# VisualMRC Paper and State of the Art

Sebastian López ·, Santiago Pineda, · Rafael Mejia , · Andrés Álvarez

Digital Signal Processing and Control Group - (GCPDS)
Universidad Nacional de Colombia
Manizales, Colombia
July 2023

# Contenido

1 VisuaL Question Answering

2 Machine Reading Comprehension

# Visual Question Answering

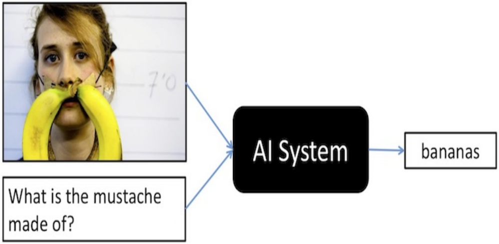
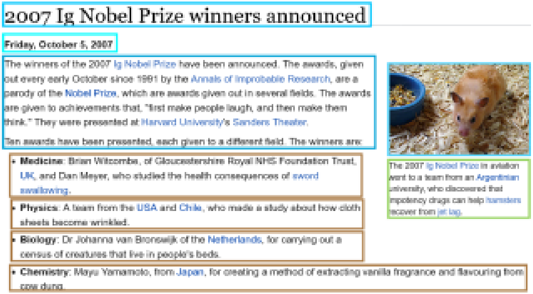


Figure: VQA Task

# Machine Reading Comprehension



## 2007 Ig Nobel Prize winners announced

**Friday, October 5, 2007**

The winners of the 2007 Ig Nobel Prize have been announced. The awards, given out every early October since 1991 by the Annals of Improbable Research, are a parody of the Nobel Prize, which are awards given out in several fields. The awards are given to achievements that, "first make people laugh, and then make them think." They were presented at Harvard University's Sanders Theater.

Ten awards have been presented, each given to a different field. The winners are:

- **Medicine**: Brian Witcombe, of Gloucestershire Royal NHS Foundation Trust, UK, and Dan Meyer, who studied the health consequences of sword swallowing.
- **Physics**: A team from the USA and Chile, who made a study about how cloth sheets become wrinkled.
- **Biology**: Dr Johanna van Bronswijk of the Netherlands, for carrying out a census of creatures that live in people's beds.
- **Chemistry**: Mayu Yamamoto, from Japan, for creating a method of extracting vanilla fragrance and flavouring from cow dung.

The 2007 Ig Nobel Prize in aviation went to a team from an Argentinian university, who discovered that impotency drugs can help hamsters recover from jet lag.

**Q**: Who were the winners of the Ig Nobel prize for Biology and Chemistry?
**A**: The winner of the Ig Nobel prize for biology was Dr Johanna van Bronswijk, and the winner for Chemistry was Mayu Yamamoto.

Figure: MRC Task

# Pipeline MRC Task

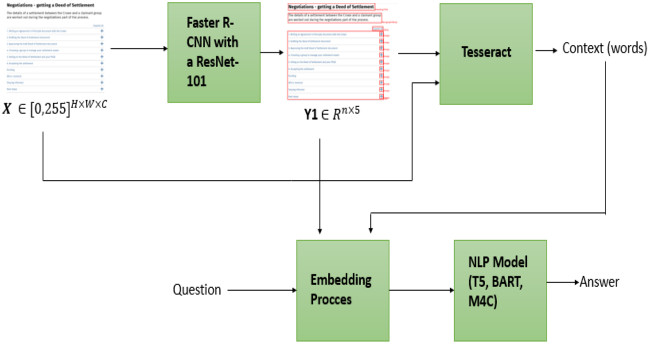

$X \in [0,255]^{H\times W\times C}$
Faster R-CNN with a ResNet-101
$Y1 \in R^{n\times 5}$
Tesseract
Context (words)
Question
Embedding Procces
NLP Model (T5, BART, M4C)
Answer

Figure: Pipeline MRC Task

¡May or may not include visual information!

# Input Sequence

$$x^{\text{token}} = \left\{ \begin{array}{l} [\mathrm{S}], q_1, \ldots, q_m, [\mathrm{SEP}], [\mathrm{L}_{r_1}], w_{1,1}, \ldots, w_{1,M}, \\ [\mathrm{L}_{r_2}], \ldots, [\mathrm{L}_{r_N}], w_{r_N,1}, \ldots, w_{r_N,M} \end{array} \right\}$$

Figure: Input Sequence Structure

Ex: [S] Who can send a congratulatory message for a 50th wedding anniversary? [sep] [Heading/Title] Get a congratulatory message [Image] [Paragraph/Body] In this guide [Subtitle/Byline] 2.

