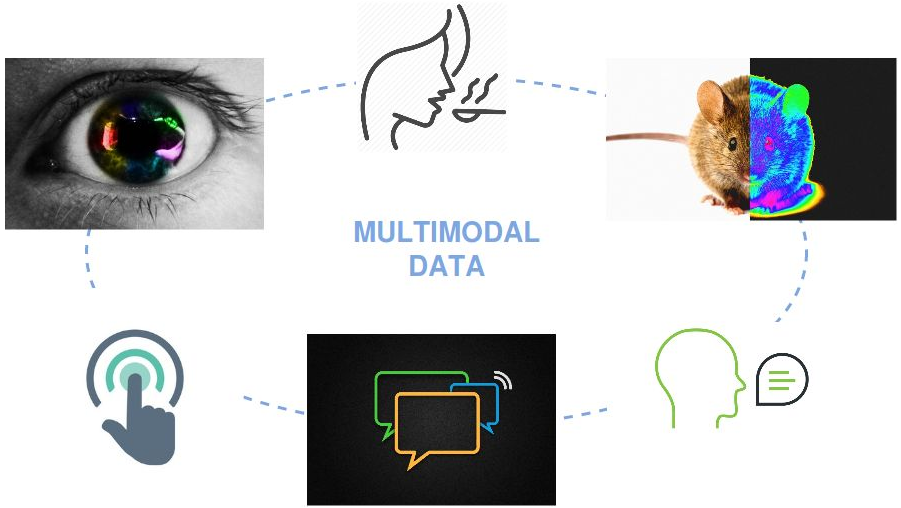
Multimodal machine learning

SIFED Talk
June 06 - 2019

Viviana Beltrán
In collaboration with: Antoine Doucet, Nicholas Journet, Mickael Coustaty, Juan Caicedo

What is Multimodal?



Sample of Modalities

- Natural language
- Visual
- Auditory
- Haptics / touch
- Smell, taste and self-motion
- Physiological signals
- Other modalities

Challenges in Multimodal ML

- **Representation**
  - Joint representations (Multimodal space)
  - Coordinated representations (Correlated projections)
- **Alignment:** Correspondances between elements in the different modalities (media description)
- **Fusion:** Early or late Fusion?
- **Translation:** Event recognition
- **Co-Learning:** Transfer learning

Samples of Real-World Tasks Tackled by Multimodal Research

Affect recognition

- Emotion

Event recognition

- Action recognition

Multimodal information retrieval

- Content based/Cross-media

Media description

- Visual Question Answering

State of the art

MML in Cross-modal Retrieval

General classification for methods addressing the task of cross-modal retrieval

- Models that find correlated subspaces: CCA based models
- Methods based on tensor factorizations and graph embeddings: Manifold learning
- Deep learning methods

Drawbacks of the state of the art methods

- Double training phases for each separate model
- Mainly, focused on the visual modality --> Textual modality modeled with methods such as BoW, topics, etc.
- Over trained to perform in single tasks

Research questions

Objective: Design and evaluate new computer systems that take advantage of different modalities through strategies in representation learning.

Research questions:

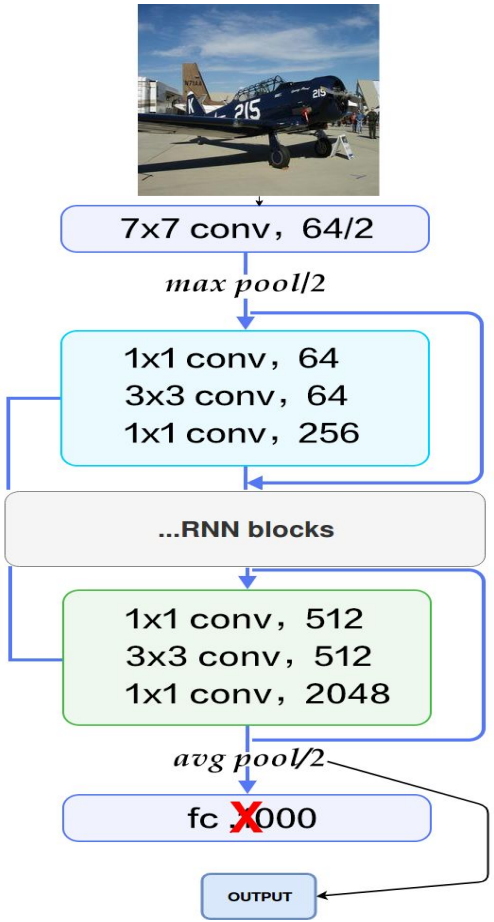
- How to leverage all the available information in databases?
- How to combine different modalities in an end-to-end system?
- How to evaluate the quality of the features obtained?

