Attention is all I need

José Ángel González[1]
jogonba2@dsic.upv.es

[1]Valencian Research Institute for Artificial Intelligence (VRAIN)
Universitat Politècnica de València

January 26, 2020

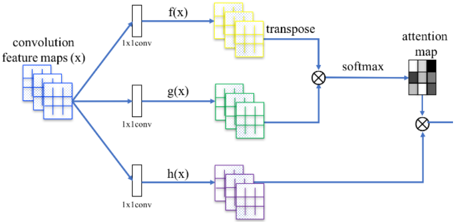
Who I am

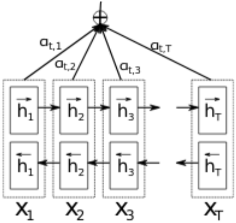
- Degree in Informatics Engineering (UPV).
- Master in Artificial Intelligence, Pattern Recognition and Digital Imaging (DSIC-UPV).
- PhD candidate (3rd year).
- Thesis: Deep Learning for Text Classification and Automatic Summarization.
- Sentiment/Emotional Analysis and Irony Detection in Social Media & Text Summarization.

Outline

Deep Learning for Sequence Modeling

- Dominance of convolutions and recurrences († 2017).
- Tendency to reduce the complexity of recurrent models:
  - GRU, Attention Mechanisms, Sequential computation ...
- Reducing the sequential computation to learn dependencies *independently of the positions* with Attention Mechanisms.
  - Transformers!





[0]Pictures from [34, 2]

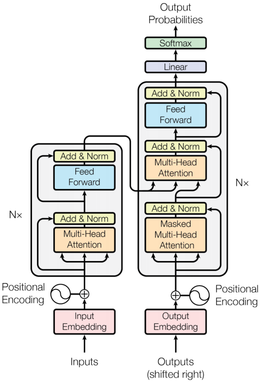
Transformers

Transformers for Text Classification

Attentional Extractive Summarization

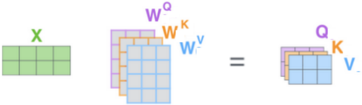
Transformers

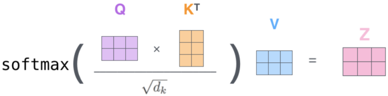
- Originally proposed as encoder-decoder model for NMT [28].
- Completely based on scaled dot-product attentions.
- More parallelizable (better suited for large and small datasets)
- The encoder is able to extract good text representations.



Self-Attention

- Learn representations from the all-vs-all interactions of the words e.g. $Z = XX^\intercal$
- Q, K, V projections for computing the self-attentions.
- Output Z is computed as a weighted sum of V.
- These weights are computed as a compatibility function of Q and K.
- Advantage: path length between X_i and X_j of $\mathcal{O}(1)$
- What happens with the word order?

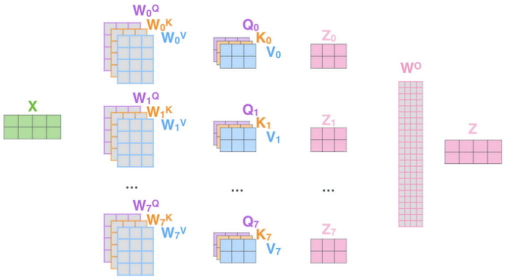




[0]Pictures from [1]

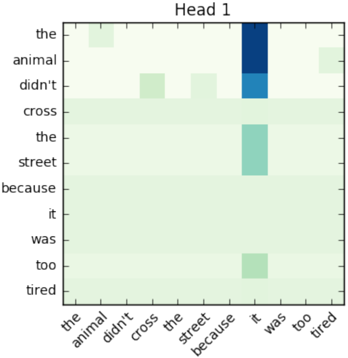
Multi-head Attention

- Self-Attention applied h times on the same input ($Z_{i \leq h}$ outputs are projected to Z)
- Allowing the model to jointly attend to information of different representation subspaces (e.g heads detecting word coreferences)
- More parallelizable (even d_q, d_k and d_v can be smaller).

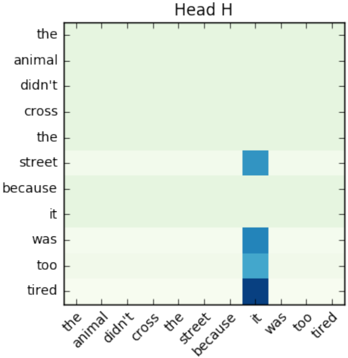


[0] Picture from [1]

Multi-head Attention



...

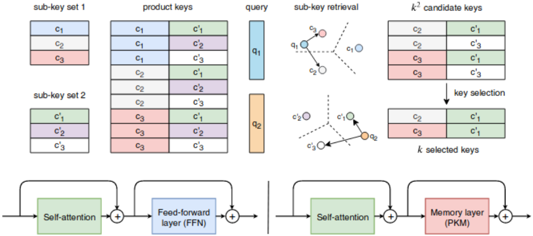


Insights

- Several strategies for adding positional information:
  - Absolute/Relative positions [26]
  - Heuristic rules / Positional embeddings [28]
- Optimization tricks for Deep Transformers:
  - Noam learning rate schedule / LAMB [28] [33]
  - Cross-Layer parameter sharing [17]
  - Factorized embeddings [17]
  - Gradient accumulation.
- Product Key Memory [16].

