UNIVERSITÀ
DEGLI STUDI
FIRENZE

Convolutional Seq2Seq Learning

For Spelling Correction

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin (2017)

Giovanni Carlucci – Deep Learning

UNIVERSITÀ DEGLI STUDI FIRENZE

Indice

1 Idea

2 Approccio

3 Architettura

4 Risultati – Ricerca Iperparametri

5 Risultati – Addestramento Completo

6 Risultati – Modello BPE

7 Riflessione e Conclusione

UNIVERSITÀ DEGLI STUDI FIRENZE

1 Idea

Il Task: Correzione come Traduzione
Il progetto affronta la correzione del testo come un problema *Sequence-to-Sequence*
Obiettivo: Mappare una sequenza "errata" in una sequenza "corretta"
→ Uso di RNN fino al 2017

1 Idea

Il Task: Correzione come Traduzione
Il progetto affronta la correzione del testo come un problema *Sequence-to-Sequence*
Obiettivo: Mappare una sequenza "errata" in una sequenza "corretta"
→ Uso di RNN fino al 2017

Superare i limiti delle RNN
Parallelizzazione Totale: Eliminazione della dipendenza temporale tipica delle LSTM - l'input è processato simultaneamente

- **Efficienza Computazionale:** Training drasticamente più veloce e utilizzo ottimale della GPU
- **Contesto Gerarchico:** Migliore cattura delle dipendenze a lungo raggio → struttura piramidale delle CNN

UNIVERSITÀ DEGLI STUDI FIRENZE

Indice

1 Idea

2 Approccio

3 Architettura

4 Risultati – Ricerca Iperparametri

5 Risultati – Addestramento Completo

6 Risultati – Modello BPE

7 Riflessione e Conclusione

2 Approccio

1

Creazione Architettura

- Ricreazione dell'architettura proposta in Gehrig et. Al. (2017)

2

Ricerca Iperparametri Ottimali

- Ricerca Iperparametri con Range derivati da letteratura (Optuna)
- 1° Ricerca con Range ampi
- 2° Ricerca con Range adattati a tentativi con migliore performance

3

Addestramento Completo

- Migliori tre tentativi

2 Approccio

1

Creazione Architettura

- Ricreazione dell'architettura proposta in Gehrig et. Al. (2017)

2

Ricerca Iperparametri Ottimali

- Ricerca Iperparametri con Range derivati da letteratura (Optuna)
- 1° Ricerca con Range ampi
- 2° Ricerca con Range adattati a tentativi con migliore performance

3

Addestramento Completo

- Migliori tre tentativi

Problemi

- Spazio disponibile sulla GPU (Papavero) variabile
- Stabilità della VPN (limite di ore in funzione)
- Librerie non aggiornate per MPS (MacOS)
- Risorse gratuite non sufficienti su Google Colab

2 Approccio

1

Creazione Architettura

- Ricreazione dell'architettura proposta in Gehrig et. Al. (2017)

2

Ricerca Iperparametri Ottimali

- Ricerca Iperparametri con Range derivati da letteratura (Optuna)
- 1° Ricerca con Range ampi
- 2° Ricerca con Range adattati a tentativi con migliore performance

3

Addestramento Completo

- Migliori tre tentativi

Problemi

- Spazio disponibile sulla GPU (Papavero) variabile
- Stabilità della VPN (limite di ore in funzione)
- Librerie non aggiornate per MPS (MacOS)
- Risorse gratuite non sufficienti su Google Colab

Soluzioni

- Ricerca finché Crash
- Addestramento con salvataggio dei pesi e caricamento successivo
- Codice eseguibile indipendentemente dal dispositivo

UNIVERSITÀ DEGLI STUDI FIRENZE

Indice

1 Idea

2 Approccio

3 Architettura

4 Risultati – Ricerca Iperparametri

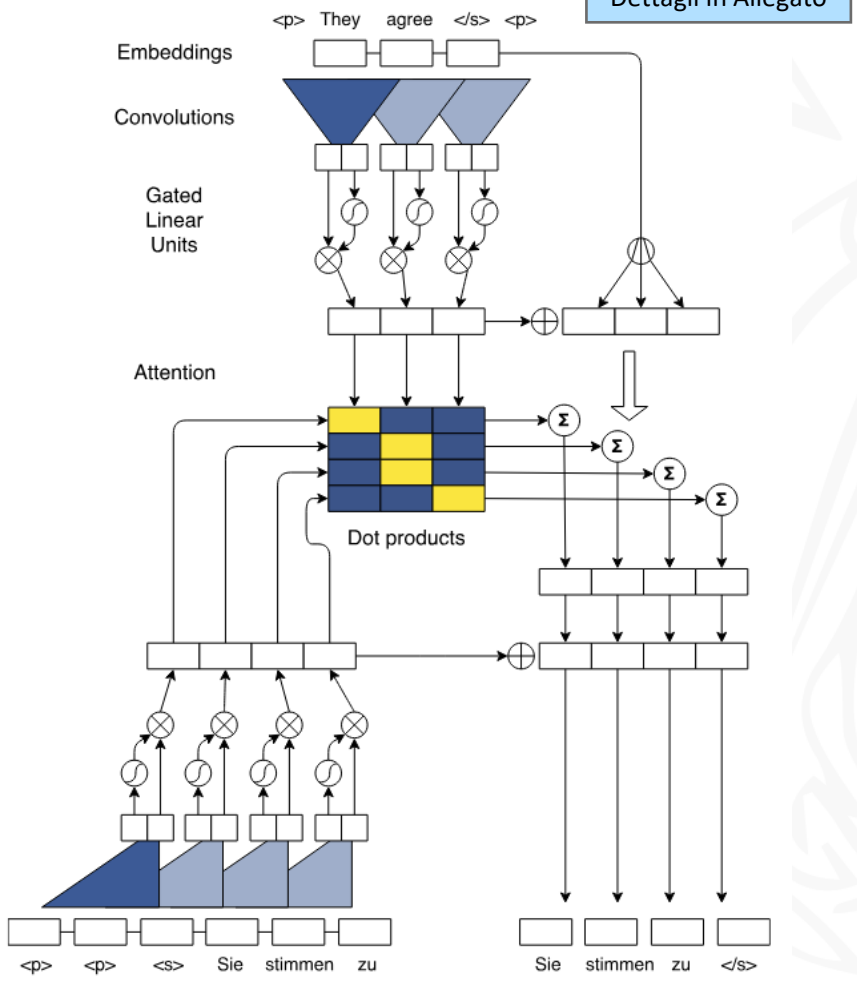
5 Risultati – Addestramento Completo

6 Risultati – Modello BPE

7 Riflessione e Conclusione

3 Architettura

Dettagli in Allegato



UNIVERSITÀ DEGLI STUDI FIRENZE

Indice

1 Idea

2 Approccio

3 Architettura

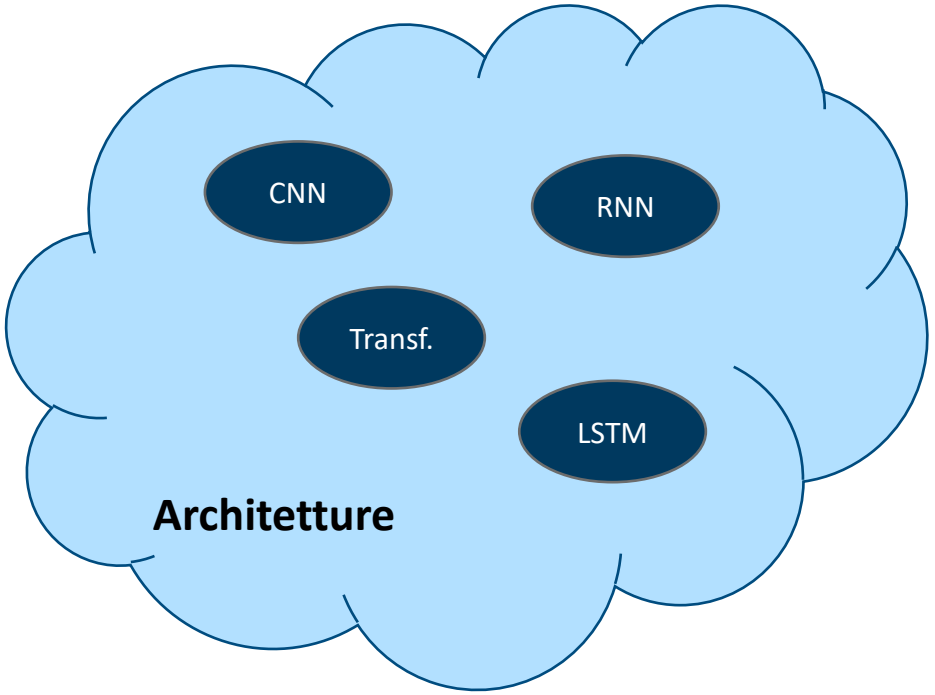
4 Risultati – Ricerca Iperparametri

5 Risultati – Addestramento Completo

6 Risultati – Modello BPE

7 Riflessione e Conclusione

4 Risultati – Ricerca Iperparametri

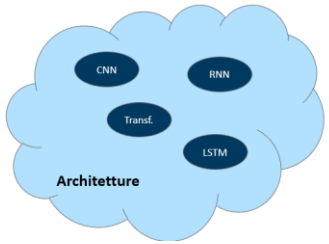


1° ottimizzazione:

learning_rate	= [0.1, 0.4]
p_dropout	= [0.1, 0.5]
hidden_dim, embedding_dim	= [128, 1028]
encoderLayer	= [2, 16]
decoderLayer	= [2, 16]
batchSize	= [32, 128]
corruption_probability	= [0.05, 0.18] (+0.02 dynamic)
sentence_repetition	= [2, 5]

UNIVERSITÀ DEGLI STUDI FIRENZE

4 Risultati – Ricerca Iperparametri



Fase optim:

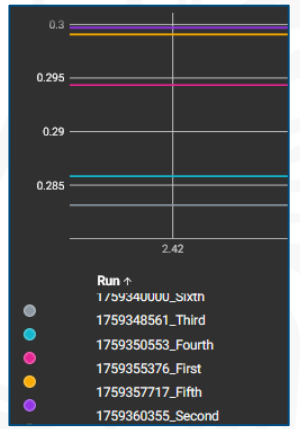
learning_rate	= [0.1, 0.4]
p_dropout	= [0.1, 0.5]
hidden_dim, embedding_dim	= [128, 1028]
encoderLayer	= [2, 16]
decoderLayer	= [2, 16]
batchSize	= [32, 128]
corruption_probability	= [0.05, 0.18] (+0.02 dynamic)
entence_repetition	= [2, 5]

Primi 7 risultati **ChrF > 35**
Dopo < 17

2° ottimizzazione:

learning_rate	= [0.2, 0.35]
p_dropout	= [0.18, 0.27]
hidden_dim, embedding_dim	= [430, 550]
encoderLayer	= [2, 6]
decoderLayer	= [9, 12]
batchSize	= [80, 110]
corruption_probability	= [0.06, 0.1] (+0.02 dynamic)
sentence_repetition	= [2, 3]

- Selezione delle 3 migliori configurazioni ChrF = [60,66]
- Selezione della 6° con ChrF 54.82 con learning rate più basso



4 Risultati – Ricerca Iperparametri

Name	Punteggio CHRF	File di Configurazione	BPE	batchSize	beamWidth	dataSet_Sentence	dataSet_probability	dataSet_repetition	decoderLayer	embedding_dim	encoderLayer	fixedNumberOfInputElements	hidden_dim	kernel_width	learning_rate	maximumlearningRateLimit	nestorovsMomentum	p_dropout	patience	renormalizationLimit	validationSet
First	**65,5042**	1761288824	0	86	5	20000	0,065947671	2	9	515	2	175	515	3	0,333374801	0,001	99	0,219018798	0	0,1	0,01
Second	**63,2923**	1761234404	0	92	5	20000	0,063779094	2	11	496	2	175	496	3	0,324367641	0,0001	0,99	0,209005332	0	0,1	0,01
Third	**60,2726**	1761282972	0	88	5	20000	0,066496007	2	12	511	2	175	511	3	0,345133123	0,0001	0,99	0,212388457	0	0,1	0,01
X_Sixth	**54,8159**	1761227621	0	87	5	20000	0,083899428	2	10	481	2	175	481	3	0,234369611	0,0001	0,99	0,212579084	0	0,1	0,01

UNIVERSITÀ DEGLI STUDI FIRENZE

Indice

1 Idea

2 Approccio

3 Architettura

4 Risultati – Ricerca Iperparametri

5 Risultati – Addestramento Completo

6 Risultati – Modello BPE

7 Riflessione e Conclusione

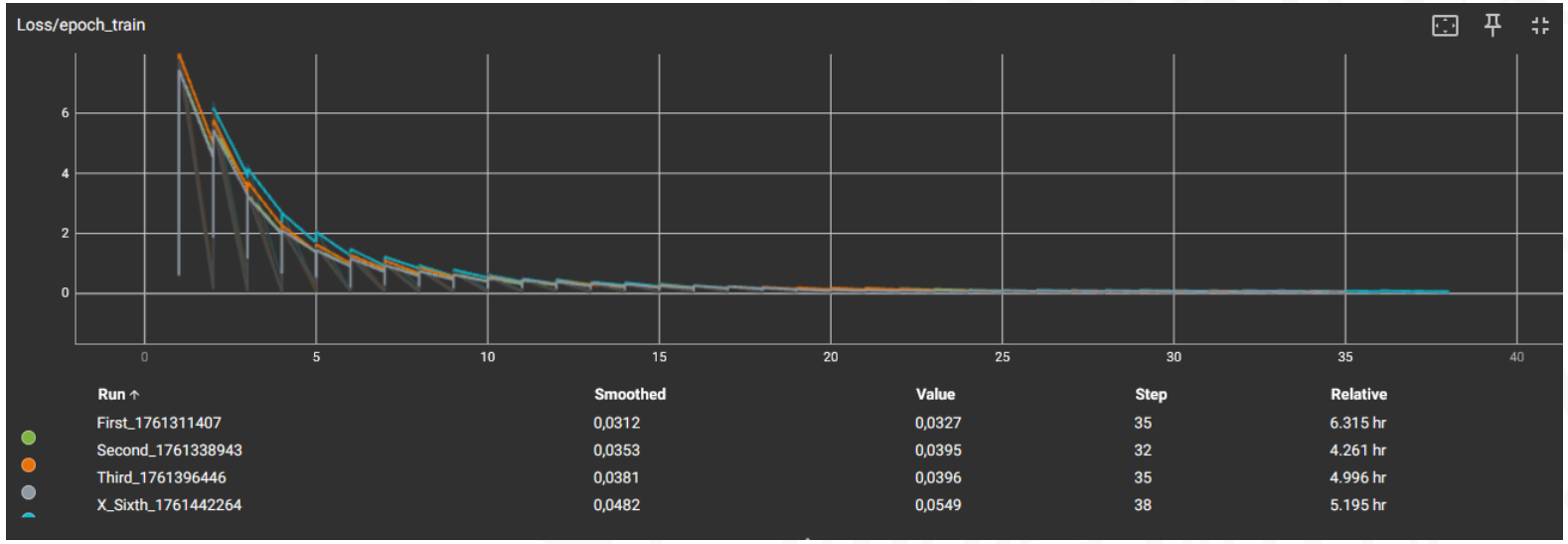
4 Risultati – Addestramento Completo

Addestramento

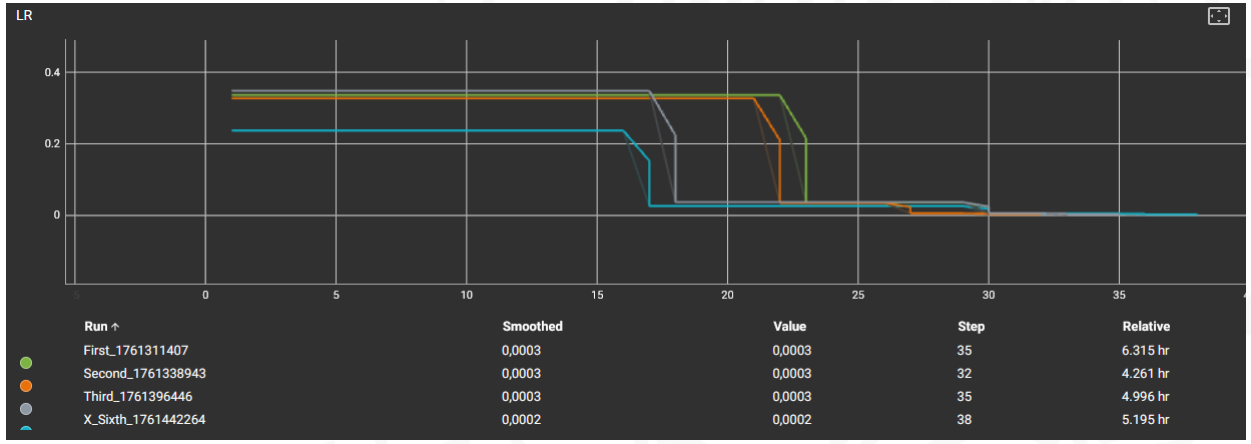
➔ Nota: Aggiunta di Patience = 2 rispetto a Paper, dato il dataset molto limitato (20.000 frasi)

