Multilabels Classification of Podcast Genres

An Experiment with Multimodal Neural Networks

Podcasts

Podcasts are an emerging medium on the internet and the number of shows is growing exponentially over the years

They provide huge amounts of data for many different tasks (e.g. audio transcription)

Streaming services (e.g. Spotify, Apple Music) usually classify shows on the basis of manual annotations and use those labels to suggest new contents to their users

Is it possible to automatically classify podcasts according to the metadata available?

Project workflow

1. Extract and prepare the data
2. Find the right architecture for each type of data (text and images)
3. Build and train a multimodal network
4. Evaluate the results

Dataset

Lack of existing dataset focusing on this kind of task (e.g. Spotify dataset is meant for audio transcription)

Extracted a dataset from Apple Music using their APIs, including nearly 30000 different shows

Each show has metadata regarding title, description, cover image, authors and primary genre

Dataset: cover images

13 Minutes to the Moon (Technology)

The Interior Design Business (Design)

Two Girls Talk Balls (Sports)

Five Point Move - U.S. Greco-Roman Wrestling (Sports)

A Delectable Education Charlotte Mason Podcast (Education)

The Christian Woman Business Podcast (Business)

The Wellness Mama Podcast (Health)

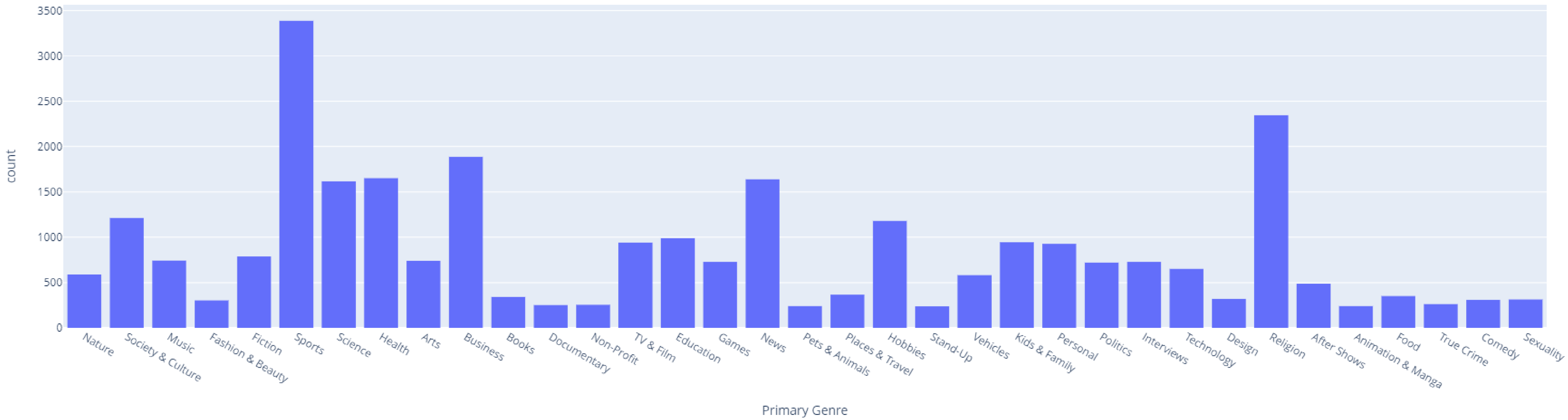
Metabolic Academy (Health)

Quran recitations (Religion)

Dataset: genres

Originally 110 different genres with many overlaps

After pre-processing only 35 different genres with a rather unbalanced distribution

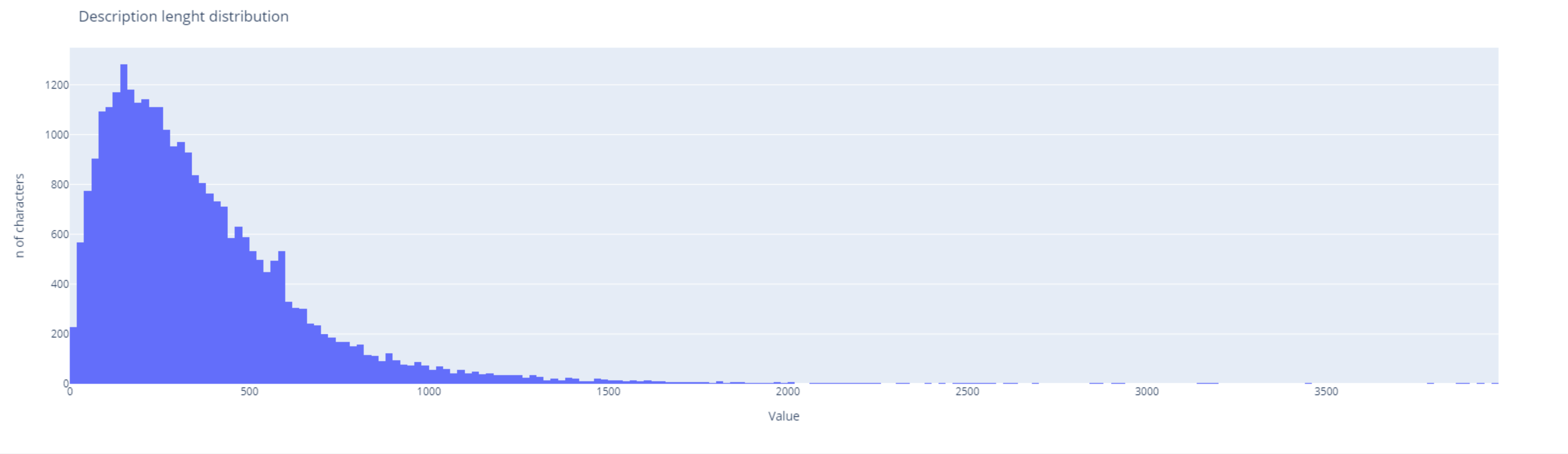


Dataset: descriptions

Two Girls Talk Balls (Sports)

Title: Two Girls Talk Balls

Description: «An alternative take on Women's Football. Covering the WSL, NWSL, Euro 2022 and Women's Champions League; Charlotte and Tamsin bring a refreshing voice with debate, banter and highlights from the best of the Women's Game. Proudly sponsored by FotMob”



