Continual Learning for Automated Audio Captioning Using The Learning Without Forgetting Approach

Jan Berg, Konstantinos Drossos

Continual Learning (CL)

- Sometimes referred to as *Incremental Learning* or *Lifelong Learning*
- Tackling the issue of catastrophic forgetting during further training models with new datasets
- Three categories
  - Regularizing
  - Rehearsal (e.g. generative models)
  - Dynamic Architectures

Motivation for CL in Automated Audio Captioning

- Disparities between datasets (e.g. because of different annotators)
  - ➢ Method optimized on a dataset will have problems when evaluated using different dataset
- Jointly training the model using all the datasets is not always possible

- Applying a continual learning method to further train the model so that performance on the model on original dataset does not degrade while also learning from the new dataset

Learning Without Forgetting (LWF)

- Regularization based Continual Learning
- Utilize output of the copy of the initial state of the model to calculate additional loss, $\mathcal{L}_{reg}$.
- Total loss becomes the sum of $\mathcal{L}_{reg}$ and $\mathcal{L}_{new}$
- λ is used to control the strength of each loss term
- In this research only common words between datasets were considered

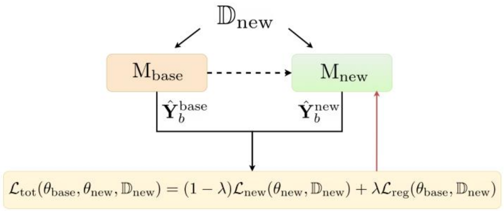

Figure 1: Learning Without Forgetting

Wavetransformer

- Transformer based model for AAC
- Utilizes the multi-head attention and positional encoding of Transformer to learn sequential information to generate Audio Captions

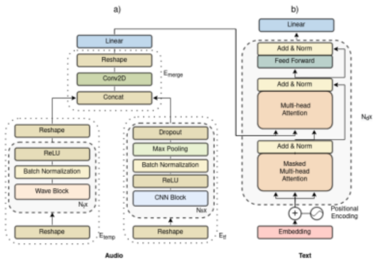

Figure 2: Wavetransformer model

Evaluation Datasets

- Clotho
  - 15-30 second audio clips
  - Five captions of eight to 20 words
  - 19 200, 5230, 5230 training, validation and evaluation samples
  - 4367 words unique words
- AudioCaps
  - AudioSet formatted
  - 10 second audio clips
  - 38 188, 2500, 4895 training, validation and evaluation samples
  - 4506 unique words

Results

- Highest scores for each of the dataset when using LwF shown in bold.
- Simple fine tuning approach:
  - $\mathbb{D}_{\mathbf{ori}}$ = 0.065, $\mathbb{D}_{\mathbf{new}}$ = 0.247
- Achieved some degree of continual learning. Example when B = 12 and λ = 0.80
  - $\mathbb{D}_{\mathbf{ori}}$ = 0.186, $\mathbb{D}_{\mathbf{new}}$ = 0.157
- Less learning on $\mathbb{D}_{\mathbf{new}}$, but keeps $\mathbb{D}_{\mathbf{ori}}$ closer to the original performance.

Baseline scenario	SPIDEr $\mathbb{D}_{\text{ori}}$	SPIDEr $\mathbb{D}_{\text{new}}$
$\text{WT}_{\text{cl-au}}$	0.182	0.108
$\text{WT}_{\text{au-cl}}$	0.318	0.102
$\text{WT}_{\text{cl-ft}}$	0.065	0.247

Figure 3: SPIDEr scores for baseline scenarios

batch size B	λ	SPIDEr $\mathbb{D}_{\text{ori}}$	SPIDEr $\mathbb{D}_{\text{new}}$
4	0.70	0.098	**0.239**
	0.75	0.102	0.215
	0.80	0.093	0.214
	0.85	0.115	0.230
	0.90	0.133	0.215
	0.95	0.155	0.192
	1.00	0.163	0.119
8	0.70	0.113	0.210
	0.75	0.119	0.223
	0.80	0.132	0.220
	0.85	0.133	0.190
	0.90	0.156	0.187
	0.95	0.178	0.157
	1.00	0.165	0.114
12	0.70	0.109	0.211
	0.75	0.160	0.197
	0.80	**0.186**	0.157
	0.85	0.171	0.179
	0.90	0.182	0.153
	0.95	0.185	0.145
	1.00	0.176	0.115

Figure 4: SPIDEr scores using LWF

Conclusions

- Achieved some degree of Continual Learning using LWF approach
- Future research potential with more sophisticated CL methods