Progressive Token Length Scaling in Transformer Encoders for Efficient Universal Segmentation




Abhishek Aich, Yumin Suh,
Samuel Schulter, Manmohan Chandraker

NEC Laboratories **America**

UC San Diego
JACOBS SCHOOL OF ENGINEERING

ICLR
International Conference On Learning Representations

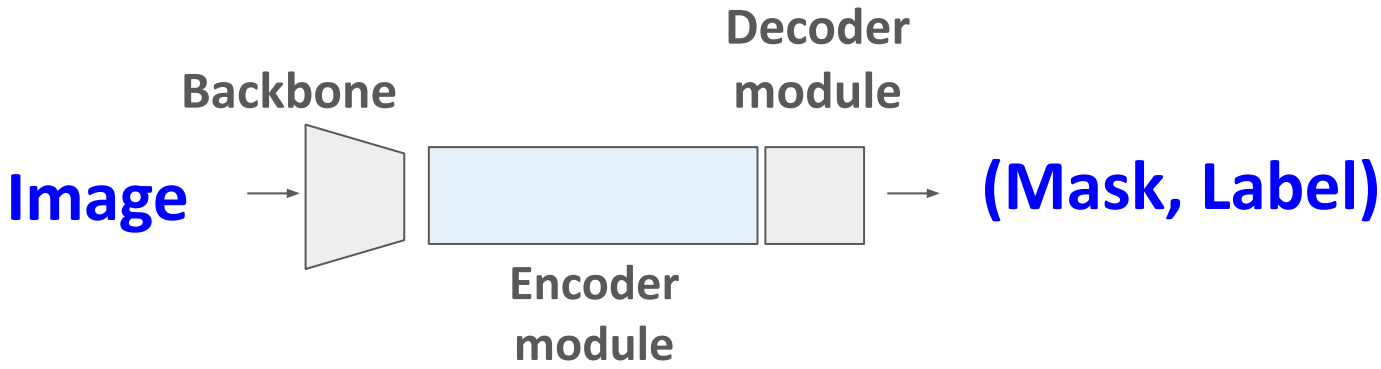
Image

Transformer Architecture

Dense perception

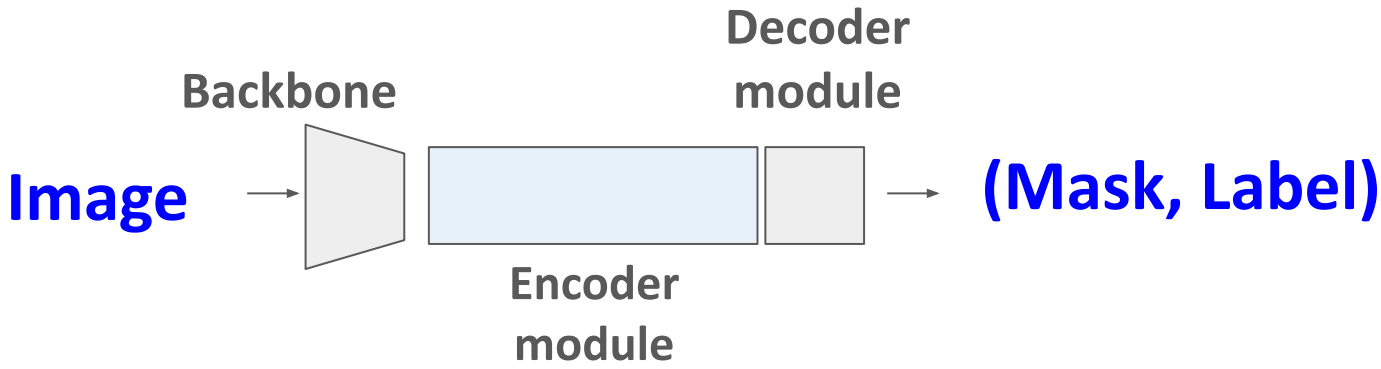
We have great performance from ViTs, can we make them efficient?

