Prediction of Variables from Cancer Reports

Candidato
Federico Magnolfi

Relatori
Prof. Paolo Frasconi
Prof. Simone Marinai

Correlatori
Leonardo Ventura, Stefano Martina

Università degli Studi di Firenze
Corso Laurea Magistrale in Ingegneria Informatica

May 31, 2021

Introduction

Tuscany cancer reports

- Doctors write reports after **oncological visits**
- **ISPRO collects reports** in Tuscany
- Experts extract **variables** from reports

Variables

- describe the **progress of the pathology**
- used to ensure that patients are receving the **correct care**

Problem

Reports are analyzed with some **years of delay** ($\sim$ 5 years)

Prediction of variables

Objective

Predict variables to **speedup** the analyses of reports

This thesis

Study the **predictability** of these variables for breast cancer reports

Previous work [1]

Dataset with all cancer types, prediction of **primary site** and **morphology**

[1] S. Martina, L. Ventura and P. Frasconi, “Classification of Cancer Pathology Reports: A Large-Scale Comparative Study”

Breast cancer dataset

Example of report

field	value
notizie	
macroscopia	Q.I.C. mammella sn (cm 12x8x5):\nT1-4) neoplasia (mm 23), distanza dai margini >mm 10; MS) margine superiore; MI) margine inferiore; MM) margine mediale; ML) margine laterale; MP) margine profondo; CU) margine cutaneo. \n(eseguita colorazione ematossilina-eosina e valutazione parametri biologici con controllo di qualitï¿½).
diagnosi	CARCINOMA DUTTALE INFILTRANTE (N.O.S.) (G2) DELLA MAMMELLA CON ASSOCIATE ESPRESSIONI INTRADUTTALI DI BASSO GRADO (T1-4)\nNON EVIDENTE PERMEAZIONE NEOPLASTICA VASCOLARE \n NESSUNA PROLIFERAZIONE CANCERIGNA NEI MARGINI DI SEZIONE CHIRURGICA (MS, MI, ML, MM, MP), NELLLA CUTE (CU), NEL LINFONODO SENTINELLA (vedi es.B 10508/12).\n(p T2 N0(OSNA -))* \n*(TNM, VIIï¿½ ed., 2009)\nParametri biologici: ER: + 90% ; PGR: + 60% ; ki67: + 10% ; Her 2: - .

field	value
id_paz	5******
anno_diagnosi	2012
sede_icdo3	C509
morfologia_icdo3	85003
dimensioni	23
tipo_T	P
metastasi	
modalita_T	E
modalita_N	E
stadio_T	2
stadio_N	0SN
recettori_estrogeni	90
recettori_progestin	60
numero_sentinella_asportati	1
numero_sentinella_positivi	0
mib1	
cerb	0
ki67	10
grading	2
anno_referto	2012
id_isto	5******

Example of report

field	value
notizie	
macroscopia	Q.I.C. mammella sn (cm 12x8x5):\nT1-4) neoplasia (mm 23), distanza dai margini >mm 10; MS) margine superiore; MI) margine inferiore; MM) margine mediale; ML) margine laterale; MP) margine profondo; CU) margine cutaneo. \n(eseguita colorazione ematossilina-eosina e valutazione parametri biologici con controllo di qualitï¿½).
diagnosi	CARCINOMA DUTTALE INFILTRANTE (N.O.S.) (G2) DELLA MAMMELLA CON ASSOCIATE ESPRESSIONI INTRADUTTALI DI BASSO GRADO (T1-4)\nNON EVIDENTE PERMEAZIONE NEOPLASTICA VASCOLARE \n NESSUNA PROLIFERAZIONE CANCERIGNA NEI MARGINI DI SEZIONE CHIRURGICA (MS, MI, ML, MM, MP), NELLLA CUTE (CU), NEL LINFONODO SENTINELLA (vedi es.B 10508/12).\n(p T2 N0(OSNA -))* \n*(TNM, VIIï¿½ ed., 2009)\nParametri biologici: ER: + 90% ; PGR: + 60% ; ki67: + 10% ; Her 2: - .

field	value
id_paz	5******
anno_diagnosi	2012
sede_icdo3	C509
morfologia_icdo3	85003
dimensioni	23
tipo_T	P
metastasi	
modalita_T	E
modalita_N	E
stadio_T	2
stadio_N	0SN
recettori_estrogeni	90
recettori_progestin	60
numero_sentinella_asportati	1
numero_sentinella_positivi	0
mib1	
cerb	0
ki67	10
grading	2
anno_referto	2012
id_isto	5******

