Assignment 2: Image-Text Multi-label Classification

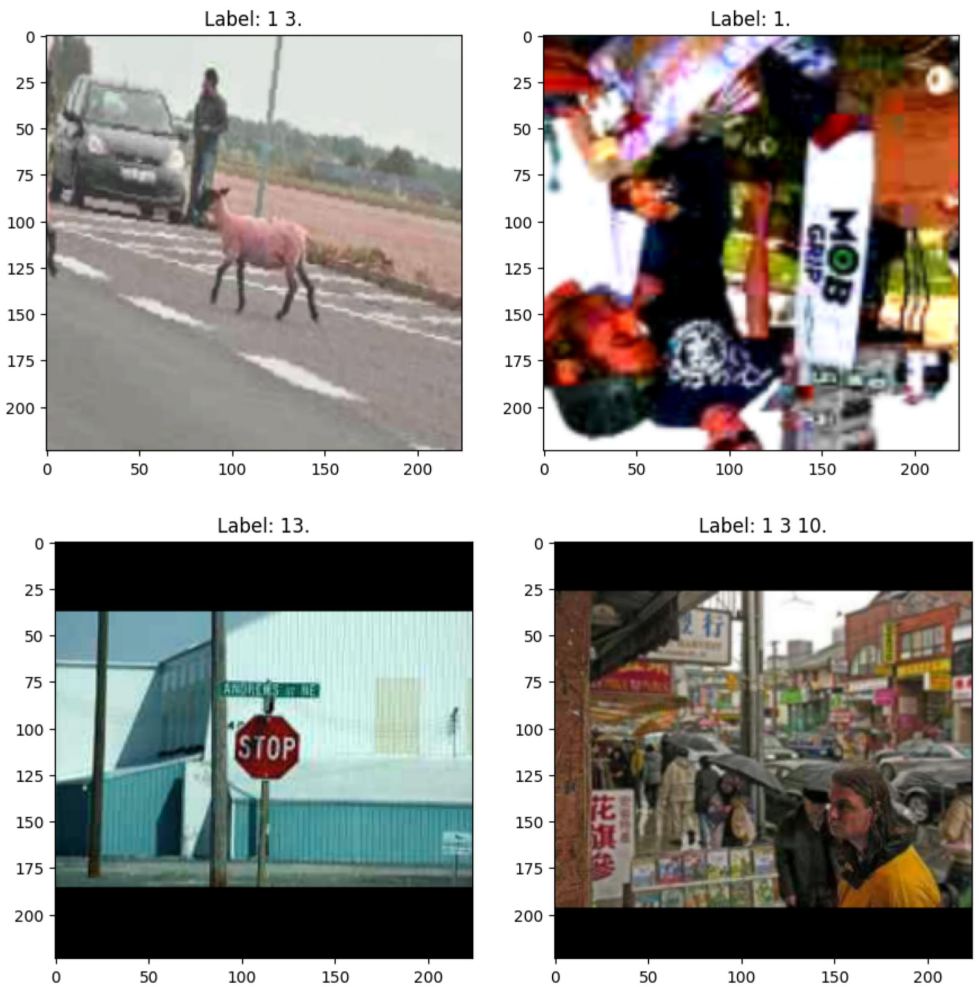
THE UNIVERSITY OF SYDNEY

Jiarui Xu 520347753

COMP5329 Deep Learning

Data augmentation and normalisation

Image



Text

	Text	Actucal input length	Labels
0	while down airborne a mountain. becomes skiing...	12	1
1	during races competition skiing downhill a ski...	10	1
2	to talking man on next a woman. a a phone	13	1
3	purple a a large hat. dog pimp wearing	12	18
4	is truck in parked of houses. front red a	12	8
5	street. busy double a decker bus a city down t...	13	1 2 3 6
6	in a hand umbrella has a women that her	11	1
7	across a surfers splashes water. as the paddle...	15	1
8	player bat ready to at a a game baseball getting	12	1
9	game front of room. and scooter a in of a a pi...	18	2 4

Unimodal models

Image classifiers

Model	Size(MB)	Train/Val	Thresh.	F1 score	Ep.	Efficiency(sec./ep.)
ResNet-18[3]	42.74	Training Validation	0.5	0.6712 0.6612	20	199.75
ResNet-34[3]	81.36	Training Validation		0.6406 0.6304		157.99
DenseNet-201[1]	70.45	Training Validation		0.6728 0.6594		179.58
ResNet-50[3]	90.12	Training Validation		0.7090 0.7063		175.89
		Training Validation		0.7283 0.7260	50	163.43

In the deployment of our unimodal models, ResNet-50 and BERT tiny have demonstrated superior performance as image and text classifiers respectively.

Text classifiers

Model	Size(MB)	Train/Val	Thresh.	F1 score	Ep.	Efficiency(sec./ep.)
TinyBert[2]	54.79	Training Validation	0.5	0.5955 0.5975	50	53.76
Bert tiny[6]	16.76	Training Validation	0.635	0.5960 0.5989		17.30

Data augmentation and normalisation are applied to enhance our model performance. For image-based data, augmentation is carried out through several methods including horizontal flipping, color alterations, and more. Subsequently, the normalised images undergo resizing and padding to ensure consistency.
In terms of text data, primarily random swapping is employed as a method of augmentation.

Multimodal models

Model	Size(MB)	Train/Val	Thresh.	F1 score	Ep.	Efficiency(sec./ep.)
DensityBert	97.71	Training Validation	0.35	0.8173 0.8173	50	191.48
MoDensityBert	97.72	Training Validation	0.38	0.8622 0.8179		178.95
CDBert-Text	93.81	Training Validation	0.29	0.7599 0.7564		181.51
CDBert-Image	91.02	Training Validation	0.461	0.8026 0.7985		147.75
WarmDBert	97.72	Training Validation	0.38	0.8505 0.8310		204.09
WarmerDBert	97.72	Training Validation	0.39	0.8567 0.8345		258.34
WarmerDBert extended	97.72	Training Validation	0.42	0.8698 0.8485	100	291.12
WWDBert	99.77	Training Validation	0.40	0.8700 0.8464		269.93
Bensity	100.83	Training Validation	0.33	0.7980 0.7980	50	190.01
Censity-Text	90.89	Training Validation	0.33	0.7905 0.7901		183.50
Censity-Image	81.14	Training Validation	0.38	0.7869 0.7801		174.49
ResT	100.92	Training Validation	0.38	0.7836 0.7766		170.06

From our experimental analysis, we can deduce the following key insights:
1. Self-attention outperforms cross-attention. 2. Image queries yield better results than text queries. 3. Layer normalisation enhances our model, but batch normalisation does not. 4. Two optimal unimodal models don't ensure a well-performing multimodal model. 5. Unfreezing more layers and widening the fully-connected layer boosts performance.

Optimal architecture

	Extractor	Number of unfrozen blocks/layers
Architecture	DenseNet-121[1]	2
	TinyBert[2]	2
	Application	**Value**
Threshold	Sigmoid function	0.39
	Dataset	**F1 score**
Performance	Training	0.8567
	Validation	0.8345
	Public leaderboard	~ 0.88