Risultati

- Nessuna grande differenza nelle configurazioni
- Addestramento più lungo per la prima configurazione (+21.6% vs media 5.192h)

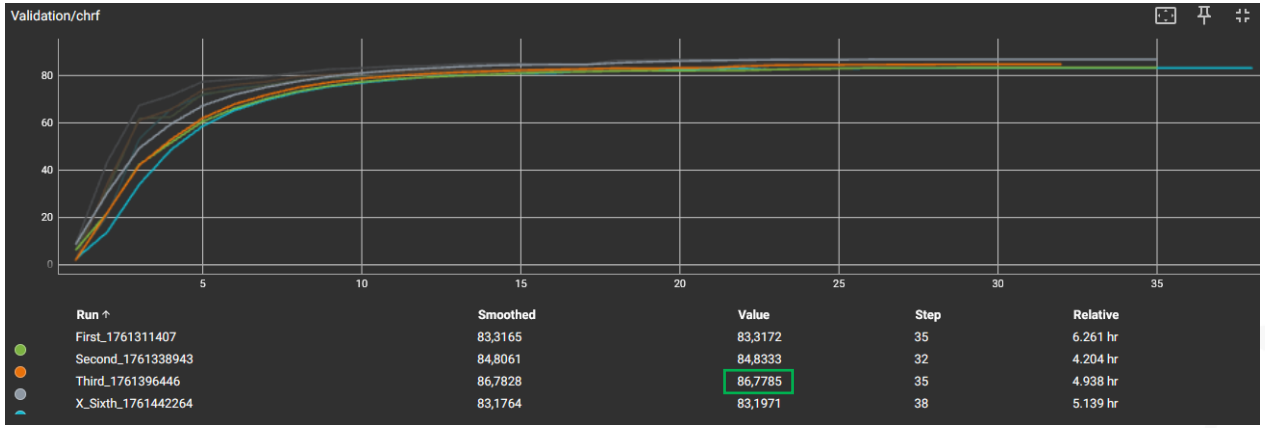


Run	Smoothed	Value	Step	Relative
First_1761311407	0,0312	0,0327	35	6.315 hr
Second_1761338943	0,0353	0,0395	32	4.261 hr
Third_1761396446	0,0381	0,0396	35	4.996 hr
X_Sixth_1761442264	0,0482	0,0549	38	5.195 hr

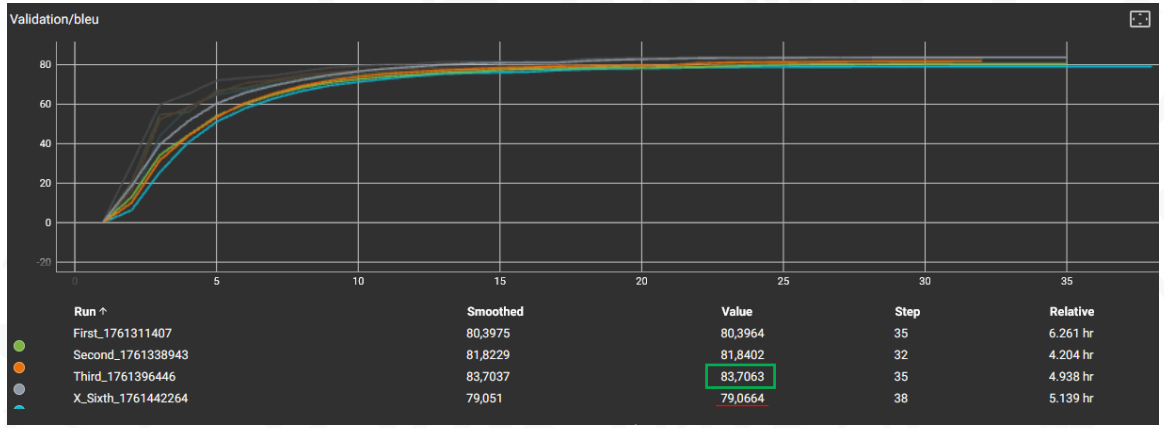


Run	Smoothed	Value	Step	Relative
First_1761311407	0,0003	0,0003	35	6.315 hr
Second_1761338943	0,0003	0,0003	32	4.261 hr
Third_1761396446	0,0003	0,0003	35	4.996 hr
X_Sixth_1761442264	0,0002	0,0002	38	5.195 hr

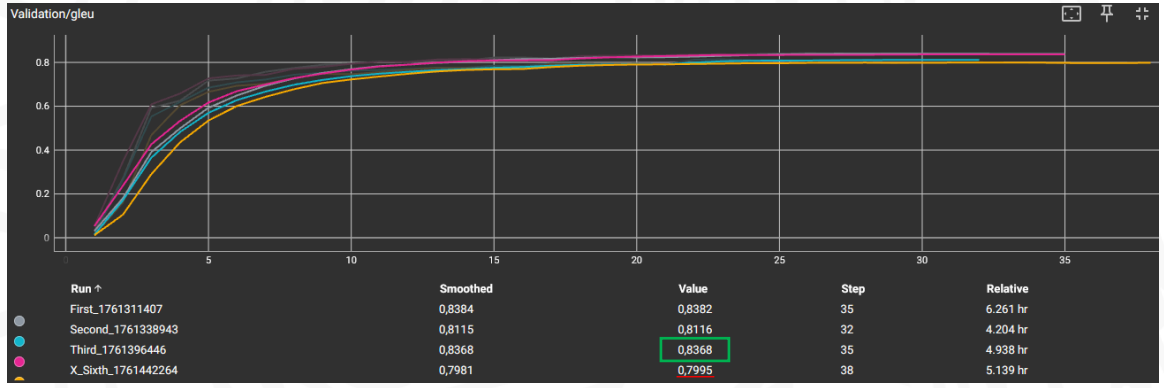
5 Risultati – Addestramento Completo



Run	Smoothed	Value	Step	Relative
First_1761311407	83,3165	83,3172	35	6.261 hr
Second_1761338943	84,8061	84,8333	32	4.204 hr
Third_1761396446	86,7828	86,7785	35	4.938 hr
X_Sixth_1761442264	83,1764	83,1971	38	5.139 hr



Run	Smoothed	Value	Step	Relative
First_1761311407	80,3975	80,3964	35	6.261 hr
Second_1761338943	81,8229	81,8402	32	4.204 hr
Third_1761396446	83,7037	83,7063	35	4.938 hr
X_Sixth_1761442264	79,051	79,0664	38	5.139 hr