Methodology

- Design/adapt and study models that can process different modalities
  - By studying and analysing strategies, mainly in representation, alignment and fusion challenges
- Evaluate those models in the context of two selected tasks:
  - Multimodal information retrieval (Application 1)
  - Media description (Application 2)

Visual Network

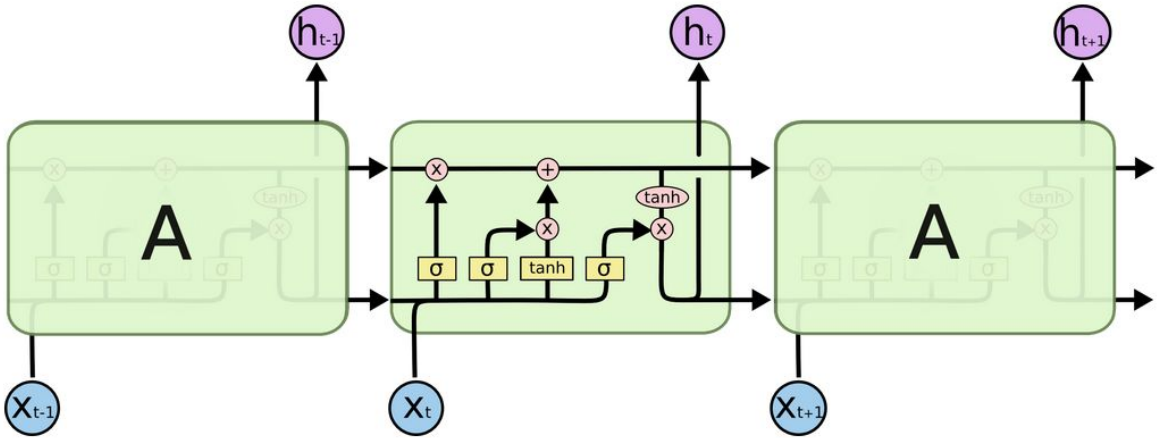
- Deeper networks
- Skip-connections
- Faster training
- Successfully applied in visual tasks [2]



Textual network

RNN

- RNNs have memory
- RNNs handle larger dependencies
- They allow us to operate over sequences of vectors
- Successfully applied in many applications involving sequences of data



The repeating module in an LSTM contains four interacting layers.

LSTM module [7]

Proposed model

By combining the previous strategies in an end-to-end system:

- Visual Network: ResNET (freezed) + 1 FC (Fully connected)
- Textual Network: GloVe initialization + 2 LSTM + 1 FC
- Multimodal Network: Fusion layer + 2 FC + 1 Classifier layer

