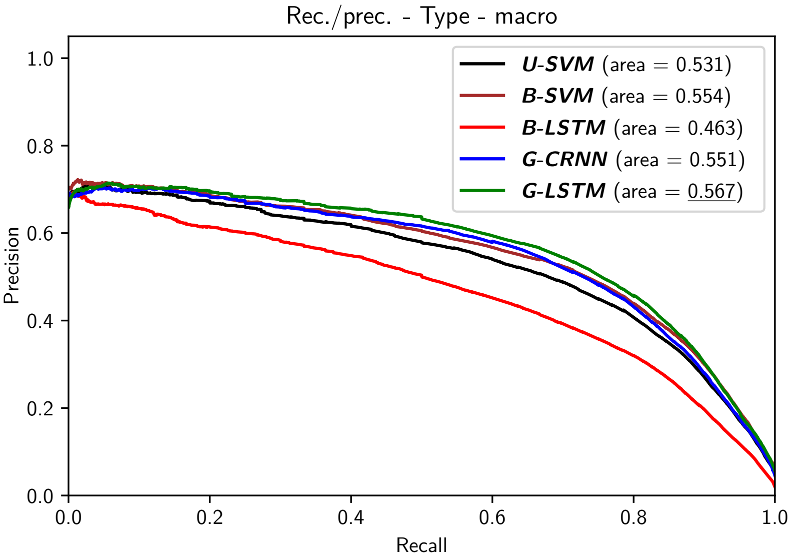


Bag-of-words VS word vectors, SVM VS deep learning

Table: Results for Type task.

		U-SVM	***B-SVM***	***B-LSTM***	***G-CRNN***	***G-LSTM***
accuracy		81.9 ± 1.9	82.9 ± 2.0	82.8 ± 1.4	84.6 ± 1.4	$\underline{84.9} \pm 1.5$
kappa		79.5 ± 2.2	80.7 ± 2.3	80.6 ± 1.6	82.7 ± 1.6	$\underline{83.0} \pm 1.7$
MAPs		87.8 ± 1.3	88.6 ± 1.4	88.7 ± 1.0	90.3 ± 0.9	$\underline{90.6} \pm 1.0$
MAPc		62.4 ± 1.6	64.4 ± 1.8	55.1 ± 3.1	64.2 ± 1.9	$\underline{65.9} \pm 1.9$
pre.	**ma.**	56.1 ± 2.4	$\underline{58.3} \pm 1.9$	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	$\underline{84.3} \pm 1.5$
rec.	**ma.**	51.1 ± 2.6	52.2 ± 2.2	47.0 ± 2.6	56.8 ± 2.2	$\underline{58.6} \pm 2.0$
	we.	81.9 ± 1.9	82.9 ± 2.0	82.8 ± 1.4	84.6 ± 1.4	$\underline{84.9} \pm 1.5$
f1s.	**ma.**	51.4 ± 2.5	52.9 ± 1.9	45.0 ± 2.9	54.6 ± 1.9	$\underline{55.5} \pm 2.3$
	we.	80.4 ± 2.0	81.7 ± 2.0	81.9 ± 1.3	83.8 ± 1.4	$\underline{84.0} \pm 1.5$

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

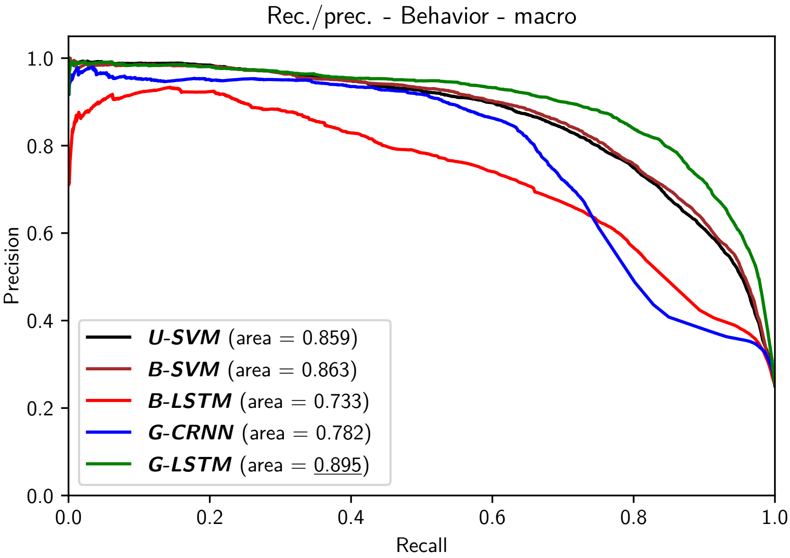


Bag-of-words VS word vectors, SVM VS deep learning

Table: Results for Behavior task.

		U-SVM	*B-SVM*	*B-LSTM*	*G-CRNN*	*G-LSTM*
accuracy		95.9 ± 1.0	96.0 ± 1.1	94.1 ± 3.0	94.4 ± 4.2	$\underline{96.5} \pm 0.8$
kappa		82.3 ± 4.6	82.8 ± 5.0	70.4 ± 25.5	67.6 ± 35.9	$\underline{85.6} \pm 3.4$
MAPs		97.7 ± 0.6	97.8 ± 0.6	96.6 ± 1.8	96.8 ± 2.5	$\underline{98.1} \pm 0.5$
MAPc		85.4 ± 5.9	85.9 ± 5.7	71.4 ± 18.4	75.5 ± 26.4	$\underline{89.5} \pm 4.2$
pre.	**ma.**	87.0 ± 5.0	$\underline{87.9} \pm 4.8$	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	$\underline{96.6} \pm 0.8$
rec.	**ma.**	78.6 ± 7.3	78.6 ± 7.4	67.6 ± 17.4	72.1 ± 25.4	$\underline{85.9} \pm 4.9$
	we.	95.9 ± 1.0	96.0 ± 1.1	94.1 ± 3.0	94.4 ± 4.2	$\underline{96.5} \pm 0.8$
f1s.	**ma.**	81.7 ± 6.3	82.0 ± 6.3	68.0 ± 18.5	72.1 ± 26.0	$\underline{85.5} \pm 4.2$
	we.	95.8 ± 1.1	95.9 ± 1.2	93.2 ± 4.8	93.1 ± 6.7	$\underline{96.5} \pm 0.8$

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



Bibliography I

[1] S. Gao, M. T. Young, J. X. Qiu, H.-J. Yoon, J. B. Christian, P. A. Fearn, G. D. Tourassi, and A. Ramanthan.
Hierarchical attention networks for information extraction from cancer pathology reports.
Journal of the American Medical Informatics Association, 25(3):321–330, Mar. 2018.

[2] V. Jouhet, G. Defossez, A. Burgun, P. Le Beux, P. Levillain, P. Ingrand, and V. Claveau.
Automated Classification of Free-text Pathology Reports for Registration of Incident Cases of Cancer:.
Methods of Information in Medicine, 51(3):242–251, July 2011.

[3] R. Kavuluru, I. Hands, E. B. Durbin, and L. Witt.
Automatic extraction of ICD-O-3 primary sites from cancer pathology reports.
In *Clinical Research Informatics AMIA symposium (forthcoming)*, 2013.

[4] J. X. Qiu, H.-J. Yoon, P. A. Fearn, and G. D. Tourassi.
Deep Learning for Automated Extraction of Primary Sites From Cancer Pathology Reports.
IEEE Journal of Biomedical and Health Informatics, 22(1):244–251, Jan. 2018.