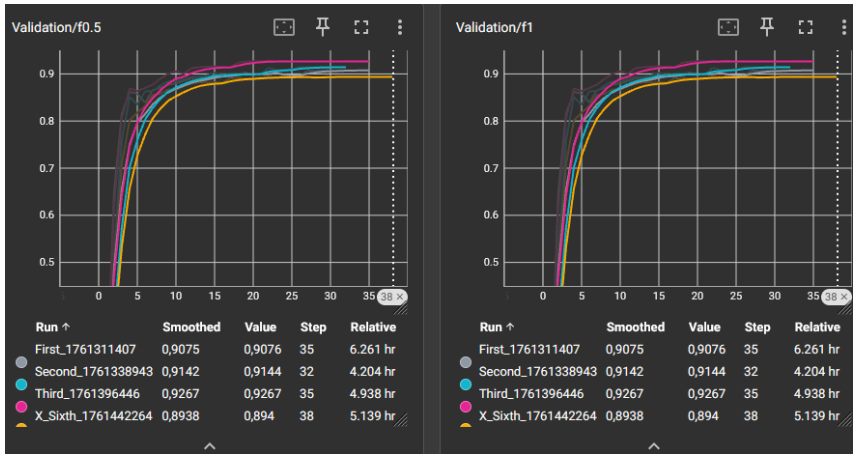


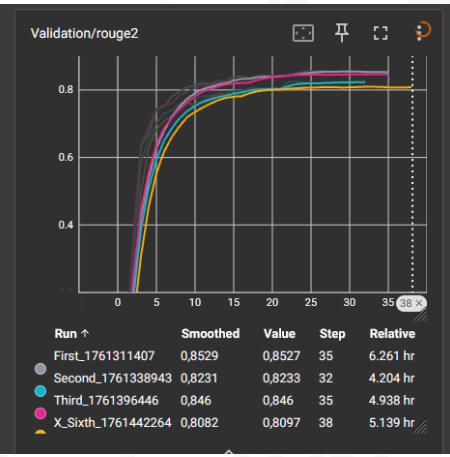
Run	Smoothed	Value	Step	Relative
First_1761311407	0,8384	0,8382	35	6.261 hr
Second_1761338943	0,8115	0,8116	32	4.204 hr
Third_1761396446	0,8368	0,8368	35	4.938 hr
X_Sixth_1761442264	0,7981	0,7995	38	5.139 hr

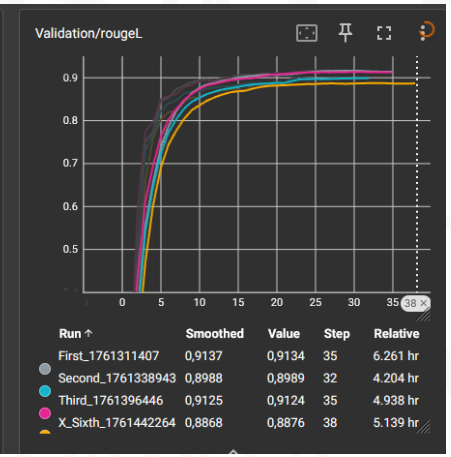
Risultati

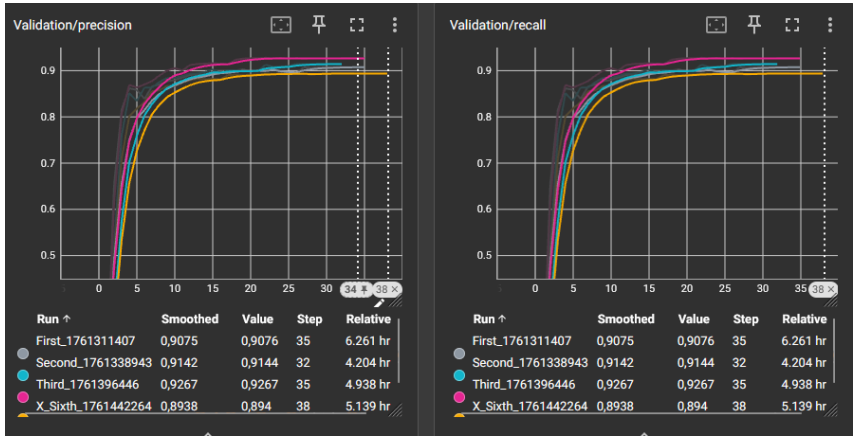
- Terza configurazione più consistente nei risultati
- Differenze minime tra le configurazioni
- Risultati molto alti rispetto a paper di riferimento
  - Overfitting?

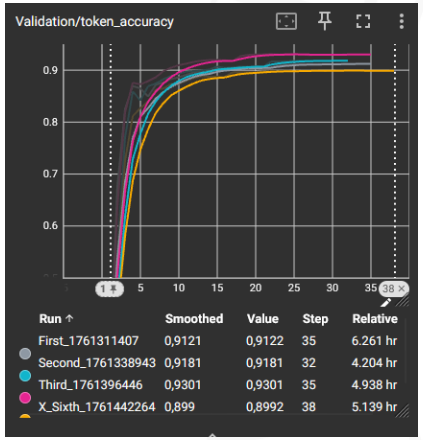
5 Risultati – Addestramento Completo









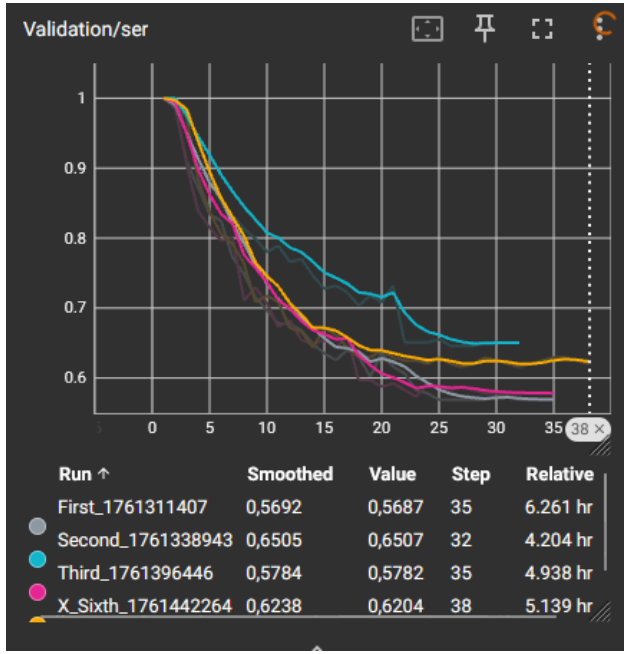


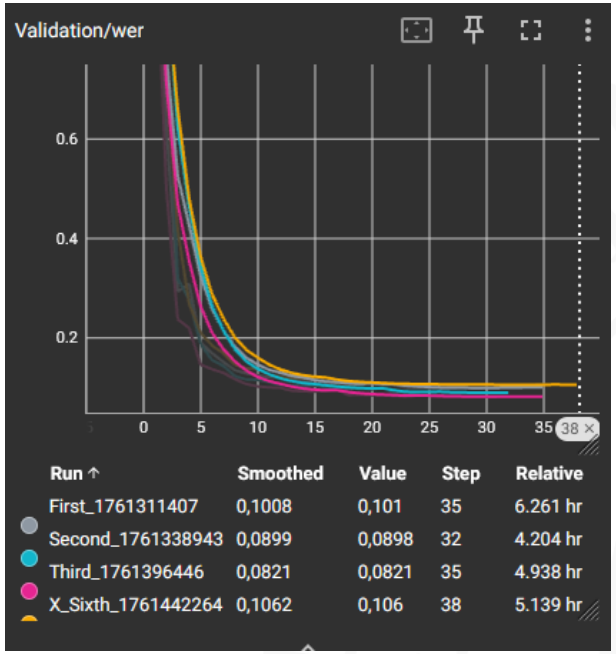
Overfitting?

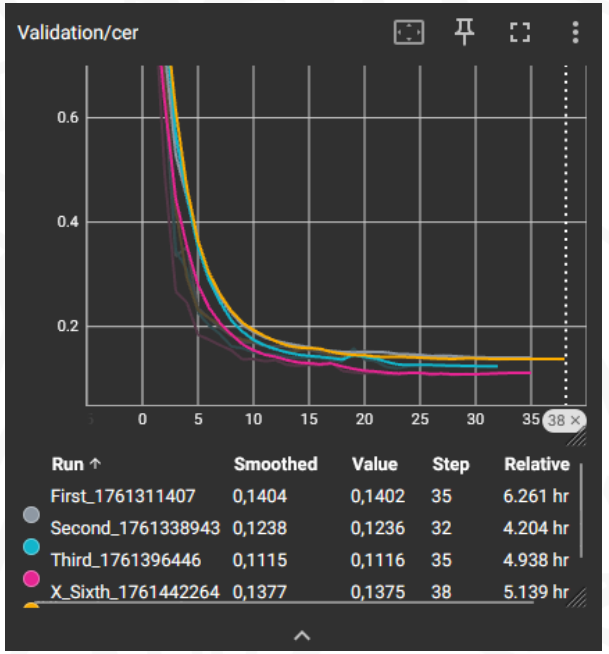
- Tutte le metriche indicano valori intorno al 0.9

➔ Decisamente alto, il modello sembra funzionare, tuttavia rimangono dubbi sulla qualità dato il dataset piccolo

5 Risultati – Addestramento Completo







Overfitting!