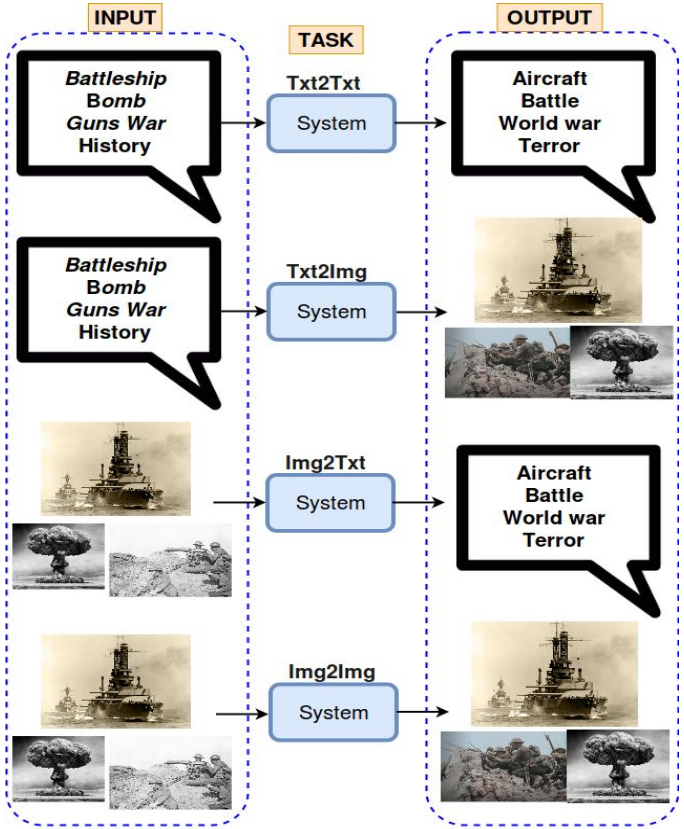
Proposed model

Advantages:

- ➢ First time to combine a ResNET + a sequential model in the context of multimodal retrieval (1st Application)
- ➢ Allow us to evaluate the embeddings from multiple modalities and at different points (different layers)
- ➢ Allow us to evaluate different retrieval tasks using the same model under the same computational framework
- ➢ Flexibility to adapt the system to address different tasks.

Application 1

Multimodal information retrieval: Cross and Uni modal



Wikipedia retrieval database

- Modalities: Images and text
- Text Length ~ 100 tokens
- 2,173 Train samples
- 693 Test samples
- 10 semantic categories

History

'arrived', 'ascended', 'atlantic', 'avoiding', 'base', 'bay', 'believe', 'belonged', 'beyond', 'camp'

Music

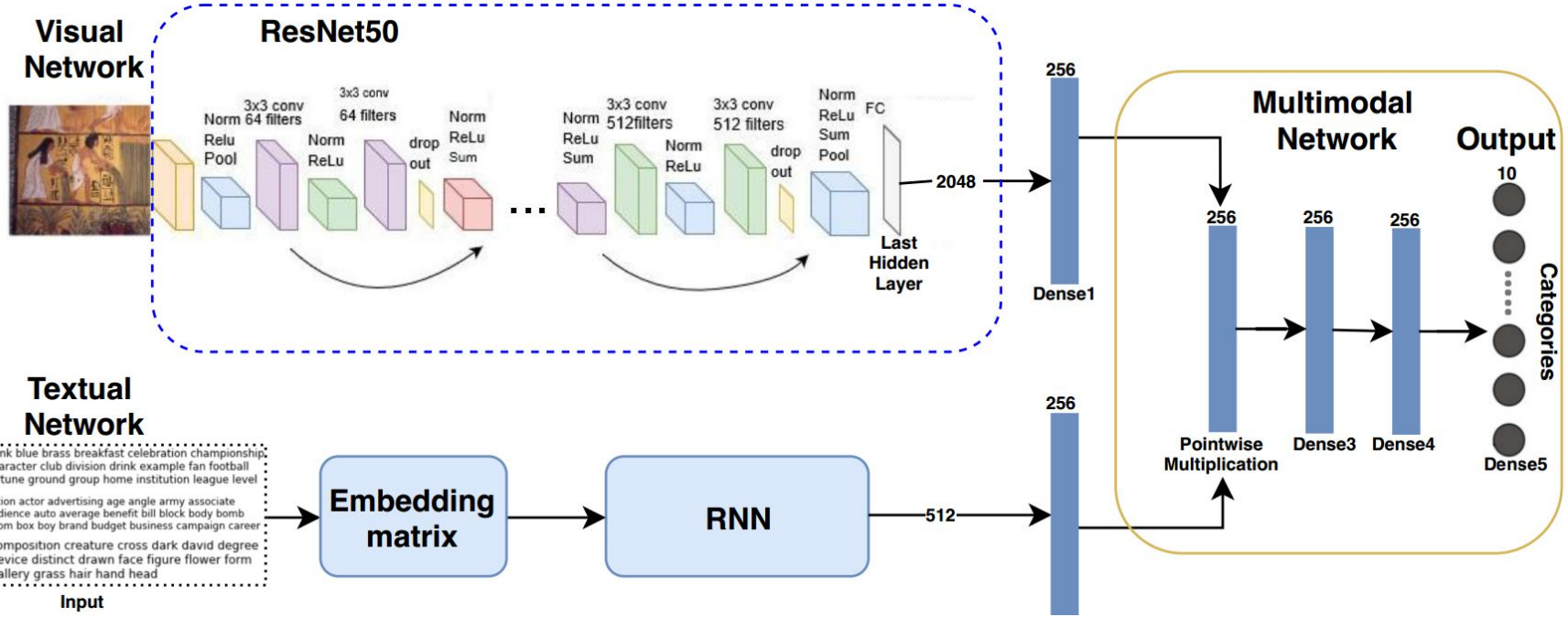
'abuse', 'accompanied', 'achieve', 'acoustic', 'adding', 'album', 'alice', 'artwork', 'bad', 'baker'