Product Key Memory

- To increase the network capacity without computational overhead [16].
- A 12-layered Transformer with one Product Key Memory can outperform a 24-layered Transformer.



[0] Picture from [16]

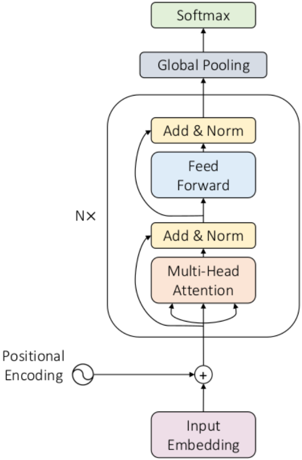
Transformers

Transformers for Text Classification

Attentional Extractive Summarization

Transformers for Text Classification

- Moving from uncontextual pre-trained embeddings [21] to contextualized finetuning models [19, 5, 24].
- What if we work in social network texts and non-english languages?
- But we want to profit the capacity of the Transformers:
  1. Contextualize pre-trained embeddings [9, 10].
  2. Adapt finetuning models to our task [8]

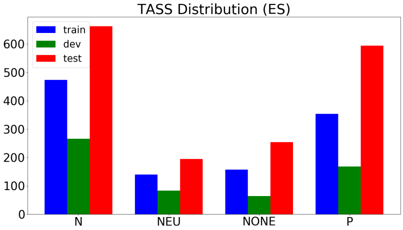


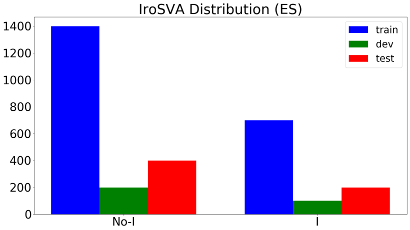
Transformers for Text Classification

- Evaluation of the Transformer Encoders for Spanish Twitter text classification tasks:
  - Sentiment Analysis (TASS 2019) [6, 10]
  - Irony Detection (IroSVA 2019) [3, 9]
- Without an extensive search of the hyper-parameters.
- Same model and resources for both tasks.
- Are they more powerful than other Deep Learning approaches?

Tasks

- TASS: Assigning a global polarity to each tweet on four classes $\mathbb{C} = \{\text{N, NEU, NONE, P}\}$
- IroSVA: Determine the ironic content of each tweet in two classes $\mathbb{C} = \{\text{No-I, I}\}$
- Spanish variants (Peru, Costa Rica, Cuba, Mexico and Uruguay)





Experimental Details

- Skip-gram word embeddings ($d_e = 300$ & 87M tweets)
- Fixed most of the hyper-parameters:
  - $L \in \{1, 2\}$, $h = 8$, $d_q = d_k = d_v = 64$ and $d_{ff} = d_e$
- Sine-Cosine Positional Encoding
- Weighted cross-entropy using $w(c) = \frac{\max_{c' \in \mathbb{C}} N(c')}{N(c)}$
- Adam + Noam Learning Rate Annealing
- Macro-F_1 for evaluating TASS and F_1 for IroSVA.
- Comparison with DAN [15] and Att-LSTM [30].

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{D}}[\mathcal{L}(f(x;\theta), y)] = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{|\mathbb{C}|} y_{ij}\ log\ f(x_i;\theta)_j\ w_j$$

Comparison

- TE outperforms DAN & Att-LSTM for all metrics.
- Same behavior for all the other Spanish variants.
- Positional relationships are not useful in these corpora.

	MP	MR	MF_1
DAN	47.66	48.46	47.94
Att-LSTM	50.00	48.14	48.83
1-TE-NoPos	52.80	**54.38**	**53.34**
1-TE-Pos	46.26	46.56	46.25
2-TE-NoPos	**52.85**	53.03	51.47
2-TE-Pos	47.31	48.79	47.71

	F_{10}	F_{11}	MF_1
DAN	85.78	69.79	77.78
Att-LSTM	81.05	66.05	73.54
1-TE-NoPos	**85.79**	**74.18**	**79.98**
1-TE-Pos	81.63	65.38	73.51
2-TE-NoPos	84.05	69.27	76.65
2-TE-Pos	82.64	62.83	72.74

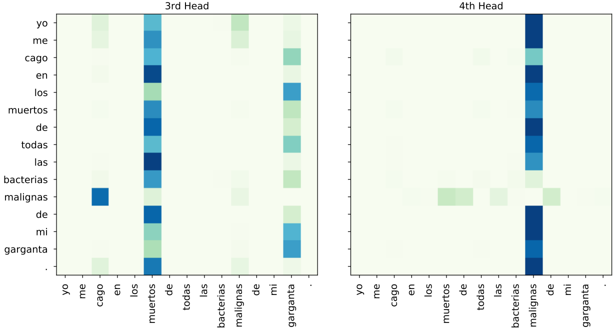
Comparison

	MF_1	MP	MR	Rank
ES	**50.70**	**50.50**	**50.80**	**1/9**
CR	49.60	49.80	49.30	2/9
PE	44.70	45.60	43.90	2/9
UY	51.50	49.70	53.60	2/7
MX	**50.10**	**49.00**	**51.20**	**1/9**