- Sentence Error Rate molto alta: Oltre la metà delle frasi contengono almeno un errore
- Word e Character Error Rate più contenute ma per un modello di correzione sono troppo alte considerando anche il livello di corruzione incluso simile

6 Risultati – Addestramento Completo

Configurazione vincente

- La terza configurazione mostra:
  - Tempi di addestramento sotto la media
  - BLEU/GLEU/ChrF migliore
  - Risultati migliori per tutte le metriche di precisioni, analisi di n-grammi etc.
  - Risultati più consistenti nell'analisi dell'error rate
- Dimostrazione che,anche con dimensioni e risorse contenute è possibile raggiungere una rappresentazione semantica delle parole e frasi
- Problema: Va testato su dataset più grandi per trovare reale efficienza e rendere utilizzabile il modello

BPE	0
batchSize	88
beamWidth	5
dataSet_Sentence	20000
dataSet_probability	0,066496007
dataSet_repetition	2
decoderLayer	12
embedding_dim	511
encoderLayer	2
fixedNumberOfInputElements	175
hidden_dim	511
kernel_width	3
learning_rate	0,345133123
maximumlearningRateLimit	0,0001
nestorovsMomentum	0,99
p_dropout	0,212388457
patience	0
renormalizationLimit	0,1
validationSet	0,01

UNIVERSITÀ DEGLI STUDI FIRENZE

Indice

1 Idea

2 Approccio

3 Architettura

4 Risultati – Ricerca Iperparametri

5 Risultati – Addestramento Completo

6 Risultati – Modello BPE

7 Riflessione e Conclusione

6 Risultati – Modello BPE

2° ottimizzazione:

learning_rate	= **[0.3, 0.36]**
p_dropout	= **[0.23, 0.35]**
hidden_dim, embedding_dim	= **[700, 950]**
encoderLayer	= [4, 6]
decoderLayer	= [6, 12]
batchSize	= **[45, 75]**
corruption_probability	= **[0.1, 0.16]** (+0.02 dynamic)
sentence_repetition	= [3, 4]

Punteggio	70,685	65,3307	52,4068
Numero	1762126535	1762122343	1761837057
BPE	1	1	1
batchSize	50	49	46
beamWidth	5	5	5
dataSet_Sentence	20000	20000	20000
dataSet_probability	0,077879684	0,076699624	0,065530666
dataSet_repetition	2	2	2
decoderLayer	3	2	3
embedding_dim	864	780	587
encoderLayer	2	3	5
fixedNumberOfInputElements	175	175	175
hidden_dim	864	780	587
kernel_width	3	3	3
learning_rate	0,393645582	0,398619933	0,307195301
maximumlearningRateLimit	0,0001	0,0001	0,0001
nestorovsMomentum	0,99	0,99	0,99
p_dropout	0,459449062	0,460141209	0,488655796
patience	2	2	2
renormalizationLimit	0,1	0,1	0,1
validationSet	0,01	0,01	0,01

6 Risultati – Modello BPE

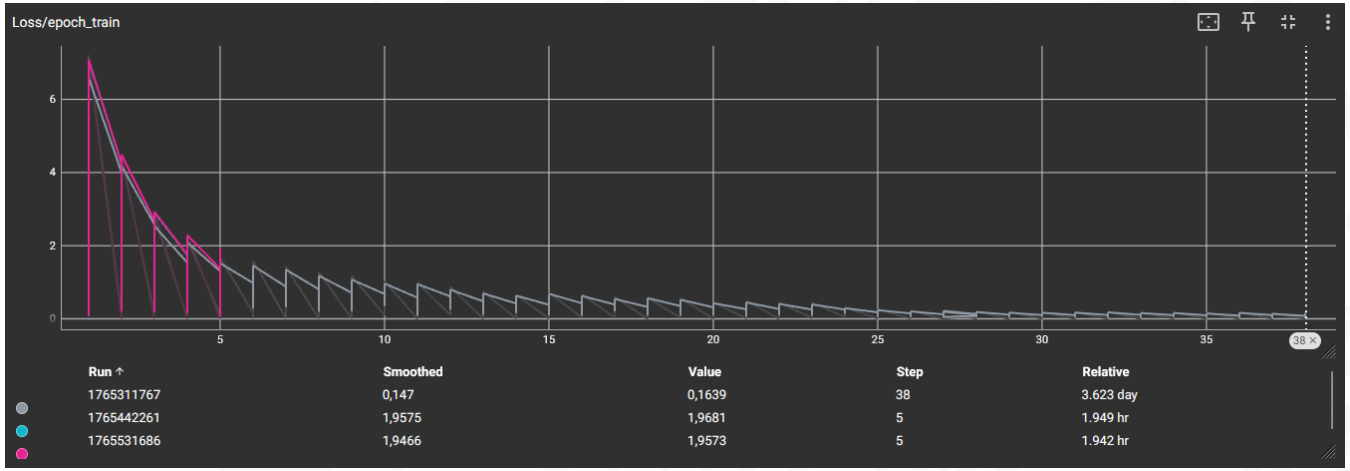
Addestramento

➔ Nota: Aggiunta di Patience = 2 rispetto a Paper, dato il dataset molto limitato (20.000 frasi)

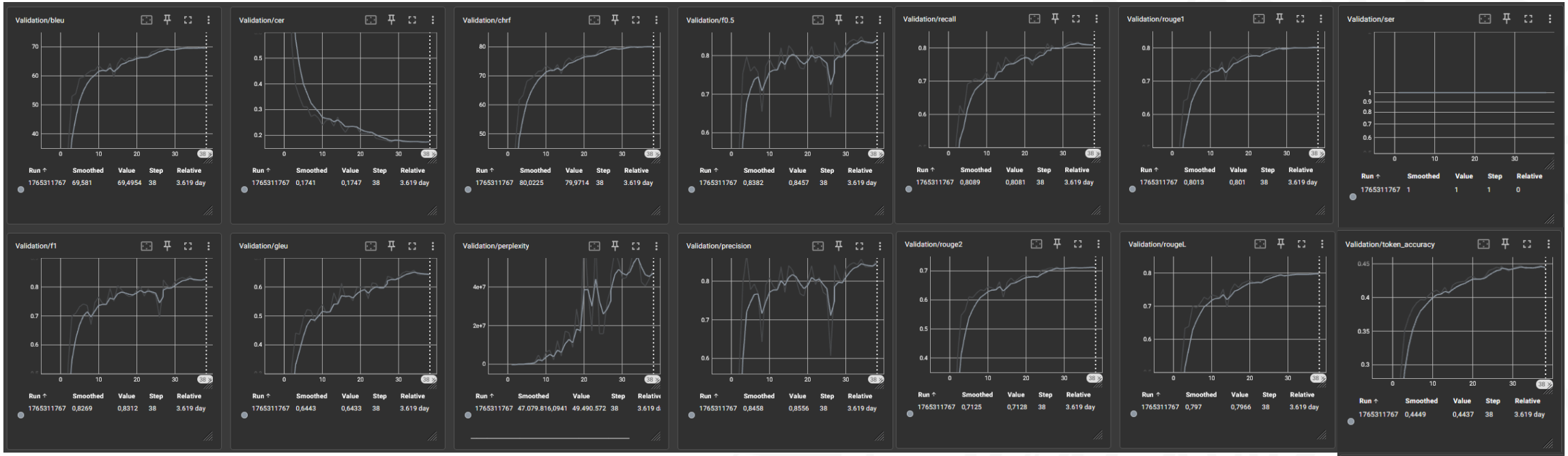
➔ Tokenizer Huggingface addestrato sui dati

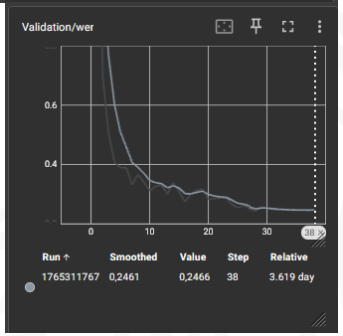
<u>Risultati</u>

- Problemi con risorse, solo un tentativo portato al termine
- Addestramento molto più lungo rispetto a WordTokenization



6 Risultati – Modello BPE





Overfitting!