Literature & theatre

'abundance', 'ancient', 'appears', 'archaeological', 'archaeologist', 'box', 'burial', 'business', 'chamber', 'chance'

Triplet of samples: Category, Image, Tokens

Proposed model



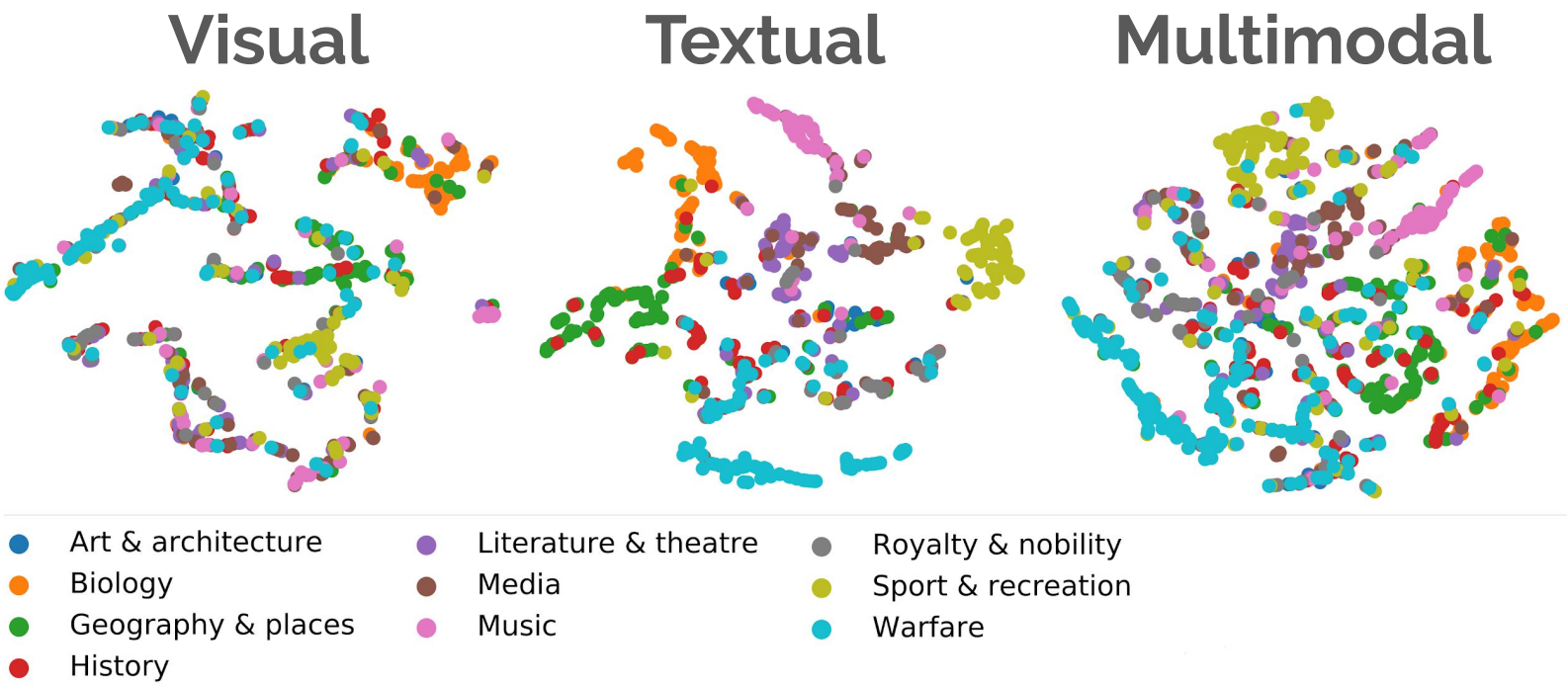
MAP Comparison from features computed from different layers