	CU	ES	MX	Avg	Rank
ELiRF-UPV	65.27	**71.67**	**68.03**	**68.32**	**1/18**
CIMAT	**65.96**	64.49	67.09	65.85	2/18
LDSE	63.35	67.95	66.08	65.79	3/18
JZaragoza	61.63	66.05	67.03	64.90	4/18
W2V Baseline	60.33	68.23	62.71	63.76	5/18

Attention Analysis

- The compatibility function between Q and K of the heads allows us to explain some properties captured by the system.
- Be A_{ijk} the attention that the word i puts in the word j in the attention head k:

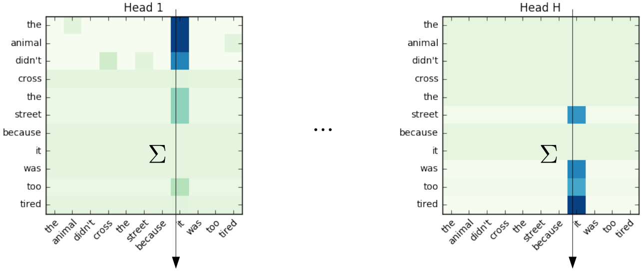


Attention Analysis

- We study the relationships captured by the multi-head self-attention mechanism.
- These relationships are task dependent:
  - Sentiment Analysis:
    - Word polarities
    - Polarity modifiers (shifters and intensifiers)
  - Irony Detection
    - Ironic attention heads
    - Impact of polarity words
    - Relevance of individual words
    - Word pair relationships

Attention Analysis

- How can we analyze if our system takes them into account?
- Computing the average attention that each word receives from all the other words for each head, averaged for all the occurrences of the word in a dataset.

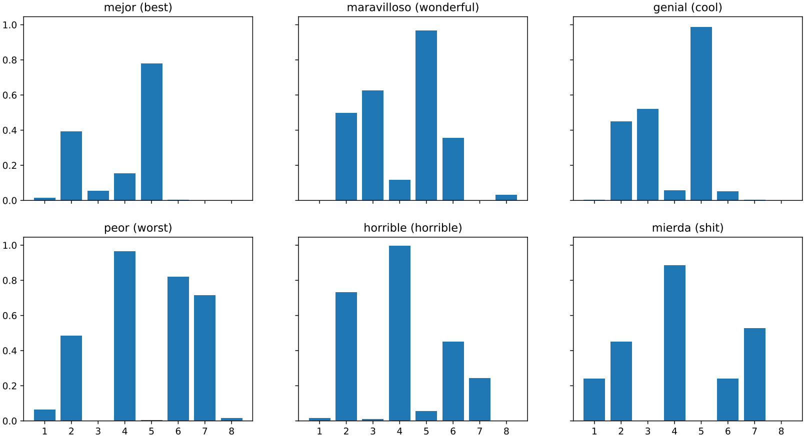


$$\alpha[\text{"}it\text{"}][1] = [...] + \alpha[\text{"}it\text{"}][1] \qquad \cdots \qquad \alpha[\text{"}it\text{"}][H] = [...] + \alpha[\text{"}it\text{"}][H]$$

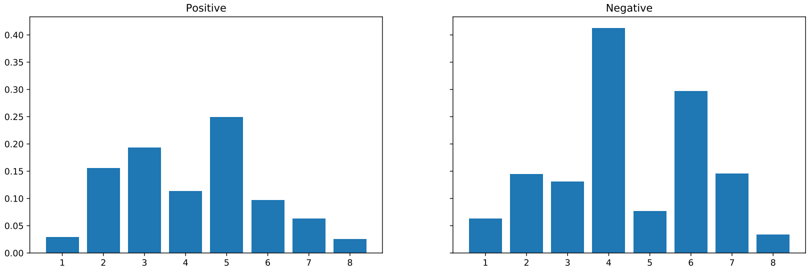
$$\vdots \qquad\qquad\qquad\qquad \vdots$$

$$\alpha[\text{"}it\text{"}][1] \qquad\qquad\qquad\qquad \alpha[\text{"}it\text{"}][H]$$

Sentiment: Detecting Word Polarity

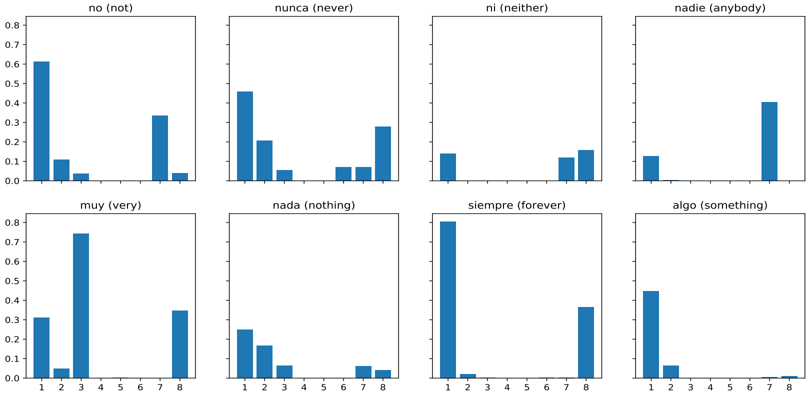


Sentiment: Detecting Word Polarity



- $\mathcal{C}(w) = \begin{Bmatrix} P & \alpha_{w4} \leq \alpha_{w5} \\ N & \alpha_{w4} > \alpha_{w5} \end{Bmatrix}$ classifies correctly the $74.75 \pm 3.17\%$ of the samples from ElHuyar lexicon.
- Only considering the distribution of two independent heads!