DNN: images

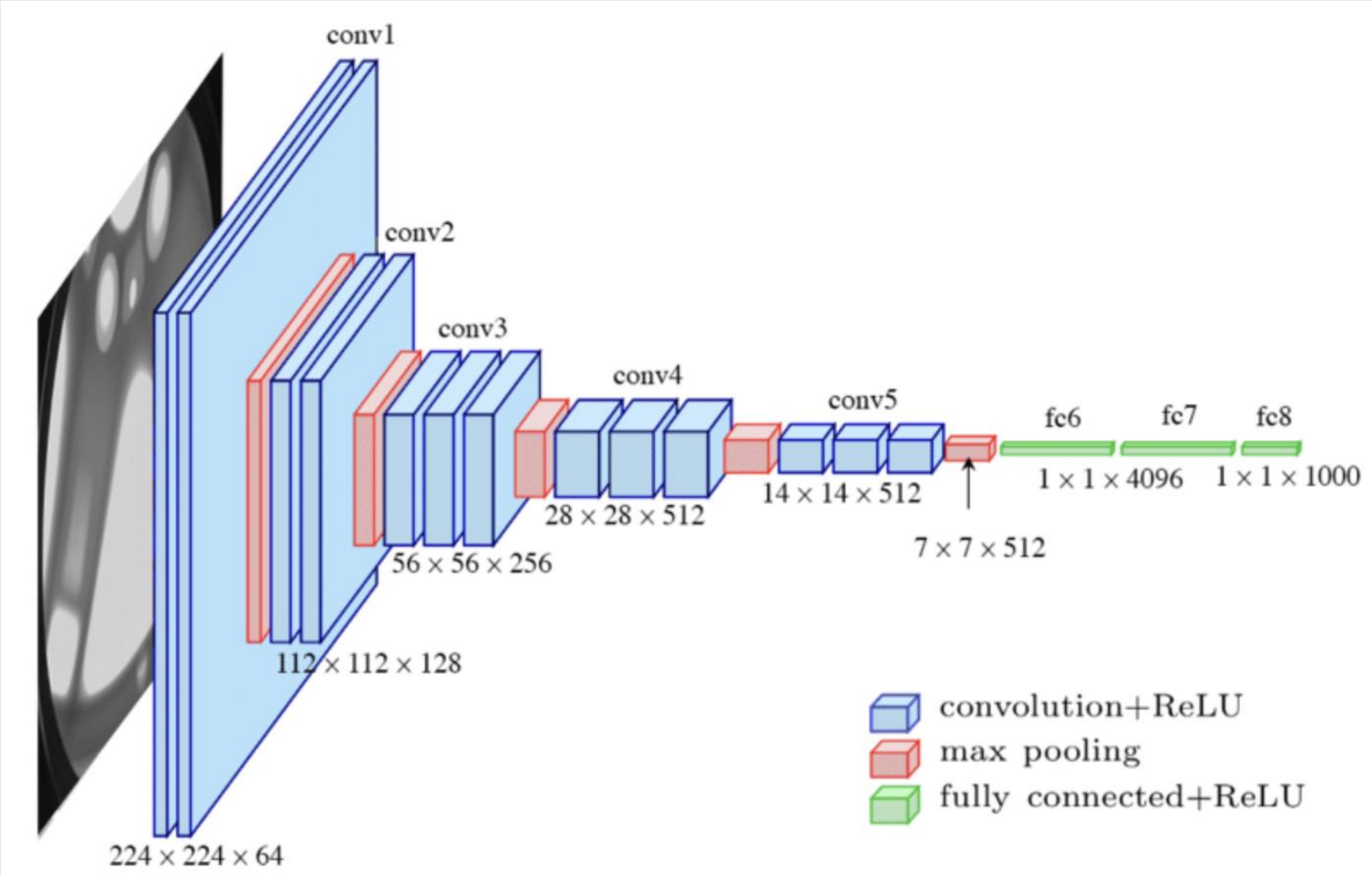
Images are coded as a numerical matrix (height x width x channels)

Each image in the dataset is in JPEG format and has a shape of 600 x 600 x 3 (RGB)

DNN can learn how to filter (parts of) an image to extract relevant features (Convolutional Neural Network)

Instead of building from scratch a CNN, better performances are achieved use a pre-trained architecture (transfer learning)