Task	Layer			
	Dense1(V) / Dense2(T)	Dense3	Dense4	Dense5
Txt2Txt	0.5615	0.5662	**0.5672**	0.5671
Img2Img	0.2555	0.2707	0.2733	**0.2831**
Img2Txt	0.3792	0.3681	0.3755	**0.4221**
Txt2Img	0.3821	**0.3863**	0.3773	0.3600

Multimodal retrieval results

Method	Task	8	50	500	100%
OURS	**Txt2Txt**	**0.681**	**0.641**	**0.576**	**0.567**
OURS	**Img2Img**	**0.455**	**0.399**	0.299	**0.283**
CMSTH [8]		-	-	**0.449**	-
OURS	**Img2Txt**	**0.451**	**0.435**	**0.426**	0.422
MDCR [9]		-	-	-	**0.435**
OURS	**Txt2Img**	**0.489**	**0.490**	0.382	0.360
CMSTH [8]		-	-	**0.387**	-
TextTopicNet [10]		-	-	-	**0.402**

Visualization of embeddings



Results

- The evaluation of embeddings from different layers demonstrates the advantages and gain in performance of using different modalities to train the system
- Competitive performance over all the state-of-the-art methods
- Although the dataset is very complex, the model achieves a global state-of-the-art performance
- The t-sne explains those categories that imposes a challenge to the model
- Further research and improvements in the visual network are required
- The results provide us with information about the strengths and weaknesses of our proposal, thus, we can work on them.

