A head que é utilizada para vqa é esta:

https://github.com/clip-vil/CLIP-ViL/blob/d8061577be7c7912564958d1995dd205c5438ad2/CLIP-ViL-Pretrain/src/lxrt/modeling.py#L727

Text

Description automatically generated

O backbone do modelo é este:

https://github.com/clip-vil/CLIP-ViL/blob/d8061577be7c7912564958d1995dd205c5438ad2/CLIP-ViL-Pretrain/src/lxrt/modeling.py#L942

Text

Description automatically generated

O encoder que é utilizado para fazer fusing das visual e language features é este:

https://github.com/clip-vil/CLIP-ViL/blob/d8061577be7c7912564958d1995dd205c5438ad2/CLIP-ViL-Pretrain/src/lxrt/modeling.py#L539

Text

Description automatically generated with medium confidence

Esse é o encoder que é utilizado para fazer fusing das modalities

Como se pode vêr a layer utilizada para fazer o cross-encoding é esta:

Text

Description automatically generated

No visual config os valores default para o l\_layers, x\_layers e r\_layers são essas:

Graphical user interface, application

Description automatically generated

Estou a ter dificuldade em responder ao professor porque está tudo dependente do config e nenhum dos valores estão diretamente no código.

A configuração que é utilizada está aqui no ficheiro pretrain.bash:

https://github.com/clip-vil/CLIP-ViL/blob/d8061577be7c7912564958d1995dd205c5438ad2/CLIP-ViL-Pretrain/scripts/pretrain.bash

Graphical user interface, text, application, email

Description automatically generated

Utilizam 12 layers para o language encoder (parâmetro –llayers) ,0 para o cross-encoder (--xlayers) e 0 para o object-relationship encoder (--rlayers) , segundo o que aí está nesse ficheiro, mas penso que isso aí é apenas para o language-encoder.

No ficheiro param.py

<https://github.com/clip-vil/CLIP-ViL/blob/d8061577be7c7912564958d1995dd205c5438ad2/CLIP-ViL-Pretrain/src/param.py#L240>

Text

Description automatically generated

Parece-me que os valores do llaeyrs, xlayers e rlayers não estão dependentes do

Não consigo ver diretamente quantas heads estão a ser utilizadas porque não consigo ver diretamente através do código, mas a config que é passada para o lxmert é a de um ficheiro do hugging face e a configuração é a “bert-base-uncased”.

Graphical user interface, text, application

Description automatically generated

<https://github.com/clip-vil/CLIP-ViL/blob/d8061577be7c7912564958d1995dd205c5438ad2/CLIP-ViL-Pretrain/src/pretrain/lxmert_pretrain.py#L102>

Parece que a config que está a ser utilizada é essa de “bert-base-uncased”.

Por isso com base no que está no VisualConfig acho que o número de layers para o Cross-Encoder é de 5 porque é o que está no VisualConfig para o xlayers e que o número de heads é o da config ‘bert-base-uncased’, ou seja de 12.

Mas vou tentar correr o código para confirmar.