- Problemi simili rispetto a WordTokenization
- Valori più bassi
- Sentence Error Rate 100%

UNIVERSITÀ DEGLI STUDI FIRENZE

Indice

1 Idea

2 Approccio

3 Architettura

4 Risultati – Ricerca Iperparametri

5 Risultati – Addestramento Completo

6 Risultati – Modello BPE

7 Riflessione e Conclusione

7 Riflessione e Conclusione

➢ **Efficienza dell'Architettura CNN:**
L'esperimento ha confermato il vantaggio computazionale delle ConvS2S. La parallelizzazione ha permesso tempi di addestramento ridotti rispetto alle architetture ricorrenti (RNN/LSTM), sfruttando meglio la GPU.

➢ **Il Collo di Bottiglia dei Dati (Data Constraint):**
L'overfitting riscontrato e l'alta *Sentence Error Rate* dimostrano che 20.000 frasi non sono sufficienti per addestrare da zero un modello così profondo. Il modello sembra tendere a "memorizzare" il rumore invece di apprendere le regole di correzione generali.

➢ **Analisi del fallimento BPE:**
L'approccio BPE richiede corpus massivi per generare un vocabolario di sottostringhe statisticamente rilevante. Su un dataset piccolo, il BPE non riesce a catturare le radici morfologiche, portando a performance peggiori rispetto alla tokenizzazione a parole/caratteri.

➢ **Sviluppi Futuri:**
- **Data Augmentation:** Incrementare il dataset sinteticamente a milioni di frasi.
- **Transfer Learning:** L'approccio moderno suggerirebbe l'uso di modelli pre-addestrati (es. Transformer/BERT) da perfezionare (fine-tuning) sul task di correzione, piuttosto che addestrare una rete da zero.

UNIVERSITÀ
DEGLI STUDI
FIRENZE

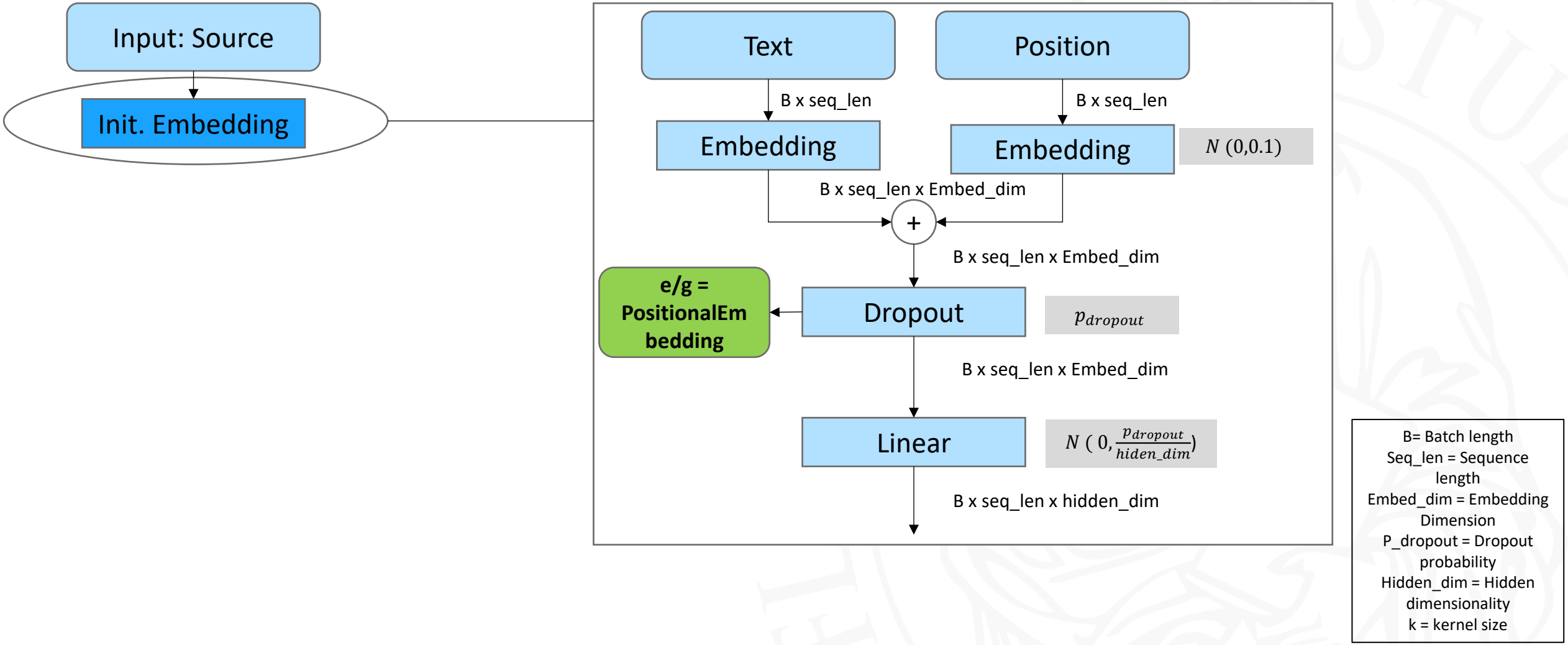
Convolutional Seq2Seq Learning

For Spelling Correction

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin (2017)