# Input Embedding

$$z_k = \mathrm{LN}(z_k^{\mathrm{token}} + z_k^{\mathrm{pos}} + z_k^{\mathrm{seg}} + z_k^{\mathrm{loc}} + z_k^{\mathrm{app}})$$

Figure: Input Embedding Structure

where:

- $\mathbf{Z_k} \in \mathbb{R}^H$ : Input Embedding.
- $\mathbf{Z_k^{token}} \in \mathbb{R}^H$ : Input sequence token.
- $\mathbf{Z_k^{pos}} \in \mathbb{R}^H$ : Input sequence position.
- $\mathbf{Z_k^{seg}} \in \mathbb{R}^H$ : Segment Embedding.
- $\mathbf{Z_k^{loc}} \in \mathbb{R}^H$ : Location Embedding.

# Input Embedding

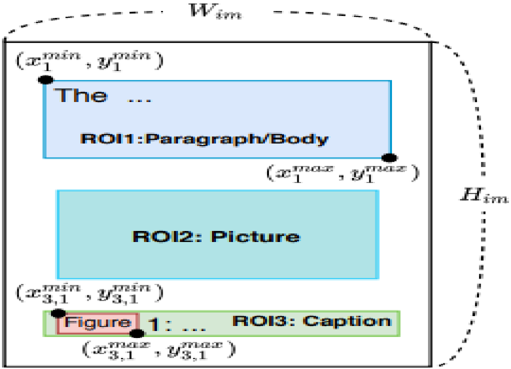
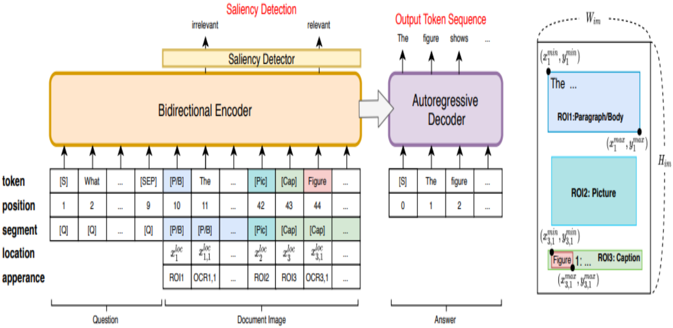


Figure: Location Embedding

- $\mathbf{Z}_{\mathbf{k}}^{\mathbf{app}} \in \mathbb{R}^H$ : Appearence Embedding.

# Input Embedding



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| token | [S] | What | ... | [SEP] | [P/B] | The | ... | [Pic] | [Cap] | Figure | ... |
| position | 1 | 2 | ... | 9 | 10 | 11 | ... | 42 | 43 | 44 | ... |
| segment | [Q] | [Q] | ... | [Q] | [P/B] | [P/B] | ... | [Pic] | [Cap] | [Cap] | ... |
| location | | | | | $x_1^{loc}$ | $x_{1,1}^{loc}$ | ... | $x_2^{loc}$ | $x_3^{loc}$ | $x_{3,1}^{loc}$ | ... |
| apperance | | | | | ROI1 | OCR1,1 | ... | ROI2 | ROI3 | OCR3,1 | ... |

| | | | |
|---|---|---|---|
| [S] | The | figure | ... |
| 0 | 1 | 2 | ... |

Figure: Main Module

# Saliency Detection and Saliency Loss

- Saliency Detection:

$$P_{i,j} = \mathrm{sigmoid}({w^s}^\top h_{w_{i,j}} + b^s)$$

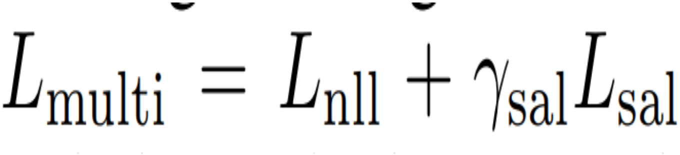
Figure: Saliency Detection

Saliency Loss:

$$L_{\mathrm{sal}} = -\frac{1}{NM}\sum_{i}^{N}\sum_{j}^{M}\left(\begin{array}{l} s_{i,j}\log P_{i,j} + \\ (1 - s_{i,j})\log(1 - P_{i,j}) \end{array}\right)$$

