**Why not using LSTM for surgical phase recognition or why should we use them?**

# Intro

In recent years, surgical phase recognition has gained more interest from researchers and the industry with new state-of-the-art results. CNNs have now since almost a decade outperformed most of previous computer techniques. While Long Short-Term Memory (LSTM) models, a special type of RNN, have brought new state-of-the-art results in video recognition. Recently, new types of architectures have arisen from research, called Transformers which are using Attention mechanism to look at particular tokens in sequential data. This approach is therefore very effective on video data. Here we are discussing the evolution of those different techniques to see which one should be good candidates for surgical phase recognition in the field of surgical data science.

# Long Short-Term Memory (1997)

LSTMs are a special type of Recurrent NN with gated parameters to choose what to remember. They mitigate the exploding and vanishing gradient problem of RNN. With three gates:

* *Input gate:* update the cell state.
* *Forget gate:* how much should we forget or keep.
* *Output gate:* what the next hidden state should be.

Those gates allow to select what to remember in previous layers. All have sigmoid or tanh activations.

### Pros:

* Capture long term dependencies through time while avoid vanishing gradient problem.
* Good for sequential patterns.
* Good for information embedded in time.

# Nicolas Padoy’s LSTM for Surgical Phase Recognition

## Less is More: Surgical Phase Recognition with Less Annotations through Self-Supervised (2018)

### Pros:

* Semi-supervised approach
  + Self-supervised pre-training
* Scalable for different surgeries.
* Pre-trained CNN and LSTM (EndoN2N).

## Weakly Supervised Convolutional LSTM Approach (2019)

### Pros:

* Weakly supervised on binary presence of tools.
* CNN + Convolutional LSTM.
* ConvLSTM outperforms the CNN baseline tracker for spatio-temporal coherence of sequential image frames (surgical video).

# Conclusion

Based on the selected papers, it seems that during those last few years the research community has shifted its attention towards Transformers in spite of the good performance of CNNs for computer vision tasks. Indeed, Transformers have shown good results for NLP and are now also very suitable for visual tasks. While, LSTMs are still showing competitive performance, they haven’t really evolved as much as other new architectures.

Another trend is going towards self-supervision learning. Some models have shown comparable performances (DINO) to supervised learning approaches. Giving the time and cost of labelling large data set, searching for new self-supervised ways seems more relevant.

# Some References

# Long Short Term memory, Sepp Hochreiter, 1997

# Addressing Some Limitations of Transformers With Feedback Memory (2020)

# Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity (2021)

# Longformer: The Long-Document Transformer (2020)

Scales linearly with the sequences length vs. quadratically.

# Rethinking Attention with Performers (2020)

Nearly unbiased estimation of the attention matrix.

# Nyströmformer: A Nyström-Based Algorithm for Approximating Self-Attention (2021)

# RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019)

# Perceiver: General Perception with Iterative Attention (2021)

# Big Transfer (BiT): General Visual Representation Learning (2019)

Good for transfer learning problem, cnn.

# MLP-Mixer: An all-MLP Architecture for Vision (2021)

# Involution: Inverting the Inherence of Convolution for Visual Recognition (2021)