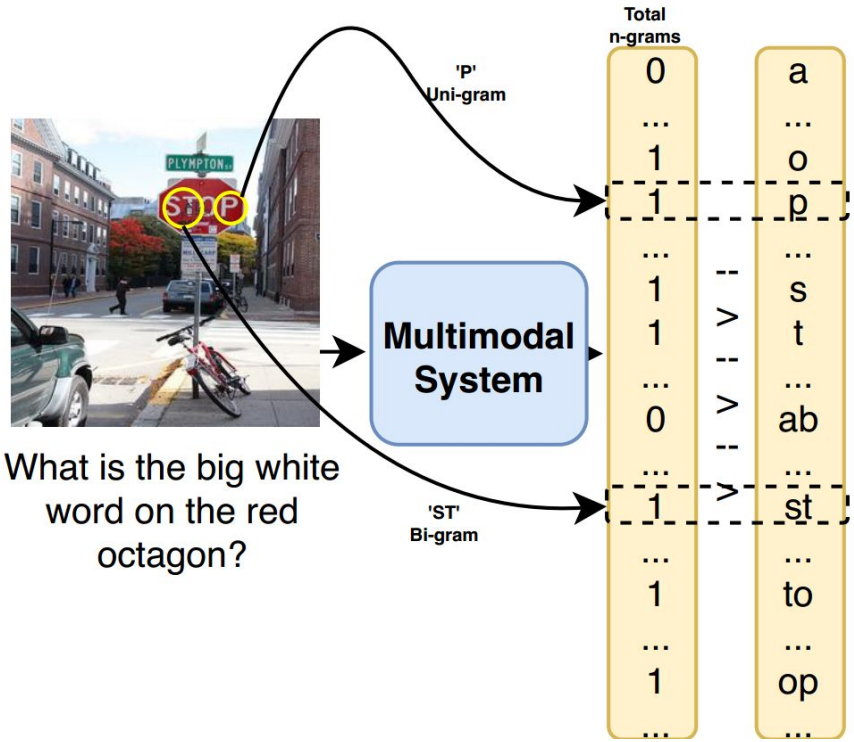
Cross-modal retrieval sample

Query Text ability actual admiral aft aircraft airfield arrived assigned assisted attacking base battleship bay beach begun bomb bombardment buried canal caroline carrier coast continuous

Top@6 retrieved images

Application 2 (ongoing work)

Scene Text Recognition via Visual Question Answering



Explorations in Textual network

Embedding matrix

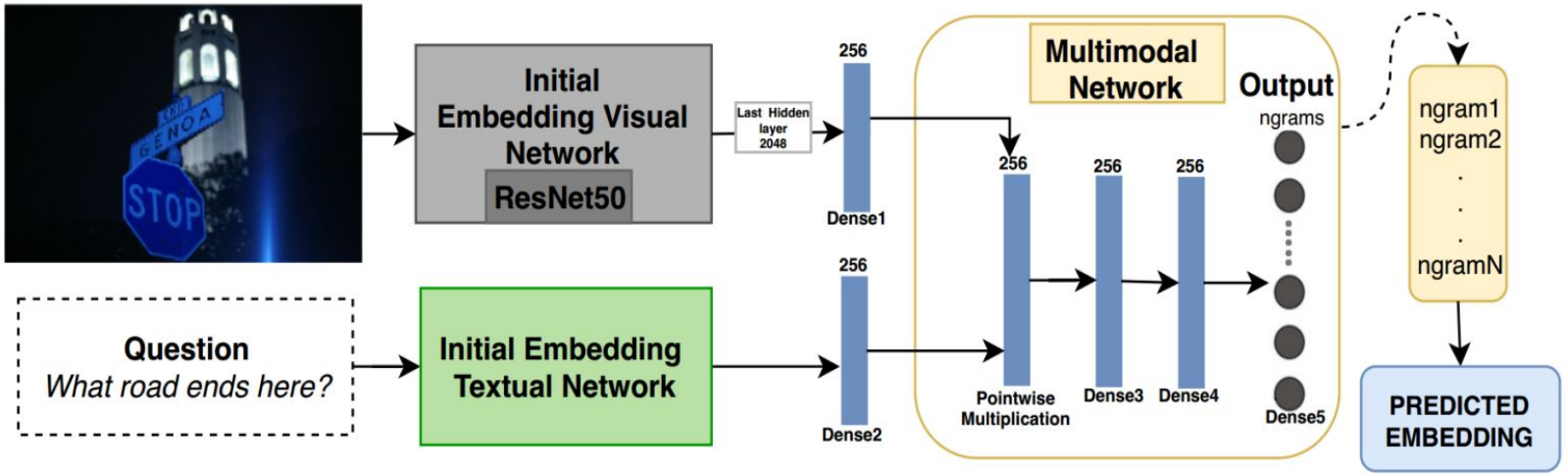
- Context-free
  - GloVe vectors [3]
- Context-based
  - BERT [6]

Sentence: "*He went to the prison **cell** with his **cell** phone to extract blood **cell** samples from inmates*"