VGG16

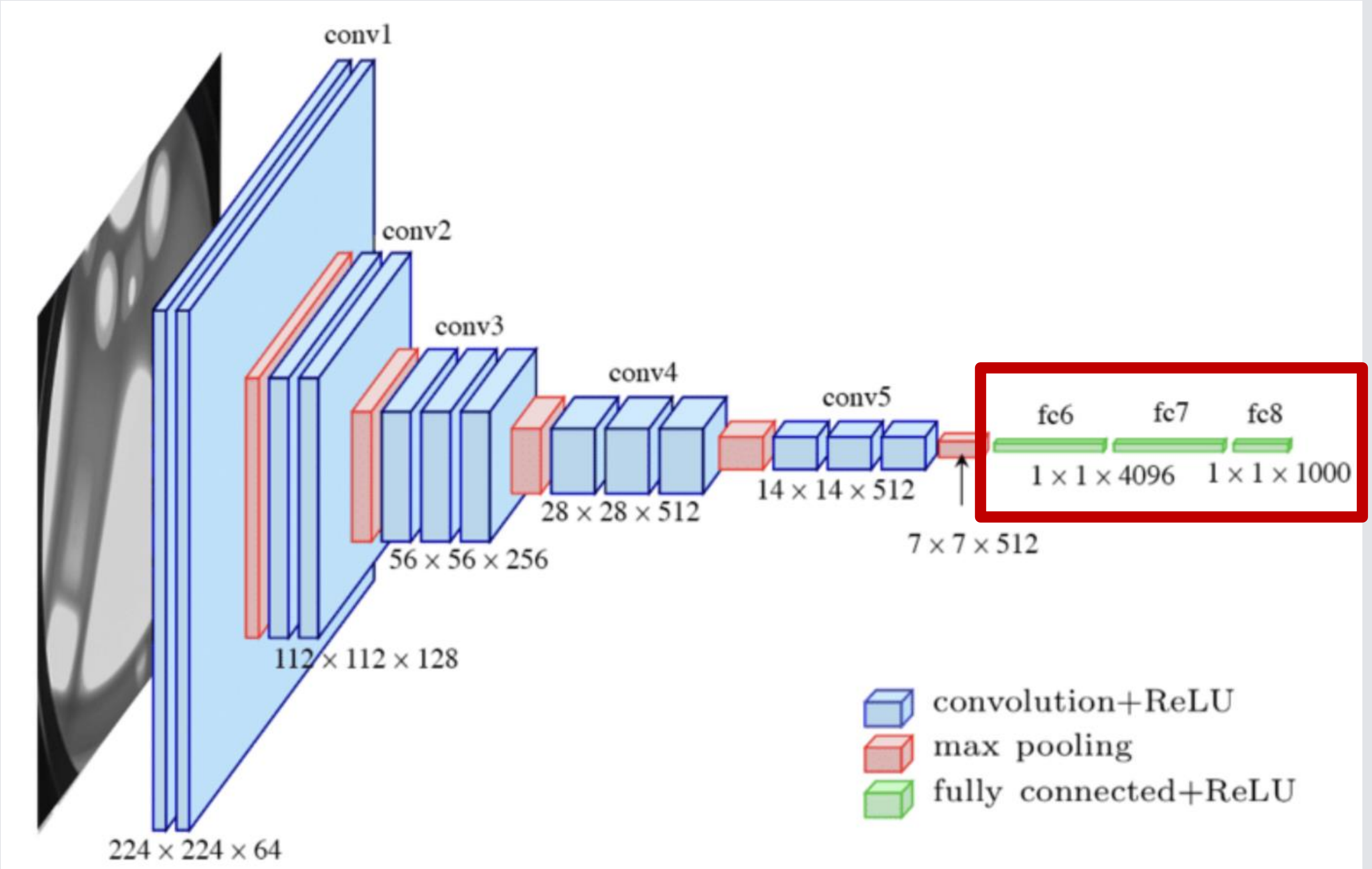


Large architecture trained on millions of images (ImageNet dataset)

Stack of convolutional layers and max pooling using ReLU as activation function

It can be used for different tasks without re-training thanks to transfer learning

VGG16



Cut the last layer (classifier head) and replace it with a dense layer with a neuron for each label (35)

Reshape images to 224 x 224 size and convert them from RGB to BGR (automatically pre-processed by Keras)

Use data augmentation to create artificially additional samples for training (random flip and random rotation)

DNN: descriptions

Text is understood as a sequence problem, thus RNN are commonly used for text classification and language modelling

In order to be fed to RNNs, text must be normalized: tokenization, stop-words elimination, lemmatisation

Textual data has to be converted into numerical arrays (assigning each token to a numerical ID and passing it to an embedding layer)

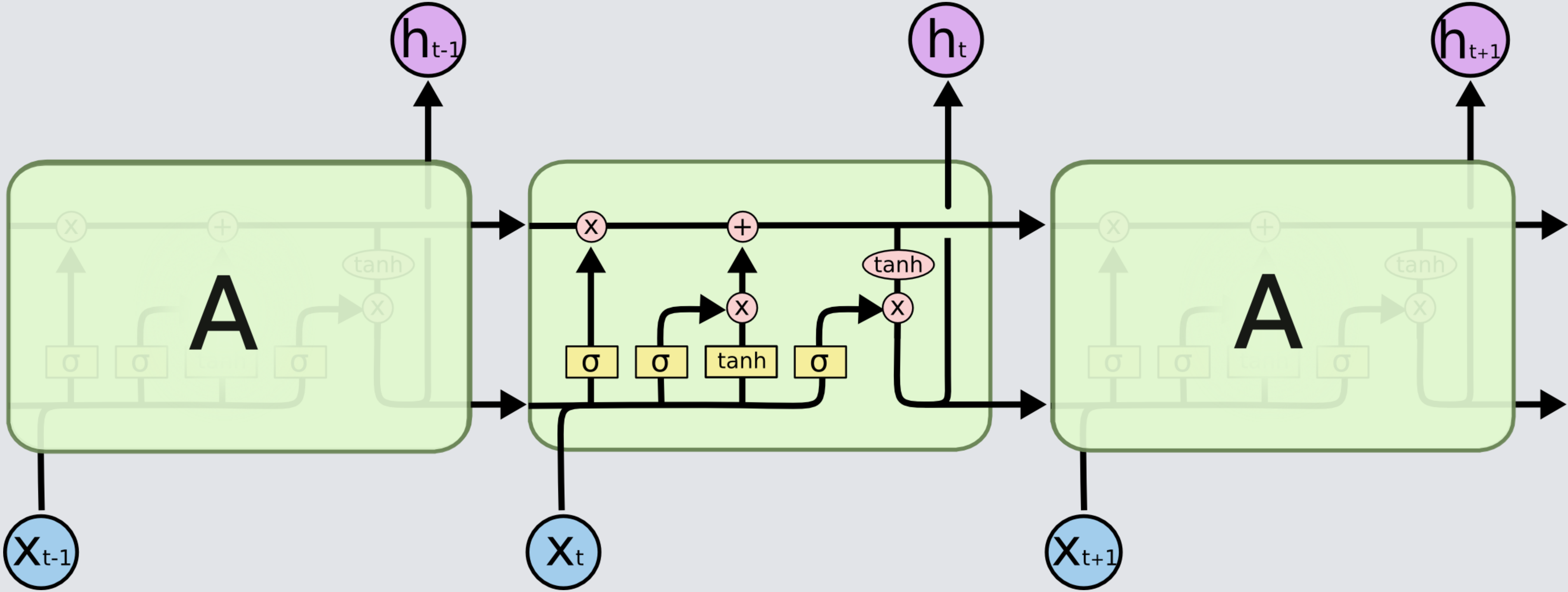
DNN for text: GRU

Standard RNNs have problems with long sequences, which is rather usual in textual data (vanishing gradient problem)