Mask2Former [Ref. 1]:



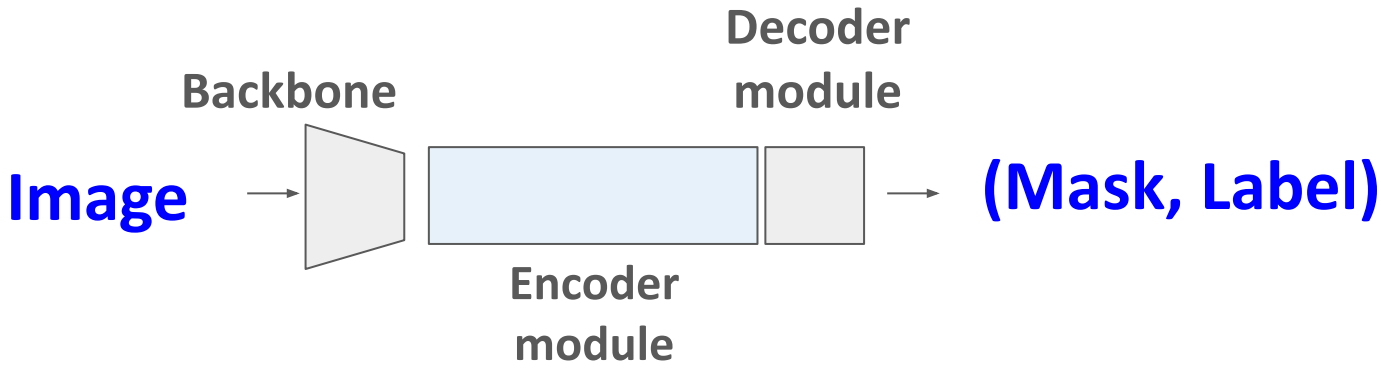
- Backbone = it's a simple feature extractor from images. For example, any CNN that provides difference scales of features.

Mask2Former [Ref. 1]:



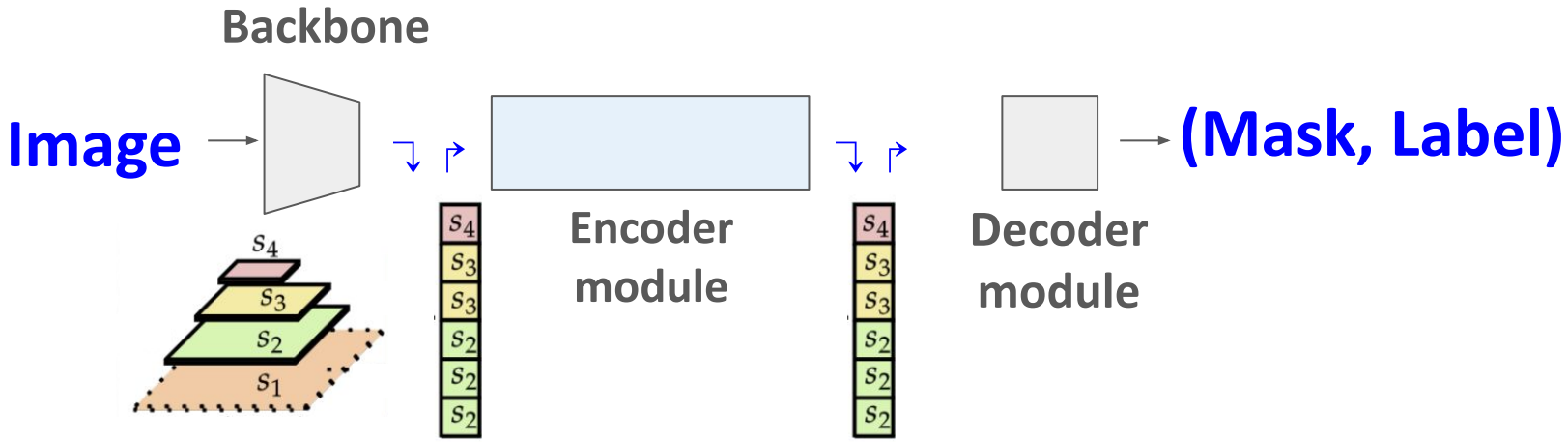
- Encoder module = Make all the multi-scale features “attend” to each other.