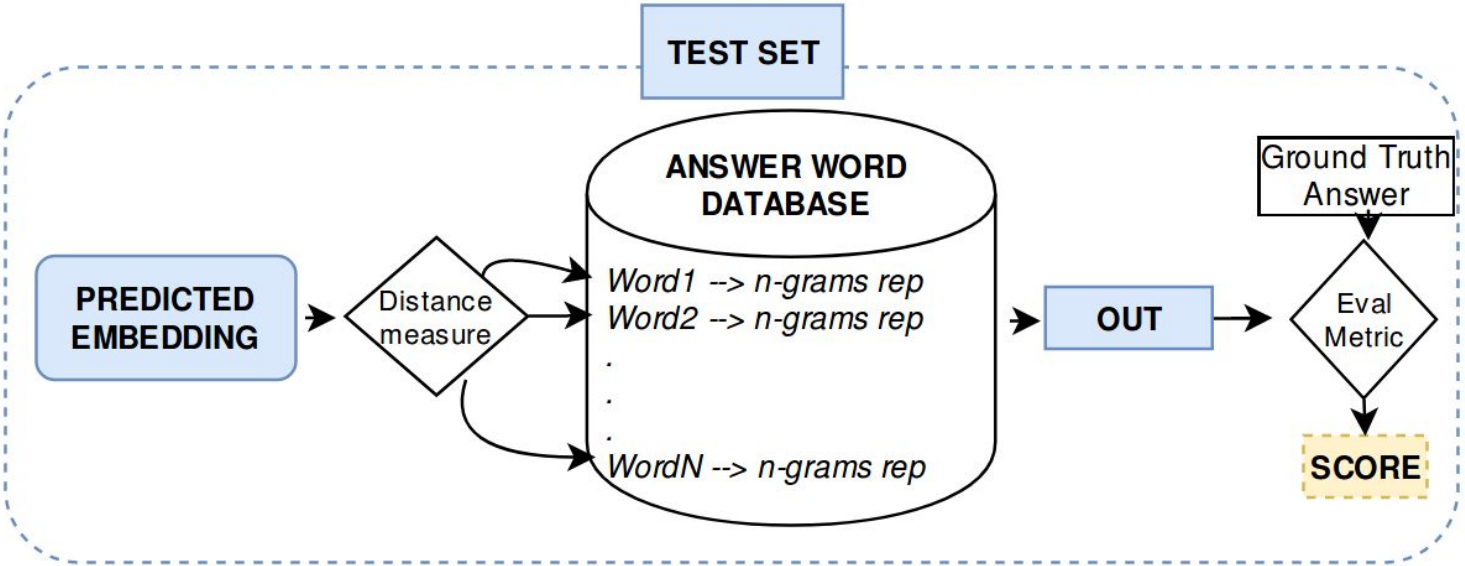
Embeddings for the word *'cell'*:

- Glove --> Only 1 embedding
- BERT --> 3 Different embeddings

Proposed model



Evaluation protocol



ICDAR Challenge:

Robust Reading Challenge on Scene Text Visual Question Answering

Textual Embedding	Score	Trimmed Score
GloVe	0.23093	0.12058
Bert_per_word	0.21814	0.11796
Bert_per_sentence	0.21622	0.11430

Conclusions and future work

- Many tasks in multimodal learning represent a challenge because the nature of the data involved, which motivates an increasing and active interest in the research community.
- We propose an end-to-end methodology that is flexible and that can be applied to address and study different tasks in the context of multimodal learning.
- Our ongoing work, as well as the few state of the art works, in the novel application of Scene Text Recognition via VQA requires much more research, due to the ambitious requirements it imposes.
- Our future research aims to strengthen the methodologies proposed by improving each component, as well as the involvement of more modalities that can contribute in finding better multimodal spaces.

References

[1] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423-443.

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[3] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

[4] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, *28*(10), 2222-2232.

[5] Sudholt, S., & Fink, G. A. (2016, October). PHOCNet: A deep convolutional neural network for word spotting in handwritten documents. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 277-282). IEEE.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[7] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[8] Xie, L., Zhu, L., Pan, P., & Lu, Y. (2016). Cross-modal self-taught hashing for large-scale image retrieval. *Signal Processing*, *124*, 81-92.

References

[9] Wei, Y., Zhao, Y., Zhu, Z., Wei, S., Xiao, Y., Feng, J., & Yan, S. (2016). Modality-dependent cross-media retrieval. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *7*(4), 57.

[10] Patel, Y., Gomez, L., Gomez, R., Rusiñol, M., Karatzas, D., & Jawahar, C. V. (2018). TextTopicNet-Self-Supervised Learning of Visual Features Through Embedding Images on Semantic Text Spaces. *arXiv preprint arXiv:1807.02110*.