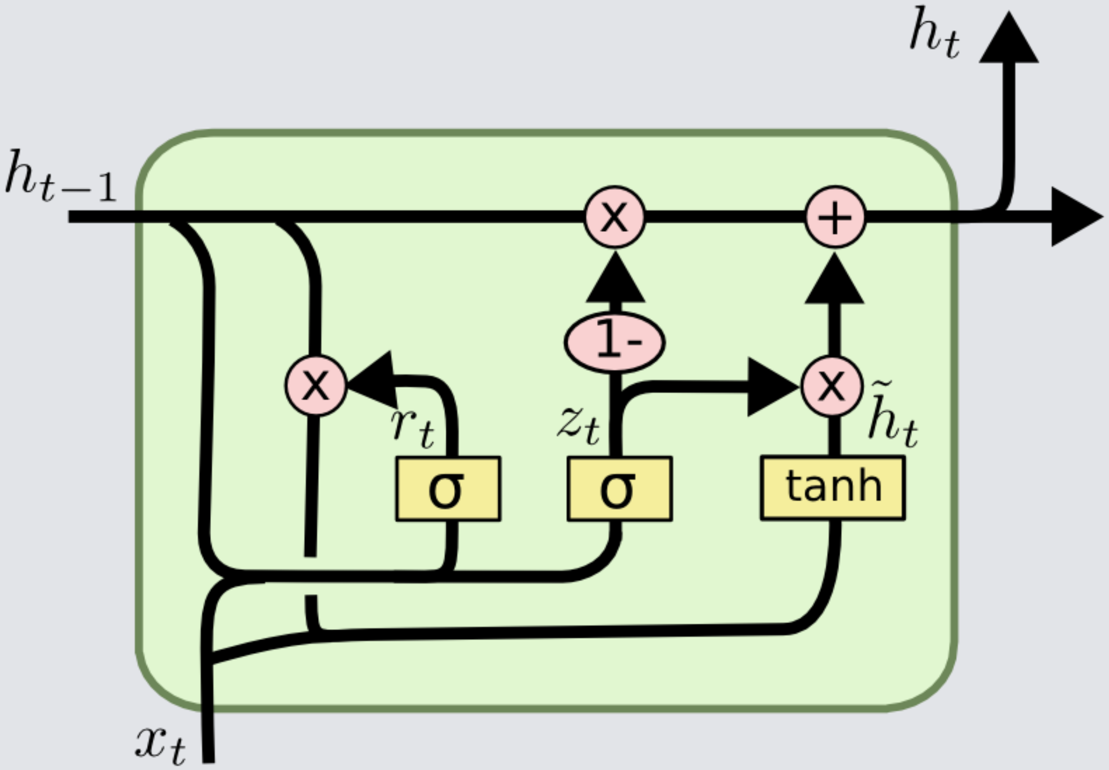
LSTM are able to solve this problem by implementing a form of long-term memory

GRU is a special kind of LSTM, faster to train and (usually) more efficient

Brief introduction to LSTM



GRU



Simpler than normal LSTM

Combines forget and input gate into a single «update layer»

Merges cell states and hidden states into one

Reset gate decides whether the previous cell is important or not

Multimodal network

Data fusion

Framework in which multiple data source are mixed to make them more useful for a specific application

It can mix information to extract more knowledge and have a better form of abstraction of the problem

The goal is to improve the system performances both in terms of accuracy and robustness

Particularly useful in robotics or computer vision in general

Late vs early fusion

In the literature the distinction between early and late fusion depends on the position of the fusion process

Late: it fuses the results of two or more different computational streams that work in parallel (this case)

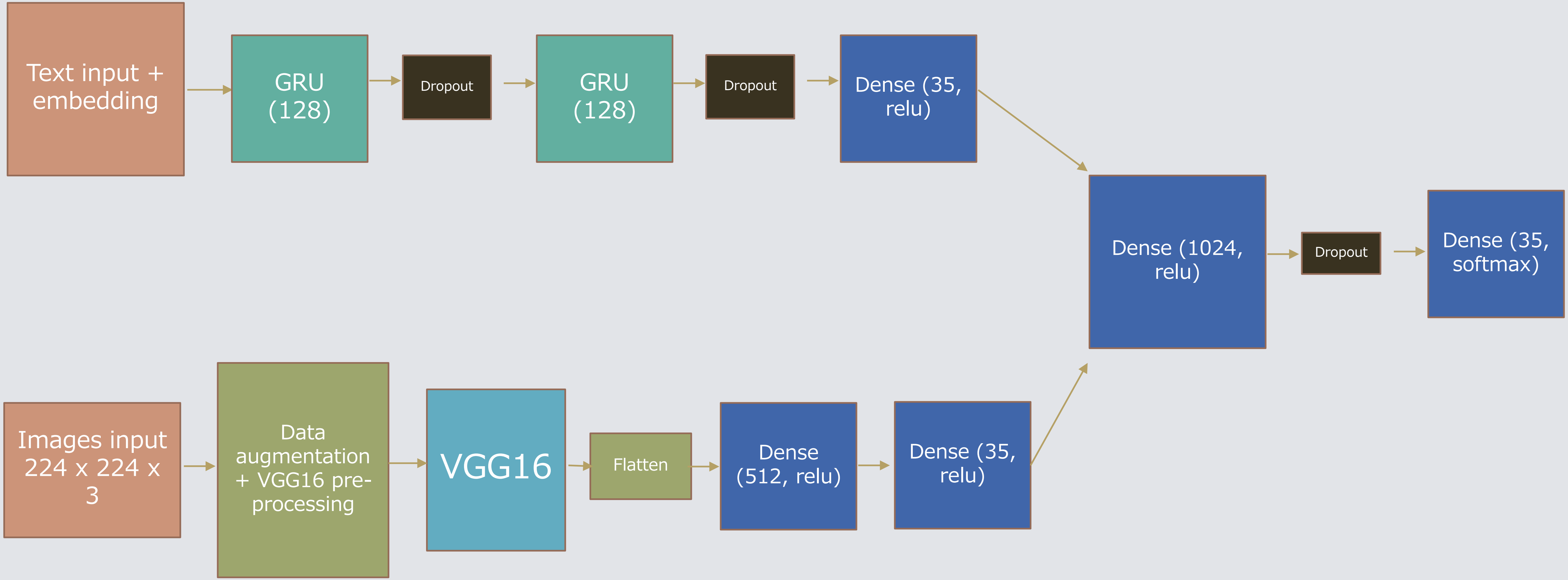
Early: it merges data sources at the beginning of the process to create a joint representation of the data