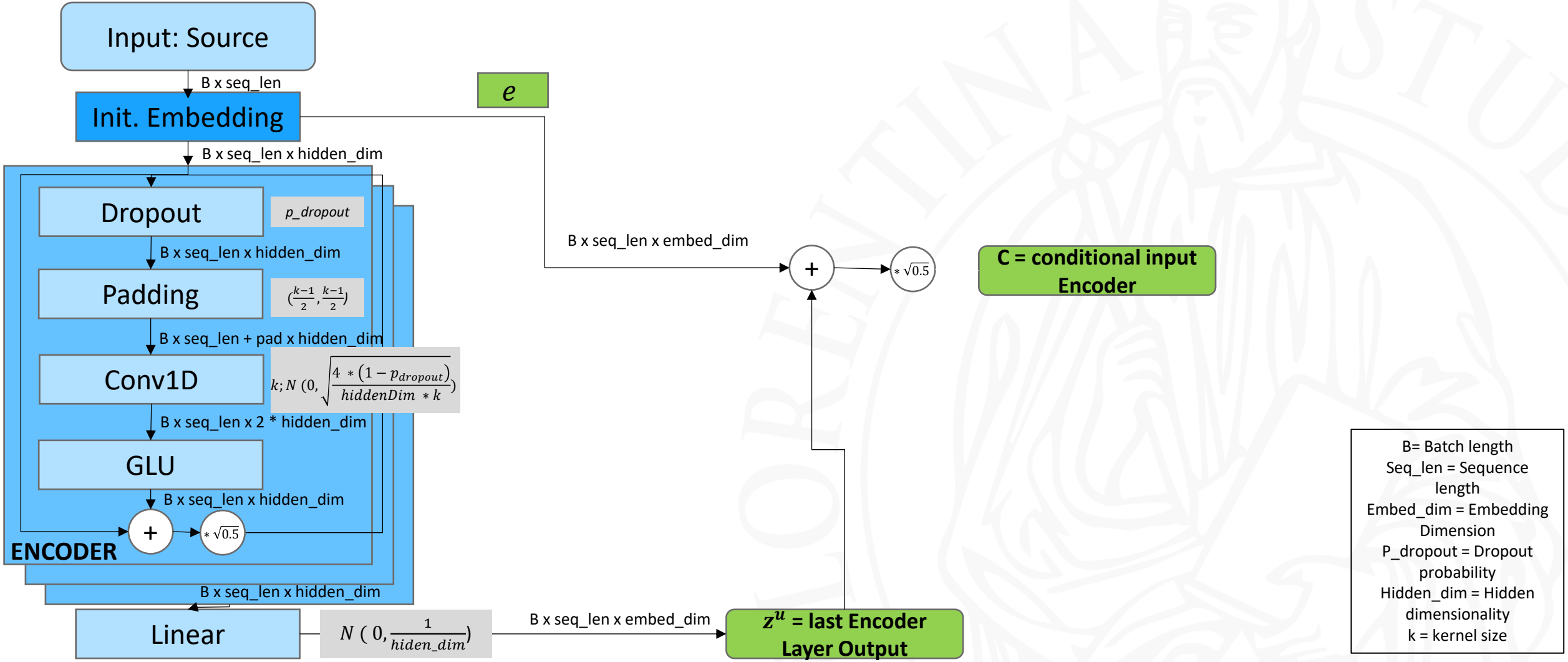
Giovanni Carlucci – Deep Learning

3 Architettura

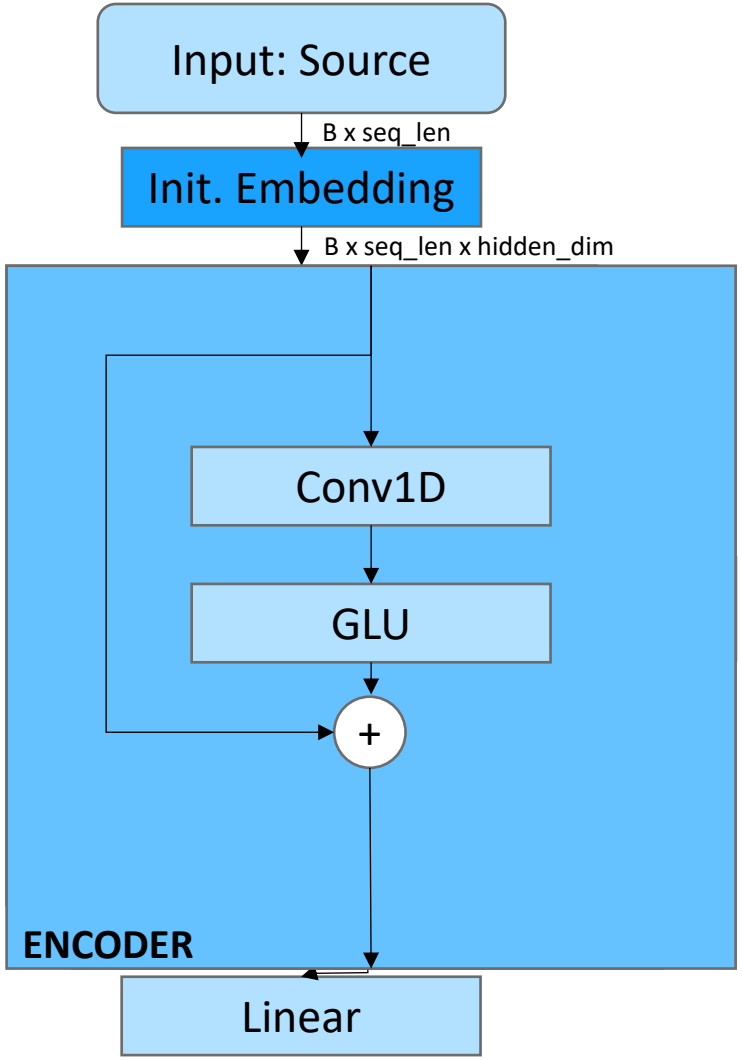


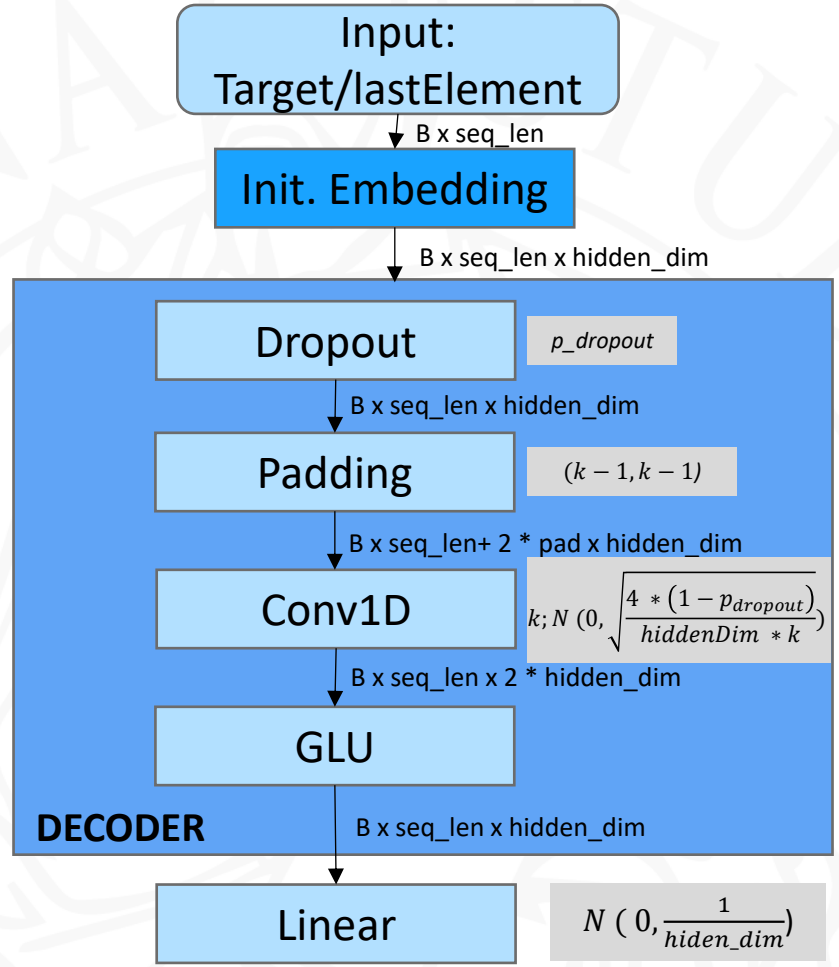
B= Batch length
Seq_len = Sequence length
Embed_dim = Embedding Dimension
P_dropout = Dropout probability
Hidden_dim = Hidden dimensionality
k = kernel size