Example of report

field	value
notizie	
macroscopia	Q.I.C. mammella sn (cm 12x8x5):\nT1-4) neoplasia (mm 23), distanza dai margini >mm 10; MS) margine superiore; MI) margine inferiore; MM) margine mediale; ML) margine laterale; MP) margine profondo; CU) margine cutaneo. \n(eseguita colorazione ematossilina-eosina e valutazione parametri biologici con controllo di qualitï¿½).
diagnosi	CARCINOMA DUTTALE INFILTRANTE (N.O.S.) (G2) DELLA MAMMELLA CON ASSOCIATE ESPRESSIONI INTRADUTTALI DI BASSO GRADO (T1-4)\nNON EVIDENTE PERMEAZIONE NEOPLASTICA VASCOLARE \n NESSUNA PROLIFERAZIONE CANCERIGNA NEI MARGINI DI SEZIONE CHIRURGICA (MS, MI, ML, MM, MP), NELLLA CUTE (CU), NEL LINFONODO SENTINELLA (vedi es.B 10508/12).\n(p T2 N0(OSNA -))* \n*(TNM, VIIï¿½ ed., 2009)\nParametri biologici: ER: + 90% ; PGR: + 60% ; ki67: + 10% ; Her 2: - .

field	value
id_paz	5******
anno_diagnosi	2012
sede_icdo3	C509
morfologia_icdo3	85003
dimensioni	23
tipo_T	P
metastasi	
modalita_T	E
modalita_N	E
stadio_T	2
stadio_N	0SN
recettori_estrogeni	90
recettori_progestin	60
numero_sentinella_asportati	1
numero_sentinella_positivi	0
mib1	
cerb	0
ki67	10
grading	2
anno_referto	2012
id_isto	5******

Some variables

- **grading**: difference between cancer cells and healthy ones
- **tumor stage**: extension of the primary tumor
- **lymph nodes stage**: involvement of lymph nodes
- **ki67**: marker of tumor cells proliferation speed
- **removed lymph nodes**: how many lymph nodes was removed
- **positive lymph nodes**: how many lymph nodes had malignant cells
- **size**: size of the primary tumor

Dataset info

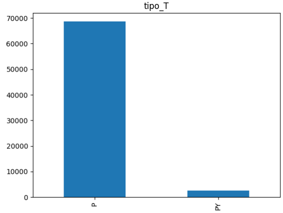
- breast cancer
- $\sim$ 25k patients
- $\sim$ 115k reports
- more than 10 variables

Dataset info

- breast cancer
- $\sim$ 25k patients $\rightarrow$ labeled from 2003 to 2015
- $\sim$ 115k reports
- more than 10 variables $\rightarrow$ many missing values