Sentiment: Detecting Polarity Modifiers



- Heads 4 and 5 do not react to non-polarity words.
- Head 1 seems to react to all polarity modifiers to a greater or lesser extent.
- The attention distributions for polarity modifiers are very similar.

Irony: Detecting Ironic Heads

- If we switch-off an attention head and the $F_1(1)$ decreases, that head is related with the Irony.
- Iterative process for masking attention heads.
- We explore incrementally the $2^h - 2$ combinations of maskings.
- The heads that appear in more combinations that worsen the $F_1(1)$ are the Ironic Heads.

Corpus	H_0	H_1	H_2	H_3	H_4	H_5	H_6	H_7
IroSVA	**16/18**	**11/18**	**13/18**	**10/18**	8/18	9/18	4/18	5/18

Irony: Impact of polarity words

- Which words are the most attended by H_{ironic} heads?
- We compute the average attention given by each head $k \in H_{ironic}$ to the word w, $\alpha[w][k]$
- If $\alpha[w][k] > \epsilon$, w is highly attended by k
- Polarity lexicons to analyze the polarity of these highly attended words.

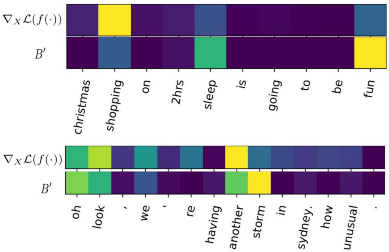
Head Set	Heads	$\lvert\alpha_w > \epsilon\rvert$	Negative	Positive	Ratio
	H_0	240	102	24	52.50%
	H_1	221	12	18	13.57%
H_{ironic}	H_2	73	22	8	41.09%
	H_3	603	140	47	31.01%
	Σ	1137	276	97	32.80%
	H_4	276	14	28	15.21%
	H_5	116	6	9	12.60%
H_{no_ironic}	H_6	281	41	11	18.50%
	H_7	237	14	18	13.50%
	Σ	910	75	66	15.50%

Irony: Impact of individual words

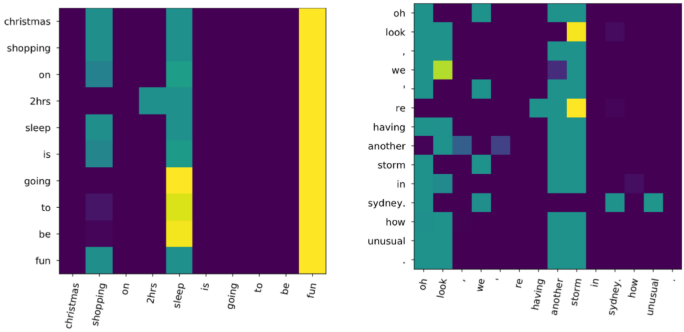
- Are there words that determine the irony?
- Two approaches:
  - Average attention given by H_{ironic}

    $$B \leftarrow \sum_{k \in H_{ironic}} softmax(\frac{f(X)_{Q_k} f(X)_{K_k}^\top}{\sqrt{d_k}}) \ ; \ B'_j \leftarrow \frac{1}{|X|} \sum_{i=1}^{|X|} B_{ij}$$
  - Euclidean norm of the input gradients:

    $$B'_j \leftarrow \|\nabla_X \mathcal{L}(f(X;\theta), y=1)_j\|$$



Irony: Word relationships



Language	Example	Top-5 Relationships
English	1	(sleep, fun), (christmas, fun) (going, fun), (2hrs, fun), (shopping, fun)
	2	(look, storm), (sydney, unusual), (', oh), (, , storm), (unusual, oh)
Spanish	1	(fallen, butter), (down, butter), (butter, side), (side, flat), (earth, side)
	2	(book, April Fool's joke), (pedro, book), ([clap emoji], book), (year, book), (seems, book)

Text Summarization

- Need to condense big amounts of unstructured information available in media platforms.
- Approaches to automatic summarization can be divided in:
  - Extractive
  - Abstractive
  - Mixed
- The most common human strategy to summarize documents consists in applying an ordered sequence of these approaches.
- The first step consists in focusing on the most relevant sentences (Extractive approach)
- Extractive Neural Summarization systems to the rescue!

Attentional Extractive Summarization

- Typical neural approaches states the problem as a sequential binary sentence classification problem.
- No corpora with this kind of labeling:
  - Suboptimal extractive oracles [4, 22, 18]
  - Reinforcement Learning [23, 35, 7, 32]
- Attentional approaches do not require a sentence labeling.
- They simplify the sequential classification problem.
- Based on the interpretation of Attentional Networks after being trained to solve a proxy binary classification task: distinguishing correct (document, summary) pairs.