Keras implementation

The project was implemented with Keras high level APIs, splitting the original dataset into training, validation and test sets

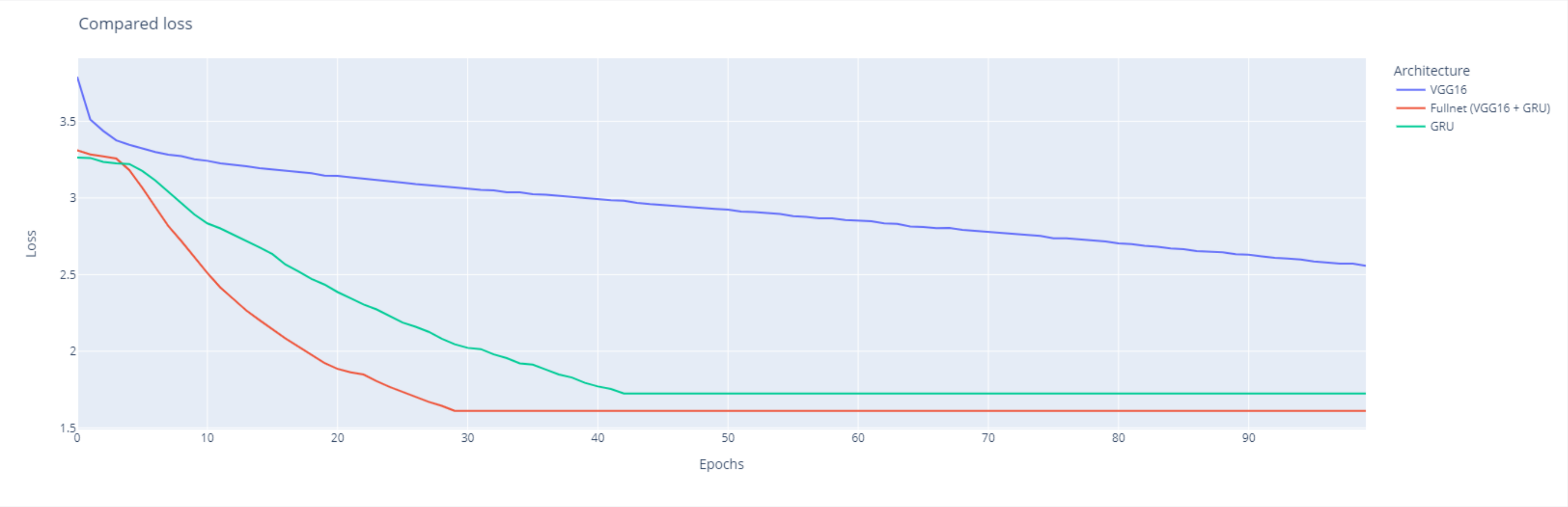
For VGG16 and the GRU model the standard Generator class was used to feed the network batches of data without saturating memory	The multimodal network required the development of a custom generator in order to feed mixed type of data to the network	Training was performed on Google Colab and Amazon Sagemaker Studio Lab (Tesla K80 vs Tesla T4)	Standard hyperparameters (Adam optimizer, batch size of 32, categorical cross-entropy as loss function) and early-stopping to avoid excessive overfitting

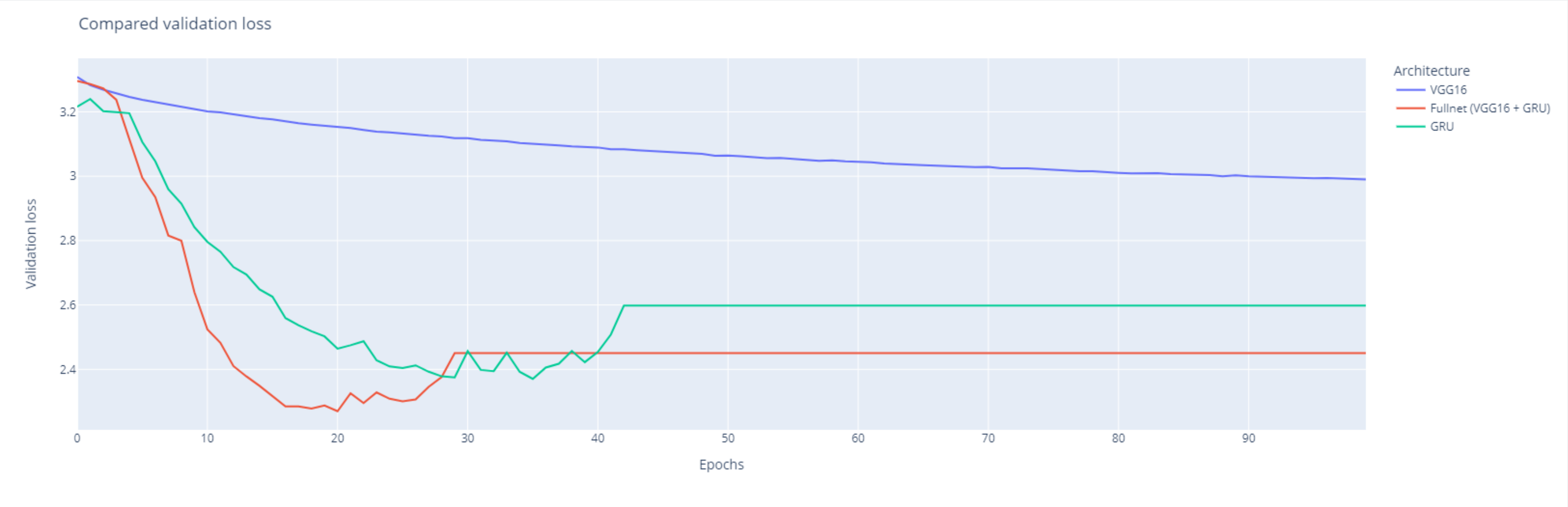
The multimodal network (VGG16 + GRU)