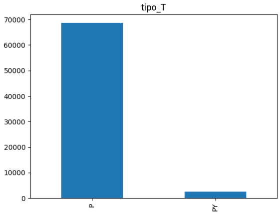
Types of variables

binary variable

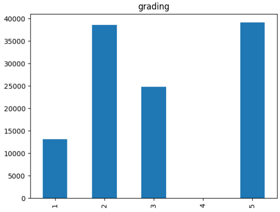


Types of variables

binary variable

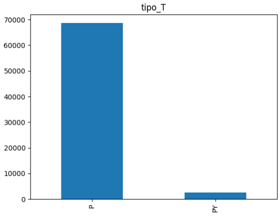


multi-class

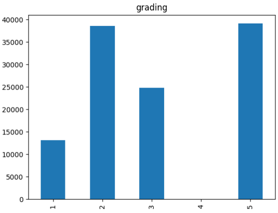


Types of variables

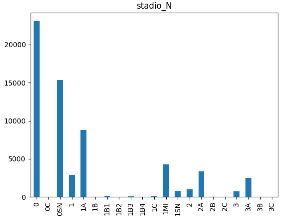
binary variable



multi-class

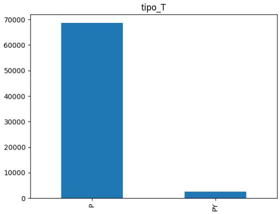


with subclasses

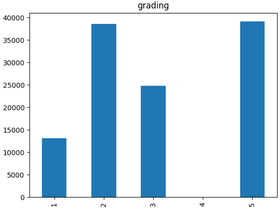


Types of variables

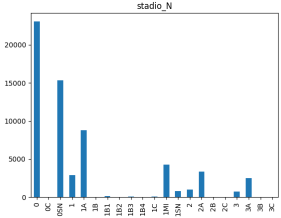
binary variable



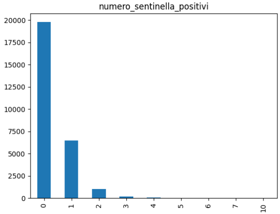
multi-class



with subclasses

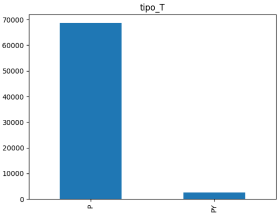


with missing values

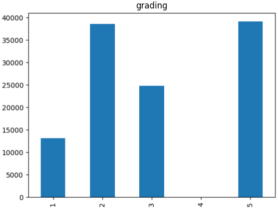


Types of variables

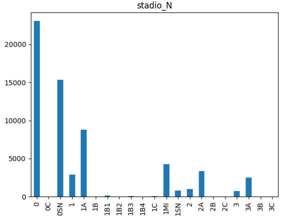
binary variable



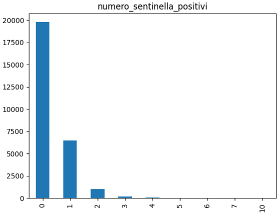
multi-class



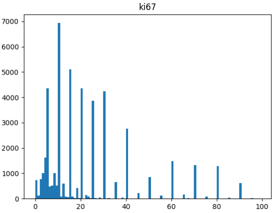
with subclasses



with missing values

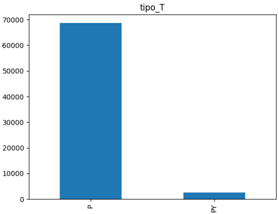


percentages

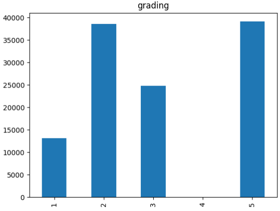


Types of variables

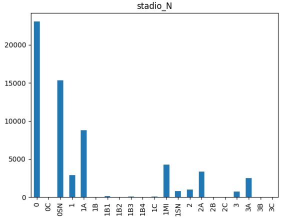
binary variable



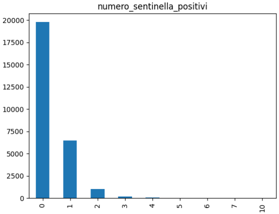
multi-class



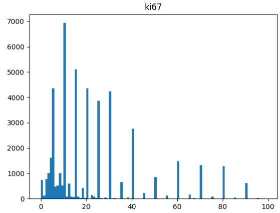
with subclasses



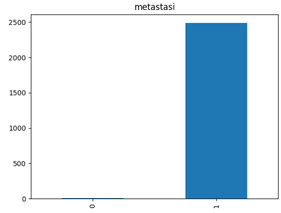
with missing values



percentages



unpredictable



Methods and models

Question we want to answer

Question

Is it possible to **predict variables** from these reports?

There are no previous references to compare

Approach

Different strategies for different variables:

1. **classification**

   Examples:

   - "*carcinoma duttale infiltrante nos* **g3**"
   - "*nessuna proliferazione neoplastica nel linfonodo sentinella*"

2. **segmentation**

   Examples:

   - "*Ki67 (clone MIB1):* **80%**"
   - "*neoplasia (***mm 23***), distanza dai margini >mm 10;*"

Approach

Different strategies for different variables:

1. **classification** → **machine learning models**

   Examples:

   - "*carcinoma duttale infiltrante nos* **g3**"
   - "*nessuna proliferazione neoplastica nel linfonodo sentinella*"

2. **segmentation** → **regex-based algorithm**

   Examples:

   - "*Ki67 (clone MIB1):* **80%**"
   - "*neoplasia (***mm 23***), distanza dai margini >mm 10;*"

Classifications: preparation steps

report → preprocess → tokenize → tokens

Example:

“CARCINOMA DUTTALE INFILTRANTE (NOS) DELLA MAMMELLA. Grado 3 secondo Elston- Ellis.”