Attentional Extractive Summarization

- All systems under this framework are based on two main ideas:

First idea

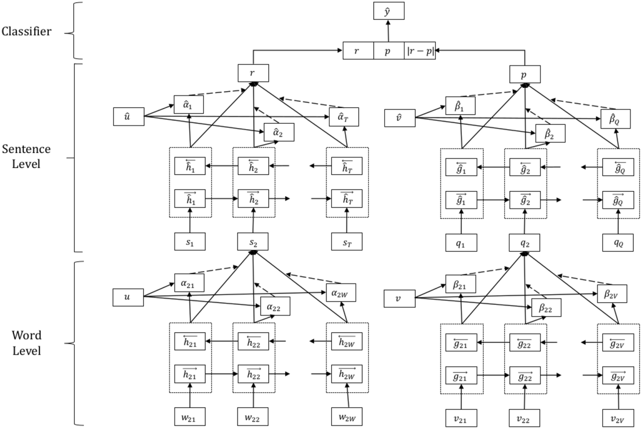
If we can say if a summary $\boldsymbol{y}$ is correct for a document $\boldsymbol{x}$ and we can look at the relevant sentences in $\boldsymbol{x}$ that led us to that decision, then we can build a summary $\widehat{\boldsymbol{y}}$, composed by the relevant sentences in $\boldsymbol{x}$, that is similar to the reference $\boldsymbol{y}$.

Second Idea

If $\boldsymbol{y}$ is a correct summary for a document $\boldsymbol{x}$, then $\boldsymbol{y}$ and $\boldsymbol{x}$ have similar semantics (similar representations) while if $\boldsymbol{w}$ is an incorrect summary for a document $\boldsymbol{x}$, then $\boldsymbol{w}$ and $\boldsymbol{x}$ have less similar semantics (less similar representations).

Siamese Hierarchical Attention Networks

- From these two ideas we proposed Siamese Hierarchical Attention Networks based on Attentional LSTM encoders [31, 12]

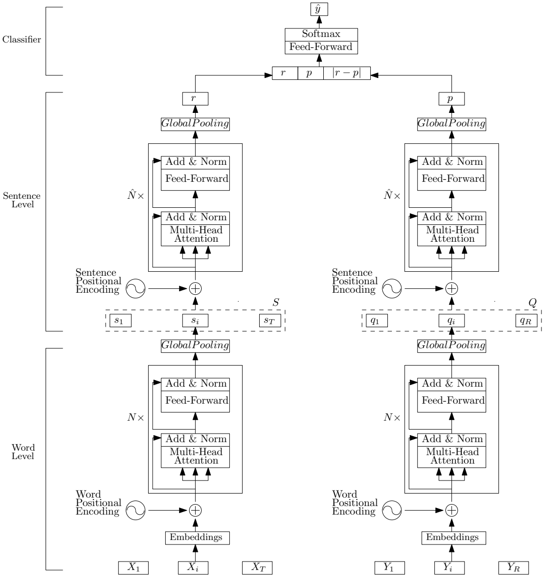


$$\mathcal{L}(\Theta) = \sum_{k=1}^{|\mathcal{D}|} \mathrm{L}(f(X_k, X'_k; \Theta), y = 1) + \mathop{\mathbb{E}}_{p(X_{j \neq k} | X_k)} [\mathrm{L}(f(X_k, X'_j; \Theta), y = 0)]$$

Siamese Hierarchical Transformer Encoders

- The attention mechanism can learn word-level relationships such as coreference [27], coherence [27], anaphora [29], etc.
- But also sentence-level relationships!
- We propose to use Transformer Encoders in a hierarchical way, to process sequences of sentences by replacing the Attentional LSTM of SHA-NN.
- The sentence relevances are implicitly computed by the multi-head self-attention mechanisms.

Siamese Hierarchical Transformer Encoders



$$\mathcal{L}(\Theta) = \sum_{k=1}^{|\mathcal{D}|} \mathrm{L}(f(X_k, X'_k; \Theta), y = 1) + \mathop{\mathbb{E}}_{p(X_{j \neq k} | X_k)} [\mathrm{L}(f(X_k, X'_j; \Theta), y = 0)]$$

Extractive Summarization with SHTE

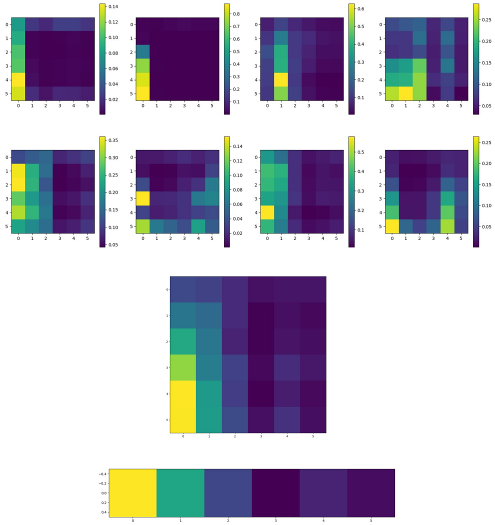
- Once the model is trained for the proxy task, the compatibility function at sentence level can be used to detect relevant sentences [11]
- Differently from SHA-NN [12], the compatibility function of these attention mechanisms do not assign a real value to each sentence:

Sentence Relevance Hypothesis

If a sentence s is greatly attended on average by all the other sentences s' for all the attention heads h, this sentence condenses a big part of the information of the sentences s', being thus, more relevant.