Results

Compared losses



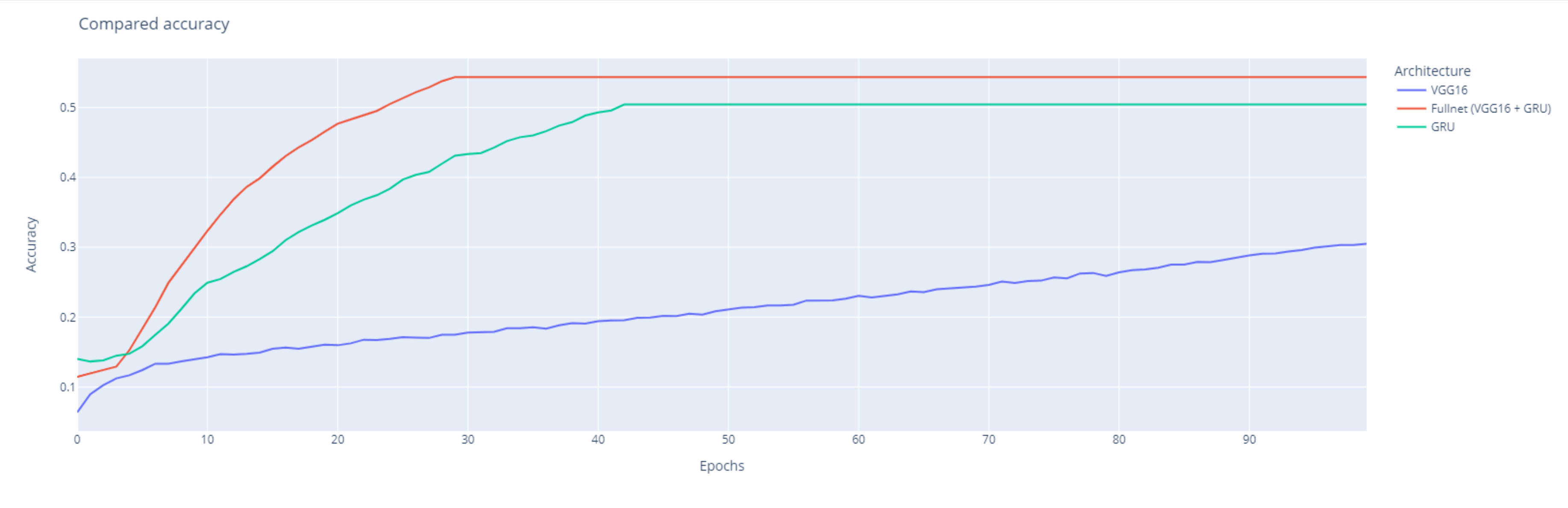


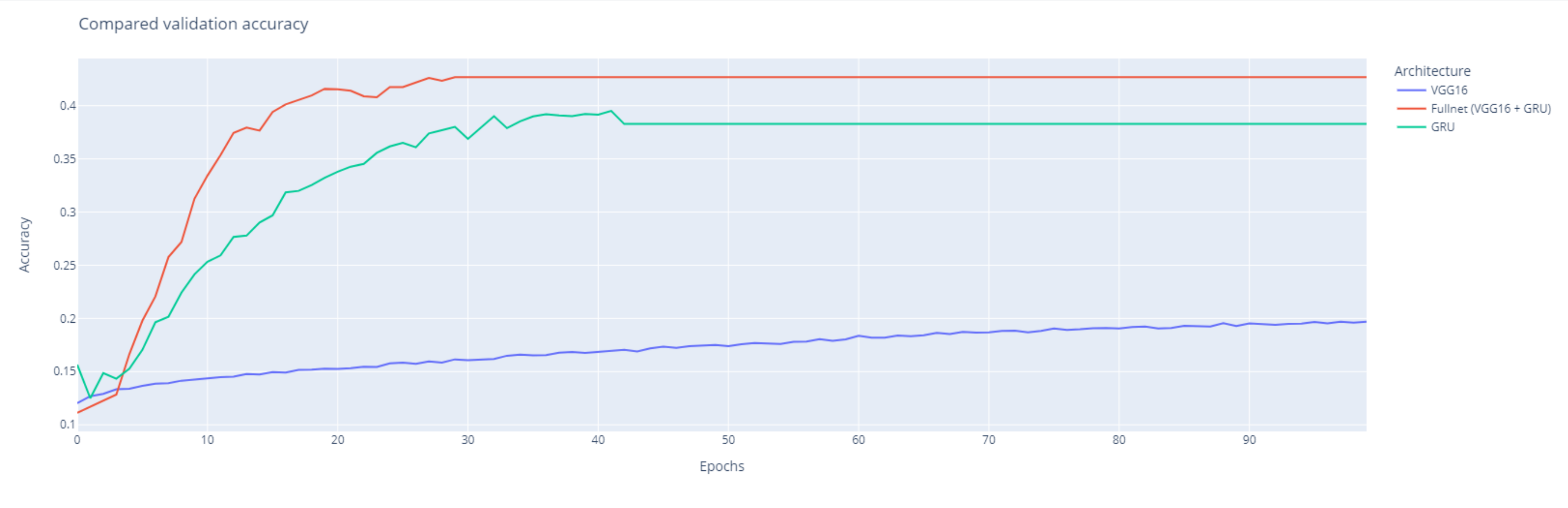
VGG16 requires a much slower learning rate to gain adequate performances in training

GRU and multimodal network are much faster and able to reach lower loss both in training and validation