↓ **preprocess**

“carcinoma duttale infiltrante (nos) della mammella . g3 secondo elston - ellis .”

↓ **tokenize**

[“carcinoma”, “duttale”, “infiltrante”, “(”, “nos”, “)”, “della”, “mammella”, “.”, “g3”, “secondo”, “elston”, “-”, “ellis”, “.”]

Classifications: models

- Transformer
- MLP
- Decision Tree
- Random Forest
- XGBoost
- Linear SVM

Transformer, MLP $\rightarrow$ Neural Networks

Random Forest, XGBoost $\rightarrow$ Trees ensemble

Transformer

- at the base of state-of-the-art in many NLP tasks
- usually take advantage of large amounts of unlabeled data...
- ...but we do not investigate this path due to the nature of the dataset.

Classifications: models

- Transformer
- MLP
- Decision Tree
- Random Forest
- XGBoost
- Linear SVM

Transformer, MLP $\rightarrow$ Neural Networks

Random Forest, XGBoost $\rightarrow$ Trees ensemble

Transformer

- at the base of state-of-the-art in many NLP tasks
- usually take advantage of large amounts of unlabeled data...
- ...but we do not investigate this path due to the nature of the dataset.

Segmentations

Example:
"CARCINOMA DUTTALE INFILTRANTE NOS G3, MULTIFOCALE. INVASIONE VASCOLARE ...
TNM 2010 VII edizione: pT1c (m), pN1 mi (sn). PARAMETRI BIOLOGICI: ER (clone SP1): POSITIVO 100% INTENSITA' DELLA COLORAZIONE: MARCATA PgR (clone 1E2): POSITIVO 70% INTENSITA' DELLA COLORAZIONE: MARCATA Ki67 (clone MIB1): 30% c-erbB-2 (policlonale A 0485): POSITIVO > 10% INTENSITA' DELLA COLORAZIONE: MODERATA. SCORE 2+."

Steps:

1. find a marker of the variable — `ki67\s?.{,10} :`
2. take a window of characters
3. cut after a foreign marker — `cerb|pgr|\ser\s|progest|estrog`
4. find a number in the window — `(\d?\d?\d)%`

Segmentations

Example:
"CARCINOMA DUTTALE INFILTRANTE NOS G3, MULTIFOCALE. INVASIONE VASCOLARE ...
TNM 2010 VII edizione: pT1c (m), pN1 mi (sn). PARAMETRI BIOLOGICI: ER (clone SP1): POSITIVO 100% INTENSITA' DELLA COLORAZIONE: MARCATA PgR (clone 1E2): POSITIVO 70% INTENSITA' DELLA COLORAZIONE: MARCATA Ki67 (clone MIB1): 30% c-erbB-2 (policlonale A 0485): POSITIVO > 10% INTENSITA' DELLA COLORAZIONE: MODERATA. SCORE 2+."

Steps:

1. find a marker of the variable `ki67\s?.{,10}` :
2. take a window of characters
3. cut after a foreign marker `cerb|pgr|\ser\s|progest|estrog`
4. find a number in the window `(\d?\d?\d)%`

Segmentations

Example:
"CARCINOMA DUTTALE INFILTRANTE NOS G3, MULTIFOCALE. INVASIONE VASCOLARE ...
TNM 2010 VII edizione: pT1c (m), pN1 mi (sn). PARAMETRI BIOLOGICI: ER (clone SP1): POSITIVO 100% INTENSITA' DELLA COLORAZIONE: MARCATA PgR (clone 1E2): POSITIVO 70% INTENSITA' DELLA COLORAZIONE: MARCATA Ki67 (clone MIB1): 30% c-erbB-2 (policlonale A 0485): POSITIVO > 10% INTENSITA' DELLA COLORAZIONE: MODERATA. SCORE 2+."

Steps:

1. find a marker of the variable — `ki67\s?.{,10}` :
2. take a window of characters
3. cut after a foreign marker — `cerb|pgr|\ser\s|progest|estrog`
4. find a number in the window — `(\d?\d?\d)%`

Segmentations

Example:
"CARCINOMA DUTTALE INFILTRANTE NOS G3, MULTIFOCALE. INVASIONE VASCOLARE ...
TNM 2010 VII edizione: pT1c (m), pN1 mi (sn). PARAMETRI BIOLOGICI: ER (clone SP1): POSITIVO 100% INTENSITA' DELLA COLORAZIONE: MARCATA PgR (clone 1E2): POSITIVO 70% INTENSITA' DELLA COLORAZIONE: MARCATA Ki67 (clone MIB1): 30% c-erbB-2 (policlonale A 0485): POSITIVO > 10% INTENSITA' DELLA COLORAZIONE: MODERATA. SCORE 2+."