Extractive Summarization with SHTE

0. before you go , we thought you 'd like these ... if you want a face lift , and you have the time and money to make that happen , go for it.
1. but if you want a non-invasive alternative to surgery to help you get younger-looking skin , you need a good device.
2. the nuface trinity is a skin care device designed with interchangeable treatment attachments to help with facial stimulation and the reduction of fine lines and wrinkles.
3. in as little as five minutes a day , you can improve your facial contour and skin tone.
4. watch beauty expert jenny patinkin show you just how it easy it is to use this device.
5. looking for something else ? check out the video below to keep shopping !



Corpora

- Automatically collected corpora from newspaper domains:
  - CNN/DailyMail [14].
  - NewsRoom (BBC, Time, Bloomberg, Telegraph, ...) [13].
- The summaries of these corpora are the highlights written manually by the editors.
- Biased towards the first article sentences.

		Sentences		Words		Words/Sentence	
Corpus	Set	Articles	Summaries	Articles	Summaries	Articles	Summaries
CNN/DailyMail	Train	31.87	3.79	750.10	51.58	23.53	13.61
	Dev	26.77	4.11	737.06	57.57	27.53	14.00
	Test	27.11	3.88	745.59	54.65	27.51	14.07
NewsRoom	Train	29.91	1.40	773.57	30.37	25.86	21.65
	Dev	29.69	1.41	767.34	30.72	25.84	21.73
	Test	29.62	1.41	765.56	30.63	25.84	21.68

Results

- Similar results to PGen+Cov [25], SummaRunner [22], DQN [32] and Refresh [23].
- Best models take profit of pre-trained language models [18], Reinforcement Learning [7, 35] or word-length strategies [20]

System	Strategy	R-1	R-2	R-L
Lead-3	Ext	40.24	17.70	36.45
SHA-NN	Ext	39.99	17.75	36.27
SHTE	Ext	39.96	17.60	36.19
SummaRunner	Ext/OC	39.60	16.20	35.30
ECS-Ext	Ext/OC	41.70	18.60	37.80
BertSumEXT	Ext/OC	43.25	20.24	39.63
Refresh	Ext/RL	40.00	18.20	36.60
DQN	Ext/RL	39.40	16.10	35.60
Latent	Ext/RL	41.10	18.80	37.40
BanditSum	Ext/RL	41.50	18.70	37.60
ECS-Comp	Mix	40.90	18.00	37.40
Latent-Comp	Mix	36.70	15.40	34.30
PGen+Cov	Mix	39.53	17.28	36.38
CopyCat	Mix	39.15	17.60	36.17

Results

- Corpus divided in 3 subsets relating to the extractive degree (density)
- Extracting $k = 2$ better than $k = 3$ (not in the abstractive subset).
- Except ECS, our proposal outperforms all the neural models.

		NR-Ext			NR-Mix			NR-Abs			NR-Full		
System	Strategy	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Lead-3	Ext	51.98	47.85	51.20	25.62	13.00	22.30	14.57	2.62	11.73	30.66	21.09	28.35
SHA-NN-3	Ext	48.29	43.54	47.42	24.62	12.32	21.37	14.22	2.57	11.43	28.99	19.42	26.69
SHTE-3	Ext	48.62	43.35	47.65	24.76	12.20	21.43	14.33	2.53	11.51	29.19	19.37	26.81
Lead-2	Ext	57.87	53.03	56.83	28.60	14.33	24.46	15.68	2.77	12.35	33.98	23.30	31.14
SHAN-2	Ext	54.83	49.25	53.72	28.03	13.79	23.85	15.67	2.74	12.29	32.78	21.86	29.85
SHTE-2	Ext	53.97	47.87	52.59	27.78	13.41	23.56	15.57	2.67	12.22	32.38	21.25	29.40
ECS-Ext	Ext	69.40	64.30	68.30	31.90	16.30	26.90	17.20	3.10	13.60	39.50	27.90	36.26
PGen+Cov	Mix	39.10	28.00	36.20	25.50	11.00	21.10	14.70	2.30	11.40	26.43	13.76	22.90
TLM	Mix	53.30	44.20	50.10	28.10	12.10	23.00	18.50	3.90	14.70	33.30	20.06	29.26
FastRL	Mix	-	-	-	-	-	-	-	-	-	21.93	9.37	19.61
ECS-Comp	Mix	68.40	62.90	67.30	31.70	16.10	27.00	17.10	3.10	14.10	39.06	27.36	36.13

Convergence

- SHTE requires visiting more samples than SHA-NN until convergence, but few training hours ($4 - 5\times$ for NewsRoom)
- It obtains lower results in terms of *Acc* on (document, summary) pairs, but similar results for ROUGE (CE mismatch [23]).
- Faster than other approaches for CNN/DailyMail:
  - BanditSum: 76 hours (single GPU Nvidia Titan Rx)
  - DQN: 10 days (single GPU Nvidia 1080 Ti)
  - Refresh: 12 hours (single GPU)

Corpora	System	Epoch	Samples	Loss	Acc	Time (h)
CNN/DailyMail	SHA-NN	82	2,624,000	0.007	99.62 $\pm$0.10	3.51
	SHTE	13	4,160,000	0.209	91.92 $\pm$0.46	2.38
NewsRoom	SHA-NN	159	5,088,000	0.083	96.16 $\pm$0.11	6.45
	SHTE	9	5,760,000	0.230	90.61 $\pm$0.17	1.65

