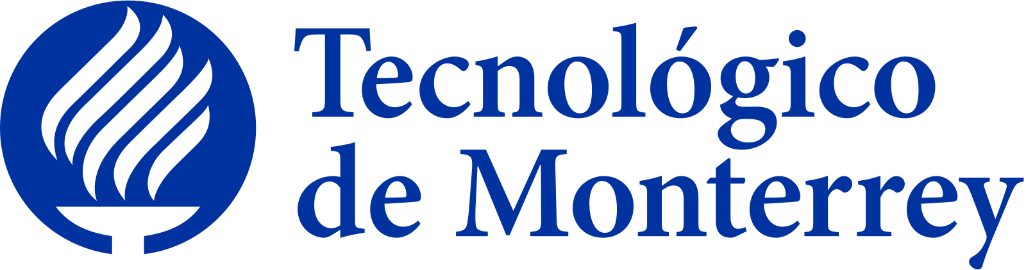
**Maesters in Applied Artificial Inteligence**

**Subject: Advanced Machine Learning Methods**



**Prof. Dr. José Antonio Cantoral Cevallos**

**Final Quiz: Design Specification for an Image Captioning Mode**

Luis Alfredo Negron Naldos A01793865

Javier Muñoz Barrios A01794423

Aurelio Antonio Lozano Rábago A01081266

22 November 2023

# Overview:

As the culmination of our deep learning course, you are tasked with designing a detailed specification for an "Image Captioning" model, i.e. given and image, the model should provide a brief description of it. Note that only the specification is required, not the code (of course you are encourage to code it as a personal project, it will be worth the effort!).

This exercise will consolidate your understanding of CNNs and RNNs, and how these models can be integrated to solve complex problems like image captioning.

# Objective:

* To create a detailed system design for an Image Captioning model.
* To demonstrate an understanding of CNNs and RNNs and their application in a real-world scenario.
* To showcase the ability to plan model architectures and training regimens.

# Task Description:

## Model Architecture:

* Image Processing Component: Specify the architecture of the CNN model to be used for image feature extraction. Detail the type and number of layers, activation functions, and any regularization techniques.
* Text Generation Component: Outline the RNN/LSTM architecture for generating captions. Specify the architecture details including the number of layers, cell type (RNN, LSTM, GRU), and any other relevant parameters.
* Integration: Explain how you will integrate the CNN and RNN components for the task of image captioning.

Text Generation

Image feature extraction



CNN

Linear

LSTM

LSTM

LSTM

LSTM

Softmax

Softmax

Softmax

Softmax

Wemb

Wemb

Wemb

### Image processing: image feature extraction

Training data: dataset that contains image name and their respective captions.

#### CNN model

To be used a model based on VGG16, Input shape required: RGB 224 x 224

1. 64 filters (3x3) padding 1 activation Relu
2. 64 filters (3x3) padding 1 activation Relu
3. Max pooling (2,2) strides 2

(now pixels are 112 x112)

1. 128 filters (3x3) padding 1 activation Relu
2. 128 filters (3x3) padding 1 activation Relu
3. Max pooling (2,2) strides 2

(now pixels are 56 x56)

1. 256 filters (3x3) padding 1 activation Relu
2. 256 filters (3x3) padding 1 activation Relu
3. 256 filters (3x3) padding 1 activation Relu
4. Max pooling (2,2) strides 2

(now pixels are 28 x28)

1. 512 filters (3x3) padding 1 activation Relu
2. 512 filters (3x3) padding 1 activation Relu
3. 512 filters (3x3) padding 1 activation Relu
4. Max pooling (2,2) strides 2

(now pixels are 14 x14)

1. 512 filters (3x3) padding 1 activation Relu
2. 512 filters (3x3) padding 1 activation Relu
3. 512 filters (3x3) padding 1 activation Relu
4. Max pooling (2,2) strides 2

(now pixels are 7 x7 and 512 channels)🡪 Volume of 25088 features

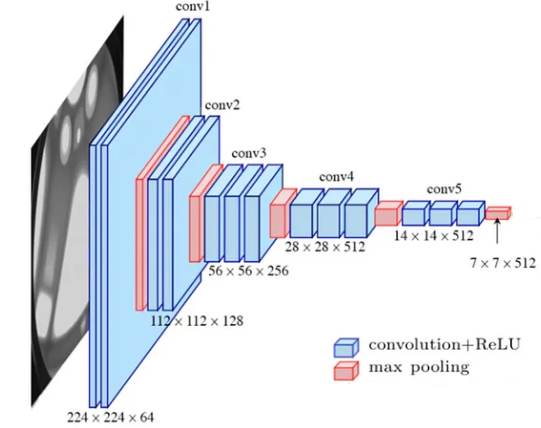


Figure 1: Architecture of VGG16 without fully conected layers. Source: Researchgate.net

#### Preprocess the spatial dimension

Flatten the output shape of the VGG16 to have a single vector of 25088

### Text generation

Now the previous input will be feed into a LSTM

#### Tokenize the words

Convertion of the words into unique vectors

#### LSTM layer

2 LSTM layers

#### Dense layer

Connect the LSTM to a fully connected, then softmax to find the highest probability of the words.

## Training Specification

### Data preprocessing

Use public dataset:

* Flicker8k\_Dataset contains 8091 images
* Flickr\_8k\_text: dataset folder containing the text files and captions

Here how to preprocess the image and text data

### Training regimen

xxepochs suggested using a batch size of xxx

### Optimizer and learning rate

Adams optimizer and learning rate of XXXX

## Performance Metrics:

* Specify which metrics you would use to evaluate the performance of your image captioning model.

## Additional Considerations:

* Discuss any potential challenges and how you might address them.