Steps:

1. find a marker of the variable — `ki67\s?.{,10} :`
2. take a window of characters
3. cut after a foreign marker — `cerb|pgr|\ser\s|progest|estrog`
4. find a number in the window — `(\d?\d?\d)%`

Segmentations

Example:
"CARCINOMA DUTTALE INFILTRANTE NOS G3, MULTIFOCALE. INVASIONE VASCOLARE ...
TNM 2010 VII edizione: pT1c (m), pN1 mi (sn). PARAMETRI BIOLOGICI: ER (clone SP1): POSITIVO 100% INTENSITA' DELLA COLORAZIONE: MARCATA PgR (clone 1E2): POSITIVO 70% INTENSITA' DELLA COLORAZIONE: MARCATA Ki67 (clone MIB1): 30% c-erbB-2 (policlonale A 0485): POSITIVO > 10% INTENSITA' DELLA COLORAZIONE: MODERATA. SCORE 2+."

Steps:

1. find a marker of the variable — `ki67\s?.{,10}:`
2. take a window of characters
3. cut after a foreign marker — `cerb|pgr|\ser\s|progest|estrog`
4. find a number in the window — `(\d?\d?\d)%`

Results

Results for multi-class classification variables

Accuracy					
	Grading	**Stadio N**	**Stadio T**	**Sentinella Asportati**	**Sentinella Positivi**
Decision Tree	92.3%	95.3%	89.6%	75.6%	87.6%
Random Forest	**94.8%**	**97.6%**	97.0%	83.5%	**92.2%**
XGBoost	94.2%	97.2%	**97.6%**	**84.6%**	90.7%
SVM	93.4%	96.8%	94.4%	83.9%	91.1%
MLP	91.6%	93.1%	94.0%	71.7%	86.8%
Transformer	94.0%	93.4%	94.6%	70.1%	86.8%
num classes	3	5	5	4	3

Observations

- *Grading*, *Stadio-N*, *Stadio-T* are easier
- *Random Forest* and *XGBoost* have the best results

Results for binary classification variables

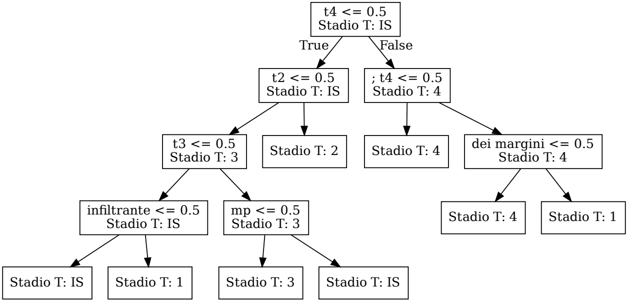
Tipo-T variable

	Accuracy	Precision	Recall	F1
Decision Tree	96.5%	48.2%	70.2%	57.1%
Random Forest	96.5%	47.8%	**77.2%**	**59.1%**
XGBoost	96.4%	46.1%	61.4%	52.6%
SVM	96.4%	47.1%	71.9%	56.9%
MLP	**96.9%**	**53.1%**	45.6%	49.1%
Transformer	95.6%	40.2%	68.4%	50.6%

Observations

- Highly unbalanced classification
- Random Forest has the best results

Predictions interpretability: example tree



Observations

- trees were greedily constructed...
- ... but learned trees are very compact

Segmentations results

Metrics

- Extracted: % of predictions on the total
- Accuracy on extracted
- Hit: % of correct predictions on the total

	Extracted	Acc. on extracted	Hit
recettori estrogeni	96.3%	71.3%	68.7%
recettori progestin	96.8%	89.9%	87.0%
mib1	100%	18.2%	18.2%
cerb	96.1%	83.6%	80.4%
ki67	97.7%	81.9%	80.0%
dimensioni	76.9%	73.7%	56.7%

Problems

- *mib1* has very few labels
- *dimensioni* has different unit of measure

Conclusions

Answering the question

Question

Is it possible to **predict variables** from these reports?