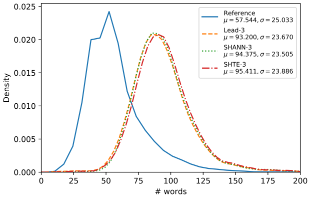
Analysis

- Two interesting observations on SHTE:
  - How affects the positional information?
    - Is it required positional information?
    - Is it required positional information on both levels?
    - What if only using sentence positional information?
  - What heads capture better the sentence relevance?
    - Are there individual heads related to condense information?
    - What about averaging heads?
- Is the word-length distribution obtained by SHA-NN & SHTE similar to the distribution on the reference summaries?

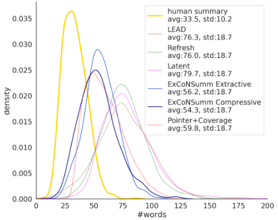
Analysis

	Head	Precision			Recall			F_1		
		R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
No Positional	1	24.28	7.92	21.79	45.06	15.15	40.38	29.75	9.80	26.68
	2	24.58	8.11	22.13	44.15	14.90	39.64	29.89	9.92	26.88
	3	**24.79**	7.97	**22.29**	43.48	14.42	38.98	29.64	9.62	26.62
	4	24.14	7.81	21.67	44.14	14.71	39.25	29.51	9.63	26.46
	5	24.49	7.94	22.02	43.40	14.39	38.90	29.61	9.66	26.58
	6	24.42	7.60	21.89	41.90	13.33	37.41	29.00	9.09	25.95
	Avg Heads	24.67	**8.23**	22.16	**45.45**	**15.53**	**40.73**	**30.20**	**10.15**	**27.10**
Sent Positional	1	27.79	11.07	25.21	51.31	20.78	47.34	34.76	13.82	31.51
	2	27.17	10.66	24.62	52.36	20.67	47.38	34.29	13.47	31.06
	3	29.19	11.71	26.53	51.74	20.86	46.98	35.83	14.39	32.55
	4	**29.84**	**12.09**	**27.15**	52.17	21.24	47.41	36.15	14.58	33.16
	5	29.12	11.87	26.48	**53.09**	**21.66**	**48.19**	36.03	14.68	32.74
	6	29.60	12.01	26.91	52.30	21.30	47.45	36.21	14.73	32.99
	Avg Heads	29.64	12.03	26.97	52.46	21.36	47.67	**36.36**	**14.76**	**33.37**
Sent-Word Positional	1	24.68	8.12	22.13	44.20	14.70	39.59	30.11	9.94	27.03
	2	23.91	7.84	21.51	44.34	14.87	39.79	29.45	9.74	26.47
	3	**25.83**	**9.69**	**23.32**	50.38	18.98	45.37	32.16	11.74	28.95
	4	23.59	7.66	21.18	43.99	14.61	39.39	28.98	9.48	25.98
	5	25.23	8.86	22.72	47.47	17.02	42.68	31.38	11.10	28.24
	6	23.94	7.49	21.56	39.29	12.76	35.82	28.35	8.94	25.49
	Avg Heads	25.33	9.42	22.84	**50.92**	**19.02**	**45.85**	**32.40**	**12.04**	**29.18**

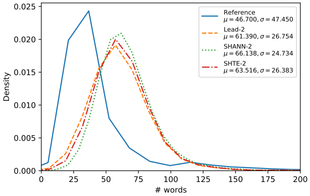
Analysis



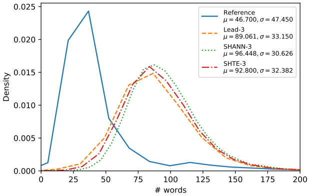
Density
words
Reference
$\mu = 57.544, \sigma = 25.033$
Lead-3
$\mu = 93.200, \sigma = 23.670$
SHANN-3
$\mu = 94.375, \sigma = 23.505$
SHTE-3
$\mu = 95.411, \sigma = 23.886$



density
#words
human summary
avg:33.5, std:10.2
LEAD
avg:76.3, std:18.7
Refresh
avg:76.0, std:18.7
Latent
avg:79.7, std:18.7
ExCoNSumm Extractive
avg:56.2, std:18.7
ExCoNSumm Compressive
avg:54.3, std:18.7
Pointer+Coverage
avg:59.8, std:18.7



Density
words
Reference
$\mu = 46.700, \sigma = 47.450$
Lead-2
$\mu = 61.390, \sigma = 26.754$
SHANN-2
$\mu = 66.138, \sigma = 24.734$
SHTE-2
$\mu = 63.516, \sigma = 26.383$



Density
words
Reference
$\mu = 46.700, \sigma = 47.450$
Lead-3
$\mu = 89.061, \sigma = 33.150$
SHANN-3
$\mu = 96.448, \sigma = 30.626$
SHTE-3
$\mu = 92.800, \sigma = 32.382$

References I

The illustrated Transformer, howpublished = http://jalammar.github.io/illustrated-transformer/, note = Accessed: 2020-01-25.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio.
Neural machine translation by jointly learning to align and translate.
In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Reynier Ortega Bueno, Francisco M. Rangel Pardo, Delia Irazú Hernández Farías, Paolo Rosso, Manuel Montes-y-Gómez, and José Medina-Pagola.
Overview of the task on irony detection in spanish variants.
In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, pages 229–256, 2019.