Figure: Saliency Loss

# Multitask Learning

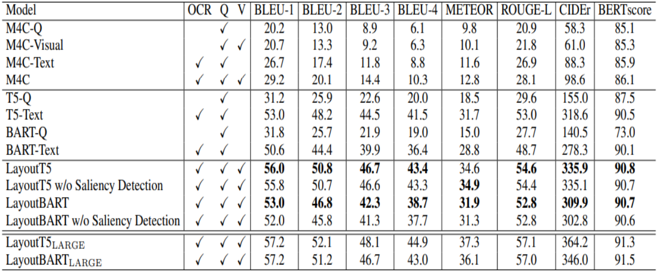


$$L_{\mathrm{multi}} = L_{\mathrm{nll}} + \gamma_{\mathrm{sal}} L_{\mathrm{sal}}$$

Figure: Multitask Learning

$$L_{nll} = -\frac{1}{T}\sum_{t} Log(P(Y_t)); t = 1, 2, ..., T \qquad (1)$$

# Experiments



| Model | OCR | Q | V | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr | BERTscore |
|---|---|---|---|---|---|---|---|---|---|---|---|
| M4C-Q | | ✓ | | 20.2 | 13.0 | 8.9 | 6.1 | 9.8 | 20.9 | 58.3 | 85.1 |
| M4C-Visual | | ✓ | ✓ | 20.7 | 13.3 | 9.2 | 6.3 | 10.1 | 21.8 | 61.0 | 85.3 |
| M4C-Text | ✓ | ✓ | | 26.7 | 17.4 | 11.8 | 8.8 | 11.6 | 26.9 | 88.3 | 85.9 |
| M4C | ✓ | ✓ | ✓ | 29.2 | 20.1 | 14.4 | 10.3 | 12.8 | 28.1 | 98.6 | 86.1 |
| T5-Q | | ✓ | | 31.2 | 25.9 | 22.6 | 20.0 | 18.5 | 29.6 | 155.0 | 87.5 |
| T5-Text | ✓ | ✓ | | 53.0 | 48.2 | 44.5 | 41.5 | 31.7 | 53.0 | 318.6 | 90.5 |
| BART-Q | | ✓ | | 31.8 | 25.7 | 21.9 | 19.0 | 15.0 | 27.7 | 140.5 | 73.0 |
| BART-Text | ✓ | ✓ | | 50.6 | 44.4 | 39.9 | 36.4 | 28.8 | 48.7 | 278.3 | 90.1 |
| LayoutT5 | ✓ | ✓ | ✓ | **56.0** | **50.8** | **46.7** | **43.4** | 34.6 | **54.6** | **335.9** | **90.8** |
| LayoutT5 w/o Saliency Detection | ✓ | ✓ | ✓ | 55.8 | 50.7 | 46.6 | 43.3 | **34.9** | 54.4 | 335.1 | 90.7 |
| LayoutBART | ✓ | ✓ | ✓ | **53.0** | **46.8** | **42.3** | **38.7** | **31.9** | **52.8** | **309.9** | **90.7** |
| LayoutBART w/o Saliency Detection | ✓ | ✓ | ✓ | 52.0 | 45.8 | 41.3 | 37.7 | 31.3 | 52.8 | 302.8 | 90.6 |
| LayoutT5$_{\text{LARGE}}$ | ✓ | ✓ | ✓ | 57.2 | 52.1 | 48.1 | 44.9 | 37.3 | 57.1 | 364.2 | 91.3 |
| LayoutBART$_{\text{LARGE}}$ | ✓ | ✓ | ✓ | 57.2 | 51.2 | 46.7 | 43.0 | 36.1 | 57.0 | 346.0 | 91.5 |

Figure: Experiments

- BART base: 6 layers
- T5 base: 12 layers
- BART large: 12 layers
- T5 large: 24 layers

# Training Hyperparameters

Table: Training Hyperparameters

| Hyperparameter | Value |
|---|---|
| $\lambda_{sal}$ | 1 |
| Batch Size | 32 |
| Epoch | 7 |
| Optimizer | ADAM |
| Learning Rate | 3e-5 |

# Thanks!