Early stopping prevents excessive overfitting

Compared accuracy

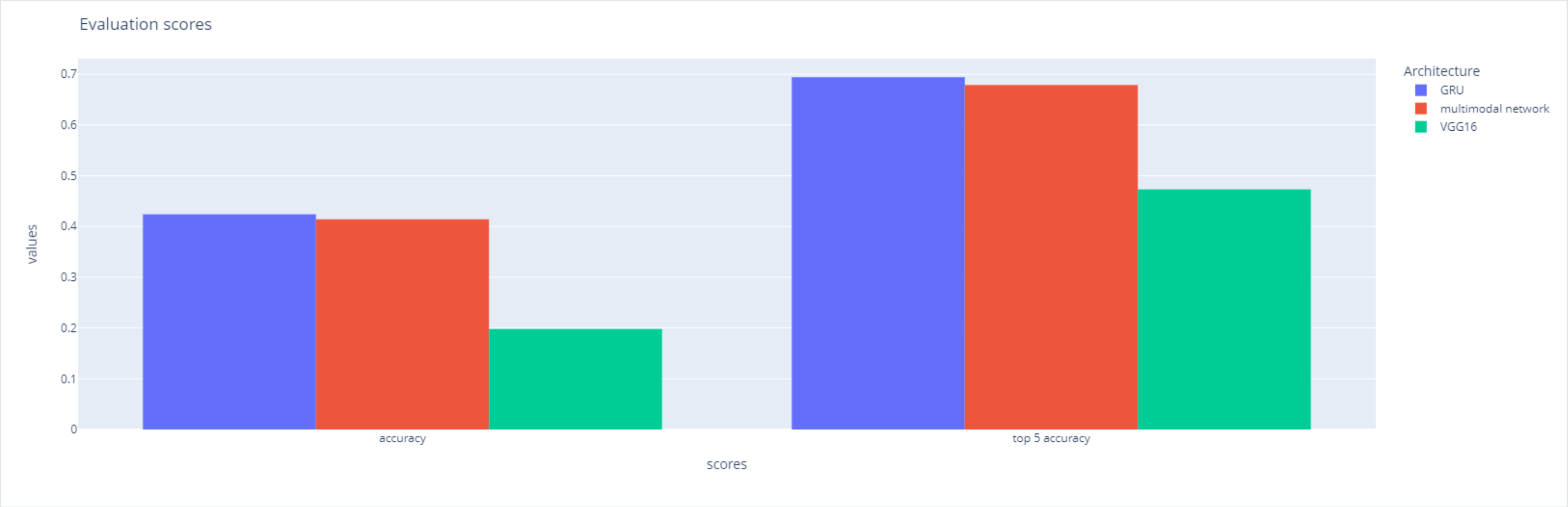


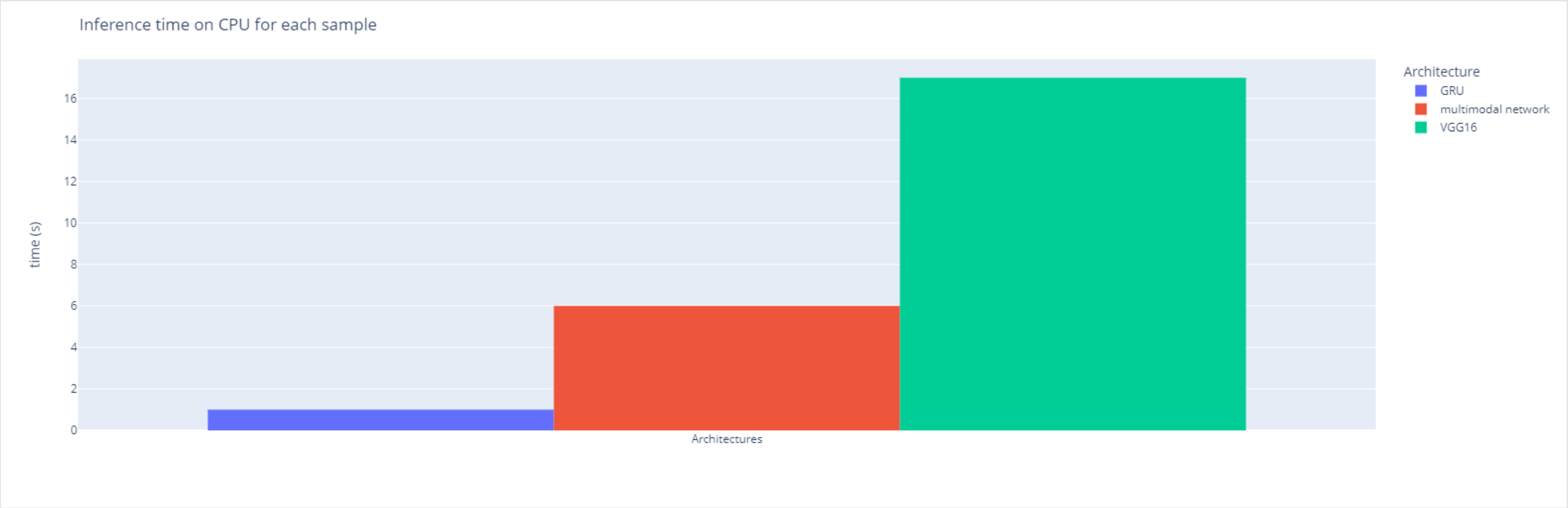


VGG16 is outperformed by both GRU and multimodal netowkr

Multimodal network is able to reach better accuracy both in training and validation

However training multimodal network is computationally more expensive than training GRU





Evaluation metrics on testing set

GRU is slightly more accurate on the testing set

If we consider top 5 accuracy the difference between GRU/multimodal and VGG16 is less significant

Inference time increase for the multimodal network, but not as dramatically expected looking at how VGG16 performs on its own