Jianpeng Cheng and Mirella Lapata.
Neural summarization by extracting sentences and words.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
BERT: pre-training of deep bidirectional transformers for language understanding.
CoRR, abs/1810.04805, 2018.

Manuel Carlos Díaz-Galiano, Manuel García Vega, Edgar Casasola, Luis Chiruzzo, Miguel Ángel García Cumbreras, Eugenio Martínez Cámara, Daniela Moctezuma, Arturo Montejo-Ráez, Marco Antonio Sobrevilla Cabezudo, Eric Sadit Tellez, Mario Graff, and Sabino Miranda-Jiménez.
Overview of TASS 2019: One more further for the global spanish sentiment analysis corpus.
In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, pages 550–560, 2019.

References II

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung.
BanditSum: Extractive summarization as a contextual bandit.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

José-Ángel González, Lluís-Felip Hurtado, and Ferran Pla.
Twilbert: Pre-trained deep bidirectional transformers for spanish twitter.
To be finished.

José-Ángel González, Lluís-Felip Hurtado, and Ferran Pla.
Elirf-upv at irosva: Transformer encoders for spanish irony detection.
In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, pages 278–284, 2019.

José-Ángel González, Lluís-Felip Hurtado, and Ferran Pla.
Elirf-upv at TASS 2019: Transformer encoders for twitter sentiment analysis in spanish.
In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, pages 571–578, 2019.

José-Ángel González, Encarna Segarra, Fernando García-Granada, Emilio Sanchis, and Llu'ıs-F Hurtado.
Extractive summarization using siamese hierarchical transformer encoders.
Journal of Intelligent & Fuzzy Systems, 2019.

José-Ángel González, Encarna Segarra, Fernando García-Granada, Emilio Sanchis, and Llu'ıs-F Hurtado.
Siamese hierarchical attention networksfor extractive summarization.
Journal of Intelligent & Fuzzy Systems, 36(5):4599–4607, 2019.

References III

Max Grusky, Mor Naaman, and Yoav Artzi.
Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom.
Teaching machines to read and comprehend.
In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA, 2015. MIT Press.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III.
Deep unordered composition rivals syntactic methods for text classification.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China, July 2015. Association for Computational Linguistics.

Guillaume Lample, Alexandre Sablayrolles, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou.
Large memory layers with product keys.
In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 8546–8557, 2019.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut.
Albert: A lite bert for self-supervised learning of language representations, 2019.

Yang Liu and Mirella Lapata.
Text summarization with pretrained encoders.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3728–3738, Hong Kong, China, November 2019. Association for Computational Linguistics.

References IV

Yijia Liu, Wanxiang Che, Yuxuan Wang, Bo Zheng, Bing Qin, and Ting Liu.
Deep contextualized word embeddings for universal dependency parsing.
ACM Trans. Asian Low-Resour. Lang. Inf. Process., 19(1), July 2019.

Afonso Mendes, Shashi Narayan, Sebastião Miranda, Zita Marinho, André F. T. Martins, and Shay B. Cohen.
Jointly extracting and compressing documents with summary state representations.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3955–3966, 2019.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean.
Distributed representations of words and phrases and their compositionality.
In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119, 2013.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou.
Summarunner: A recurrent neural network based sequence model for extractive summarization of documents.
In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3075–3081, 2017.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata.
Ranking sentences for extractive summarization with reinforcement learning.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

References V

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever.
Language models are unsupervised multitask learners.
2019.

Abigail See, Peter J. Liu, and Christopher D. Manning.
Get to the point: Summarization with pointer-generator networks.
In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics, 2017.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani.
Self-attention with relative position representations.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Dario Stojanovski and Alexander Fraser.
Coreference and coherence in neural machine translation: A study using oracle experiments.
In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 49–60, Brussels, Belgium, October 2018. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Kaiser, and Illia Polosukhin.
Attention is all you need.
In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, USA, 2017. Curran Associates Inc.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov.
Context-aware neural machine translation learns anaphora resolution.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia, July 2018. Association for Computational Linguistics.

References VI

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao.
Attention-based LSTM for aspect-level sentiment classification.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas, November 2016. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy.
Hierarchical attention networks for document classification.
In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics.

Kaichun Yao, Libo Zhang, Tiejian Luo, and Yanjun Wu.
Deep reinforcement learning for extractive document summarization.
Neurocomputing, 284, 02 2018.

Yang You, Jing Li, Jonathan Hseu, Xiaodan Song, James Demmel, and Cho-Jui Hsieh.
Reducing BERT pre-training time from 3 days to 76 minutes.
CoRR, abs/1904.00962, 2019.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena.
Self-attention generative adversarial networks.
In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7354–7363, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou.
Neural latent extractive document summarization.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Attention is all I need

José Ángel González[1]
jogonba2@dsic.upv.es

[1]Valencian Research Institute for Artificial Intelligence (VRAIN)
Universitat Politècnica de València

January 26, 2020