Answer

It's possible to predict them with high accuracy in most cases.

Whether these accuracies are enough can be verified only using these models in practice.

Recap

- *RF* and *XGBoost* are **good classifiers** for some variables
- *NN* do not obtain better results
- in [1] complex models were not much better than simpler ones
- regex-based algorithm is a **good baseline** for many variables
- further **improvements** are possible

[1] S. Martina, L. Ventura and P. Frasconi, "Classification of Cancer Pathology Reports: A Large-Scale Comparative Study"

Conclusions

Considerations

- the aim is not to replace human experts
- **two registers** that proceed at different speed

Future works

- predict **presence** of the variable
- access to **similar datasets**
- numbers extraction as a **learning** problem
- focus on **interpretable** models

Thank you for the attention.
Are there any questions?

Macro F1 on the test set for the multi-class classification variables.

Macro F1					
	Grading	**Stadio N**	**Stadio T**	**Sentinella Asportati**	**Sentinella Positivi**
Decision Tree	91.5%	92.3%	84.5%	66.3%	75.5%
Random Forest	**94.1%**	**96.5%**	93.3%	73.1%	80.4%
XGBoost	93.5%	95.9%	**94.4%**	**77.2%**	81.2%
SVM	92.5%	95.3%	89.8%	75.4%	**82.0%**
MLP	90.6%	87.7%	82.8%	62.6%	69.0%
Transformer	93.3%	89.7%	91.6%	64.3%	72.8%
num classes	3	5	5	4	3

Important tokens for Random Forest

	Grading	Stadio N	Stadio T
1	g2	n1	t2
2	g3	n0	p t2
3	g1	n1 a	p t1
4	nos g3	n2 a	t1
5	(g2	n2	ptis
6	scarsamente differenziato	p n1	t1 c
7	scarsamente	p n1 a	p t1 c
8	infiltrante nos g3	n3 a	: p t2
9	(g3	n0 (	t2 ,
10	moderatamente differenziato	n3	infiltrante

Data format

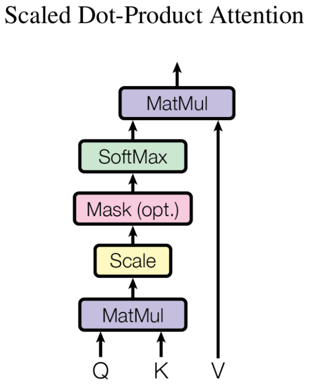
Transformer:

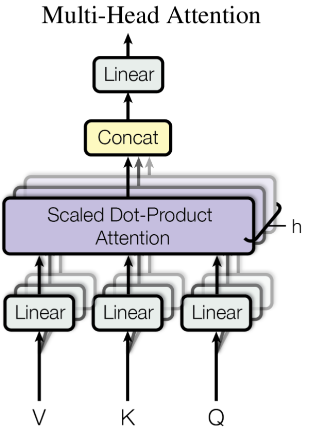
- tokens indices
- vector of integers

Other models:

- bag of *n*-grams of tokens
- vector of booleans

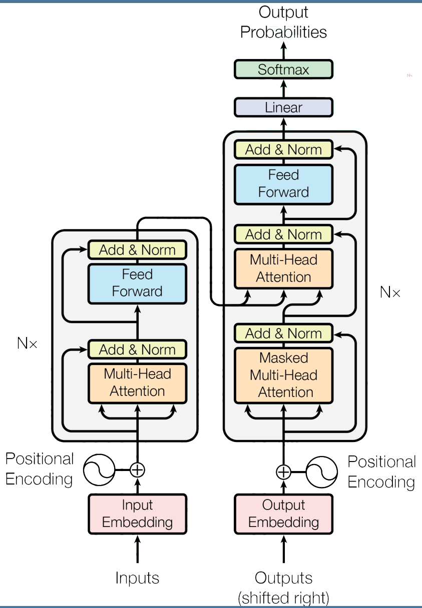
Attention


Scaled Dot-Product Attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V

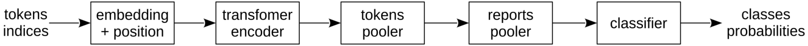

Multi-Head Attention
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q

Transformer architecture

- we use only the left part (encoder)



Pipeline of Transformer-based model



Numbers segmentation as position regression

Ground truth location can be ambiguous