Mask2Former [Ref. 1]:

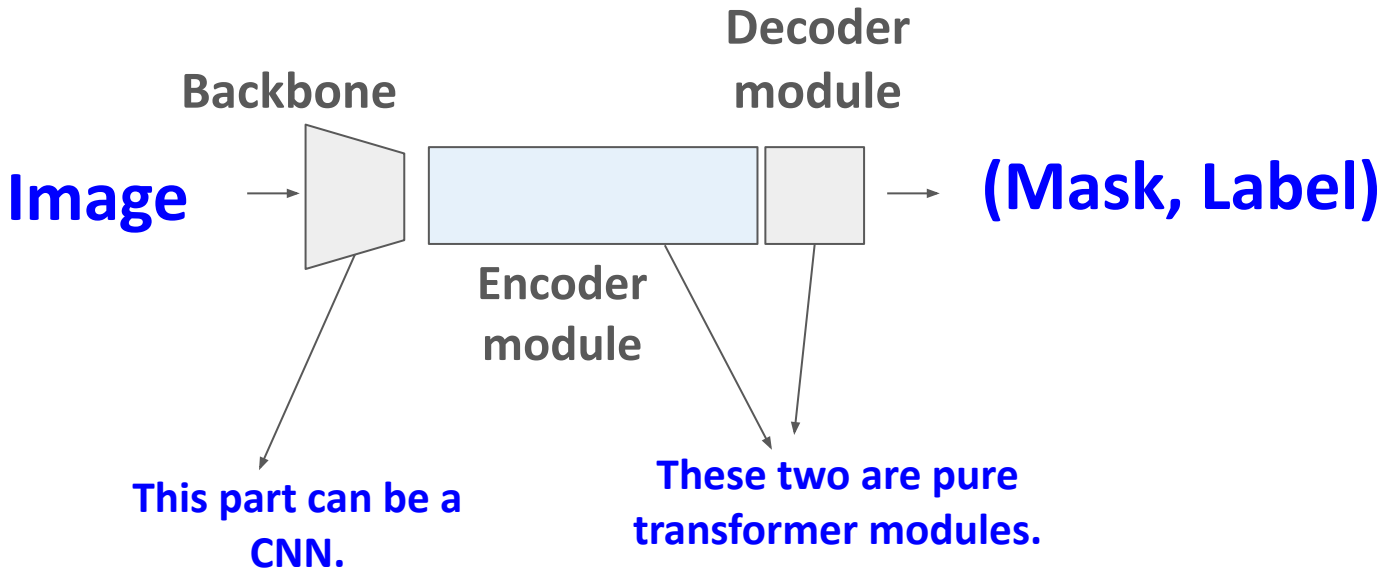


- Decoder module = Take the “enhanced” features from encoder and decode them into masks and corresponding labels.

Mask2Former [Ref. 1]:

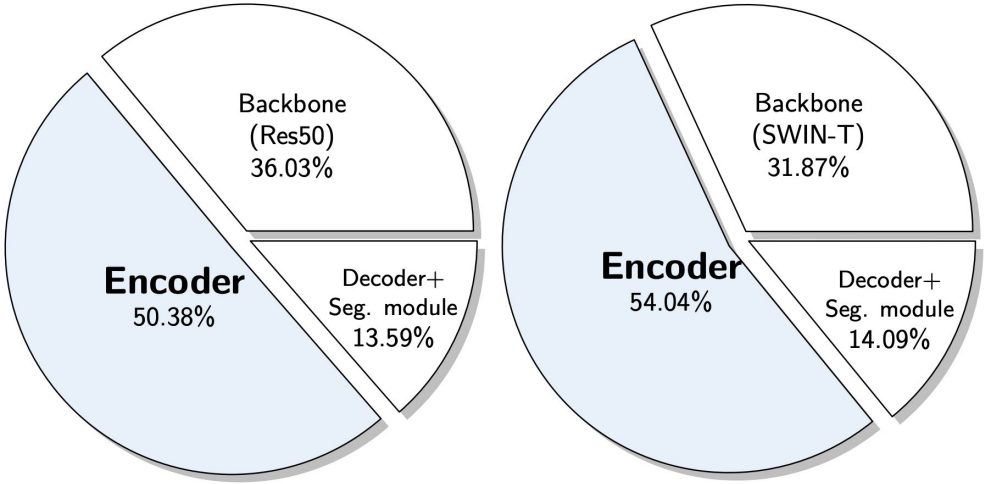


Mask2Former [Ref. 1]:



Mask2Former (M2F):

- **Unfortunately, for Mask2Former, good performance comes at a price of expensive computations.**

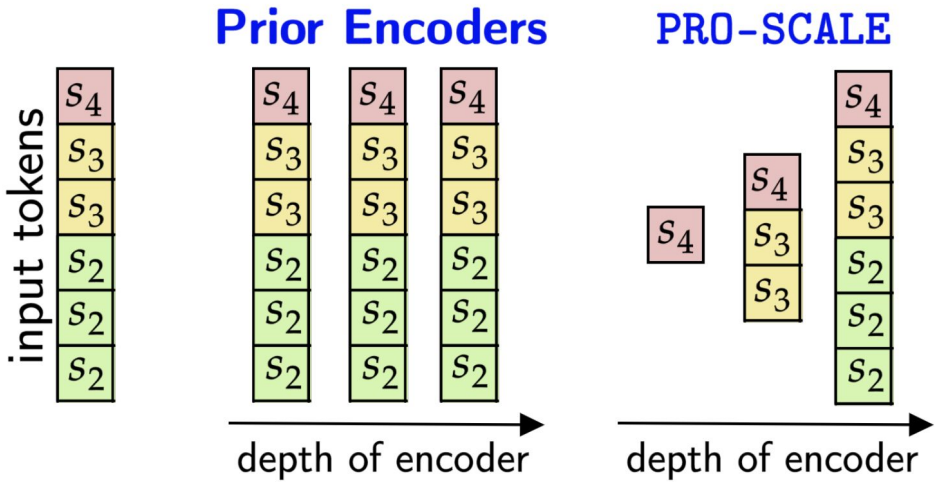


PRO-SCALE

- We propose a strategy termed PROgressive Token Length SCALing for Efficient transformer encoders (PRO-SCALE) that can be plugged-in to the Mask2Former segmentation architecture to significantly reduce the computational cost

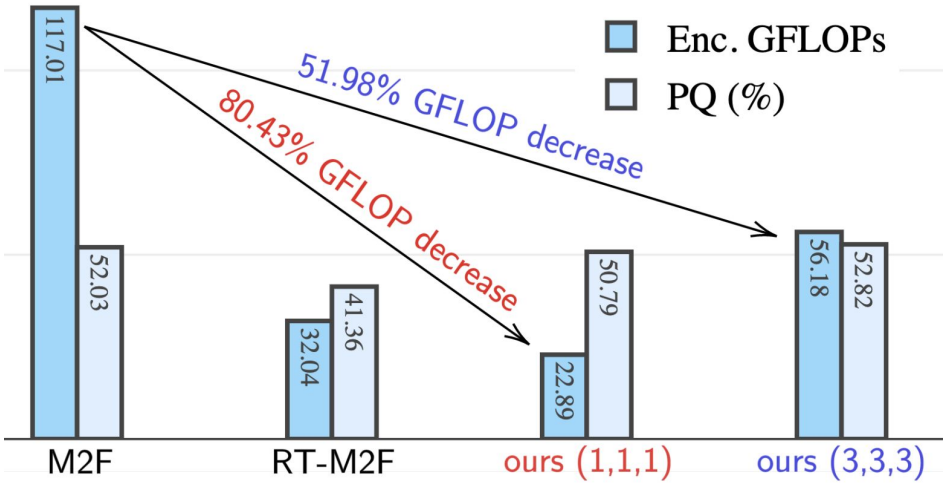
PRO-SCALE

- **PRO-SCALE: PROgressive Token Length SCALing for Efficient transformer encoders**



PRO-SCALE

- **PRO-SCALE works great !**
- **Maintains performance while reducing computations.**



Thank you!

For code and paper,

scan this QR code to Github ⇨