3 Architettura

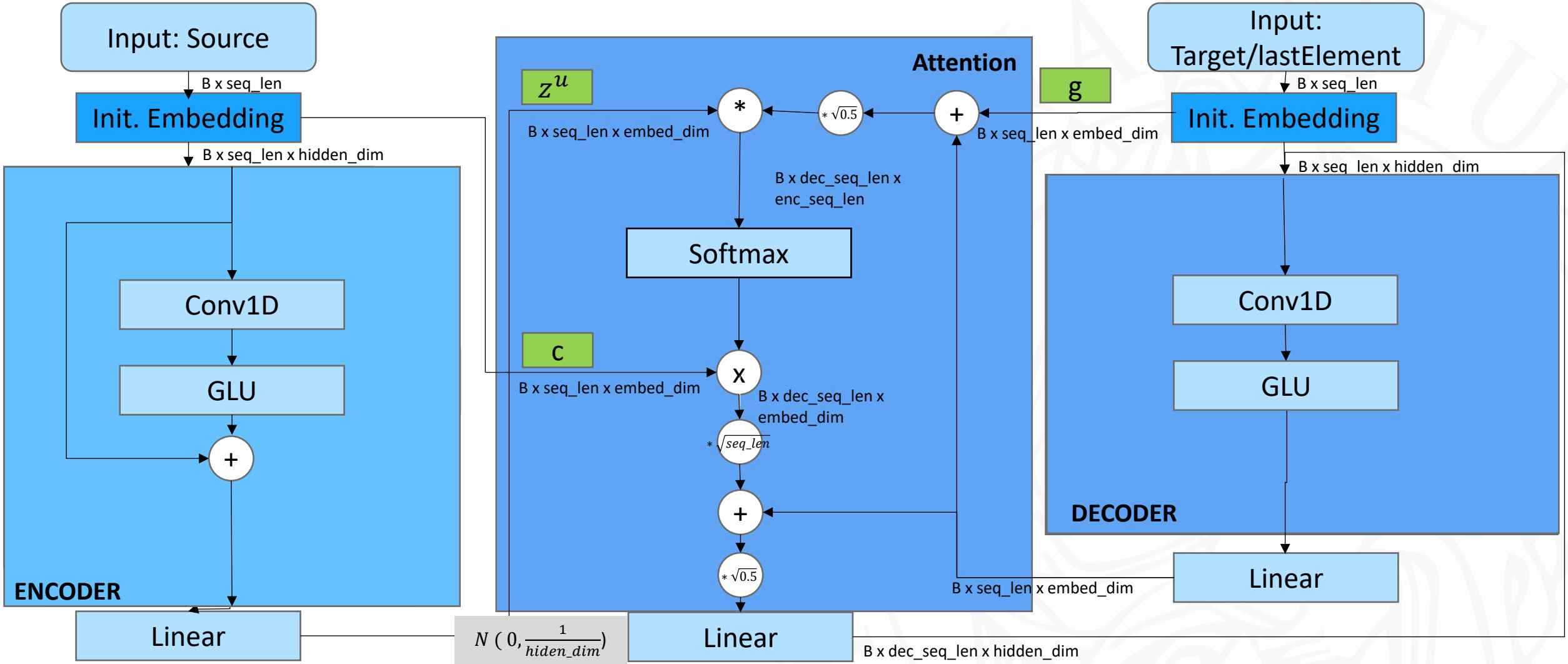


3 Architettura

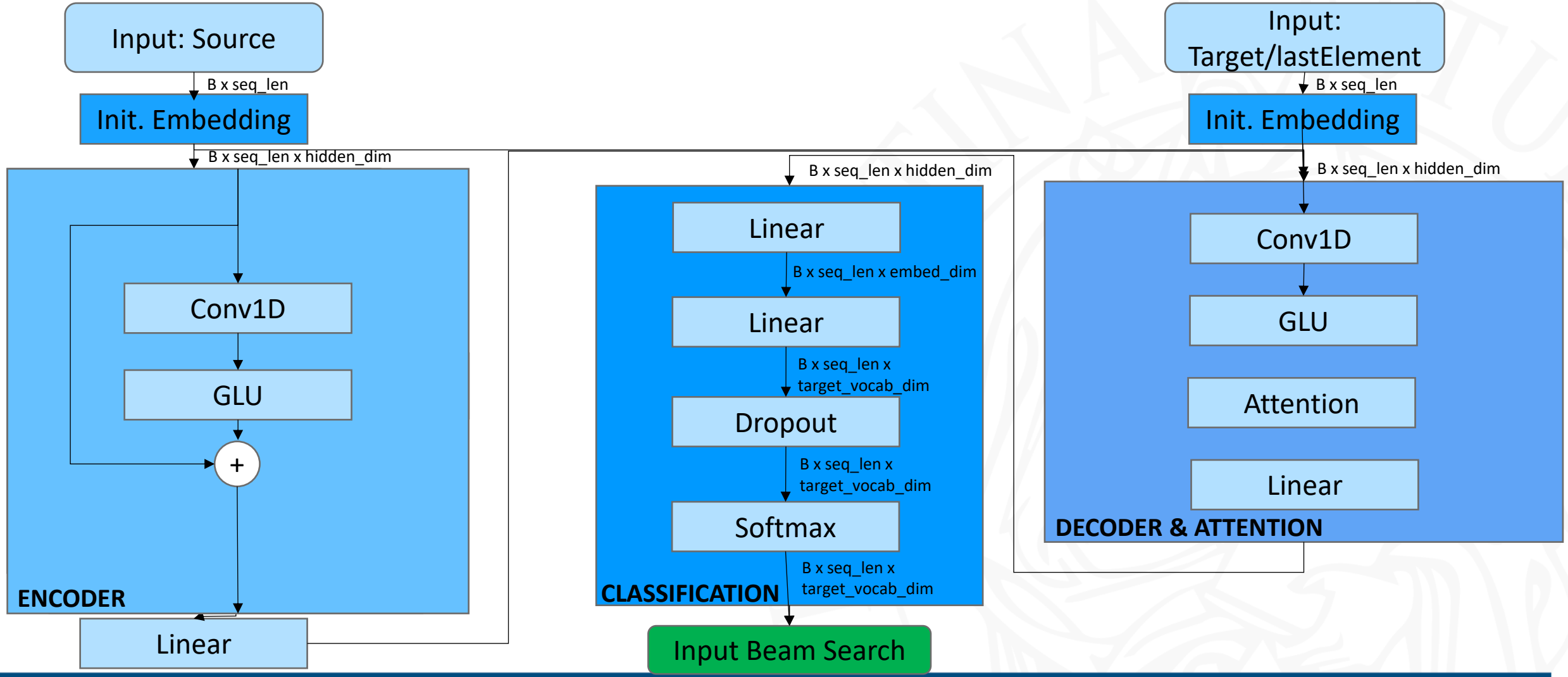




3 Architettura



3 Architettura



UNIVERSITÀ DEGLI STUDI FIRENZE

Results SecondOptim WordTokenization

Punteggio CHRF	File di Configurazione	BPE'	'batchSize'	'beamWidth'	'dataSet_Sentence'	'dataSet_probability'	'dataSet_repetition'	'decoderLayer'	'embedding_dim'	'encoderLayer'	'fixedNumberOfInputElements'	'hidden_dim'	'kernel_width'	'learning_rate'
65,5042	1761288824	0	86	5	20000	0,065947671	2	9	515	2	175	515	3	0,333374801
63,2923	1761234404	0	92	5	20000	0,063779094	2	11	496	2	175	496	3	0,324367641
60,2726	1761282972	0	88	5	20000	0,066496007	2	12	511	2	175	511	3	0,345133123
59,5501	1761284528	0	89	5	20000	0,067882593	2	12	520	2	175	520	3	0,349234525
58,5921	1761287227	0	87	5	20000	0,065760503	2	12	517	2	175	517	3	0,33227263
54,8159	1761227621	0	87	5	20000	0,083899428	2	10	481	2	175	481	3	0,234369611
52,5104	1761235844	0	107	5	20000	0,060465687	2	11	509	2	175	509	3	0,332351094
50,8719	1761264287	0	90	5	20000	0,064165302	3	12	443	2	175	443	3	0,310061362
49,61	1761229220	0	85	5	20000	0,085101217	2	11	486	2	175	486	3	0,245853882
48,3061	1761273409	0	98	5	20000	0,095621998	3	12	469	5	175	469	3	0,317869957
44,4201	1761231822	0	94	5	20000	0,081068845	2	11	504	3	175	504	3	0,22693989
43,8266	1761226174	0	82	5	20000	0,09160843	2	10	361	3	175	361	3	0,21594946
35,6449	1761261788	0	93	5	20000	0,07984424	2	12	470	4	175	470	3	0,228254374
35,1565	1761208792	0	90	5	20000	0,159920812	5	6	439	4	175	439	3	0,205476263
17,0612	1761205030	0	41	5	20000	0,071313806	3	10	139	6	175	139	3	0,201501337
16,163	1761212067	0	87	5	20000	0,1414883	2	2	563	15	175	563	3	0,272560172
12,7336	1761279450	0	102	5	20000	0,074530923	3	12	453	4	175	453	3	0,273247874
12,3295	1761219917	0	52	5	20000	0,052699027	2	2	358	13	175	358	3	0,291942263
11,7215	1761269862	0	95	5	20000	0,066458144	2	12	454	4	175	454	3	0,274425974
11,5871	1761230887	0	82	5	20000	0,086126995	2	13	362	2	175	362	3	0,268026254
11,0111	1761233379	0	101	5	20000	0,081901379	2	11	513	3	175	513	3	0,226394017
10,8956	1761216789	0	48	5	20000	0,074761363	4	4	245	9	175	245	3	0,116047087
10,7179	1761275418	0	103	5	20000	0,094110463	3	12	456	6	175	456	3	0,323267088
9,329	1761221139	0	70	5	20000	0,100392642	3	7	535	5	175	535	3	0,132974973
9,0038	1761271370	0	91	5	20000	0,087369644	3	9	448	5	175	448	3	0,201305256
6,9863	1761217736	0	75	5	20000	0,14302235	3	9	425	15	175	425	3	0,223870941
6,4004	1761215354	0	65	5	20000	0,083628789	2	12	279	10	175	279	3	0,243836804
5,9781	1761214674	0	86	5	20000	0,050068868	4	6	137	7	175	137	3	0,16319792
1,1059	1761218979	0	77	5	20000	0,110768103	4	12	213	2	175	213	3	0,164005449