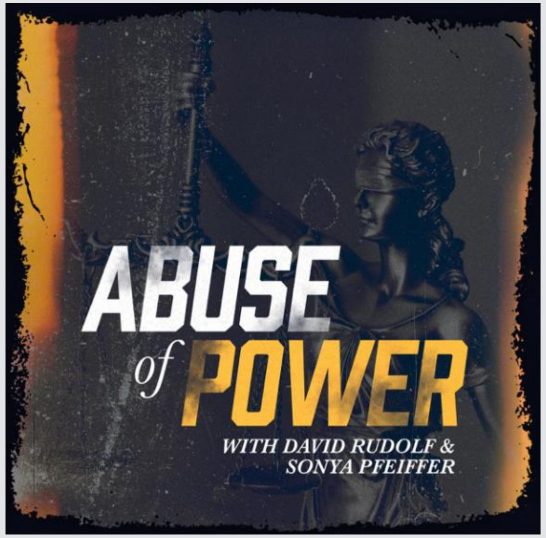
Predictions: VGG16

Predicted: Sports
Real: News

Predicted: Religion
Real: Vehicles

Predicted: Society & Culture
Real: Religion

Predicted: Sports
Real: Documentary



Predicted: Business
Real: Religion

Predicted: Sports
Real: Society & Culture

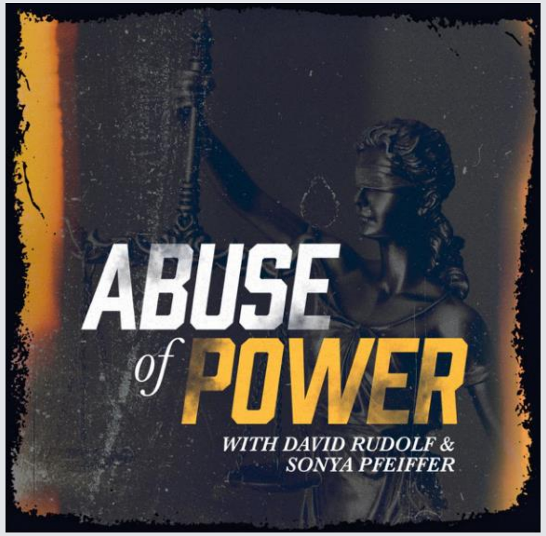
Predictions: GRU

Predicted: Sports
Real: News

Predicted: Religion
Real: Vehicles

Predicted: Religion
Real: Religion

Predicted: Interviews
Real: Documentary



Predicted: Fiction
Real: Religion

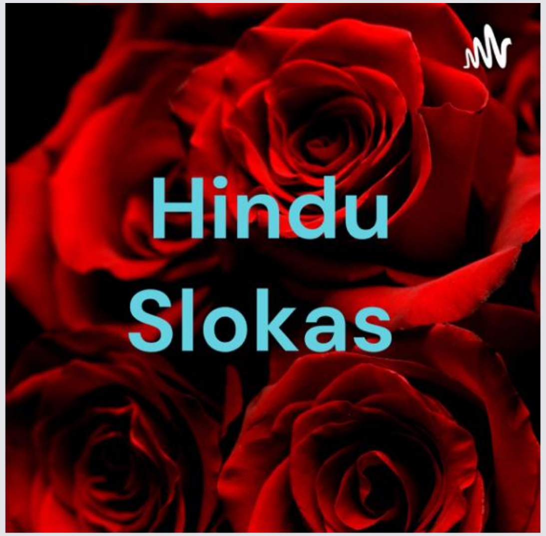
Predicted: TV & Film
Real: Society & Culture

Predictions: multimodal network

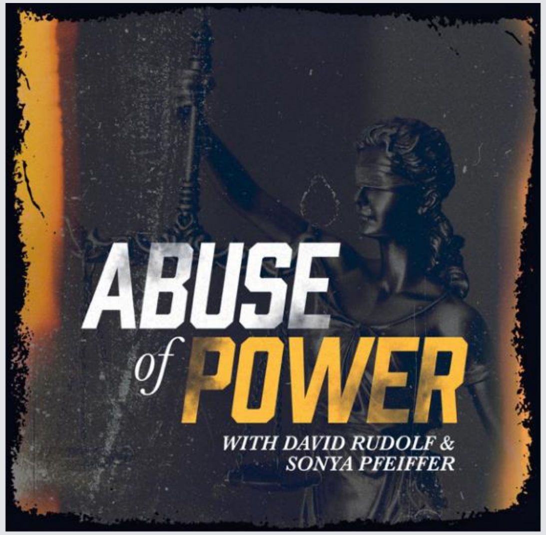
Predicted: True Crime
Real: News

Predicted: Vehicles
Real: Vehicles

Predicted: Vehicles
Real: Religion



Predicted: Arts
Real: Documentary



Predicted: Religion
Real: Religion

Predicted: Politics
Real: Society & Culture

Balance vs unbalanced metadata

«Chasing the Horizon is a podcast by, about and for motorcyclists. We talk to motorcycle industry figures, technical experts and riders just like you no matter what bike they love to ride. Please sign up on the mailing list at http://tinyletter.com/chasingthehorizon and subscribe in popular podcast apps or listen to every episode at http://chasingthehorizon.us»

VGG16: Religion
GRU: Religion
Multimodal: Vehicles

«Hindu slokas»

VGG16: Society & Culture
GRU: Religion
Multimodal: Vehicles

Conclusions

The multimodal model converges faster (in terms of epochs) during training and it tends to have better accuracy

However, the GRU model performs a little better on the testing set: this might be due to the fact that the dataset is largely unbalanced and so is the testing set

The multimodal approach seems to help in cases in which the two features are likely mixed, but it might lead to mistakes in unbalanced cases (e.g. less text and confusing image)

Labels were manually and arbitrarily aggregated; we might get better results depending on how genres were merged

Further experiments with larger and more balanced dataset could improved performances

Bibliography

Miller, Stuart J., et al, «Multi-Modal Classification Using Images and Text», SMU Data Science Review 3.3 (2020): 6.

Gadzicki, Konrad, et al., «Early vs late fusion in multimodal convolutional neural networks», 2020 IEEE 23rd International Conference on Information Fusion (FUSION), IEEE, 2020.

Li, Jonathan, Di Sun, and Tongxin Cai, «Genre Classification via Album Cover».

Ofli, Ferda, Firoj Alam, and Muhammad Imran, «Analysis of social media data using multimodal deep learning for disaster response», arXiv preprint arXiv:2004.11838 (2020).

Bibliography

Yin, Wenpeng, et al., «Comparative study of CNN and RNN for natural language processing», arXiv preprint arXiv:1702.01923 (2017).

Ahmed, Eman, and Mohamed Moustafa, «House price estimation from visual and textual features», arXiv preprint arXiv:1609.08399 (2016).

Nawaz, Shah, et al., «Are these birds similar: Learning branched networks for fine-grained representations», 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ), IEEE, 2019.