Suggested Modifications

MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes

Let the original reward be defined as

$$r = \frac{1}{3}\Big(R_1^F + R_2^F + R_L^F\Big), \tag{1}$$

where R_1^F, R_2^F, and R_L^F denote the ROUGE-1, ROUGE-2, and ROUGE-L F-scores, respectively. To discourage overly long summaries, a length penalty is applied:

$$R_{\text{len}} = \frac{r}{T+1}, \tag{2}$$

with T representing the number of extracted sentences. However, this formulation does not explicitly penalize redundancy. To promote diversity, we introduce a penalty based on the cosine similarity between sentence embeddings. Let $\mathbf{e}(s_i)$ denote the embedding of sentence s_i produced by the Local Sentence Encoder (LSE), and define the cosine similarity function as

$$\phi(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}. \tag{3}$$

For the sentence s_{a_t} selected at time step t, compute its maximum similarity with any sentence in the current partial summary $\mathcal{S}_{t-1}$:

$$\gamma_t = \max_{s \in \mathcal{S}_{t-1}} \phi\Big(\mathbf{e}(s_{a_t}), \mathbf{e}(s)\Big). \tag{4}$$

The overall diversity penalty is then the average maximum similarity:

$$\Delta = \frac{1}{T} \sum_{t=1}^{T} \gamma_t. \tag{5}$$

Incorporating the diversity term, the modified reward becomes

$$r_{\text{div}} = r - \lambda \, \Delta, \tag{6}$$

where $\lambda > 0$ is a hyperparameter controlling the influence of the diversity penalty. Finally, combining both the length and diversity adjustments, the final return is defined as

$$R_t = \frac{r - \lambda \, \Delta}{T+1}. \tag{7}$$

This revised formulation encourages the MemSum model to select sentences that not only achieve high ROUGE scores but also minimize redundancy, leading to summaries that are both informative and diverse.

Self-Supervised Facial Representation Learning with Facial Region Awareness

Let the cosine similarity between a projected pixel feature and a facial mask embedding be defined as

$$S(m,u,v) = \frac{\boldsymbol{f}_{uv}^{\top}\boldsymbol{q}_m}{\|\boldsymbol{f}_{uv}\|\,\|\boldsymbol{q}_m\|}, \tag{8}$$

where $\boldsymbol{f}_{uv} \in \mathbb{R}^D$ denotes the projected feature at a pixel (for example, at location (u,v)) and $\boldsymbol{q}_m \in \mathbb{R}^D$ is the facial mask embedding corresponding to the m-th facial region, with $m = 1,\ldots,N$.

To generate the heatmap $M \in \mathbb{R}^{N\times H\times W}$, a temperature-scaled softmax is applied over the N facial regions:

$$M(m,u,v) = \frac{\exp(S(m,u,v)/\tau)}{\sum_{k=1}^{N}\exp(S(k,u,v)/\tau)}, \tag{9}$$

where $\tau = 0.1$ is a fixed temperature parameter. A smaller τ enforces a peaky (hard) assignment to a specific region, whereas a larger τ results in a smoother (softer) distribution.

In contrast to approaches that monitor per-pixel entropy, we promote diversity via a global measure. First, the average assignment probability for each region across the image is computed as

$$\bar{M}(m) = \frac{1}{HW}\sum_{u=1}^{H}\sum_{v=1}^{W} M(m,u,v). \tag{10}$$

Then, the global diversity score δ is defined as the entropy of the average assignment distribution:

$$\delta = -\sum_{m=1}^{N} \bar{M}(m)\log \bar{M}(m). \tag{11}$$

Maximizing δ encourages balanced utilization of all facial regions.

To integrate this mechanism into the training process, let the original self-supervised loss be denoted as L_{SS}. The overall modified loss is defined as

$$L_{\mathrm{mod}} = L_{\mathrm{SS}} - \lambda\,\delta, \tag{12}$$

with $\lambda = 0.01$ serving as a hyperparameter that controls the influence of the diversity term. This loss formulation not only guides the model to learn discriminative local features via L_{SS} but also ensures balanced region assignments by penalizing overly concentrated distributions.

Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation

To enhance the Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for medical image segmentation, we propose two modifications: (1) a combined binary cross-entropy (BCE) and Dice loss function, and (2) attention gates in skip connections. These changes aim to improve segmentation performance while remaining easy to implement within the existing framework. The original R2U-Net employs binary cross-entropy (BCE) loss, defined as

$$\mathrm{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right], \tag{13}$$

where $y_i \in \{0, 1\}$ is the ground truth label for pixel i, $\hat{y}_i \in [0, 1]$ is the predicted probability, and N is the total number of pixels. While effective, BCE may struggle with imbalanced datasets common in medical imaging, where foreground regions (e.g., lesions) are smaller than the background. To address this, we introduce the Dice coefficient, a standard metric for segmentation overlap, defined as

$$\delta = 2\frac{|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|}, \tag{14}$$

where Y is the ground truth mask and $\hat{Y}$ is the predicted mask. The Dice loss is then $1 - \delta$. We propose a combined loss function:

$$L = \alpha \cdot \mathrm{BCE} + \beta \cdot (1 - \delta), \tag{15}$$

where $\alpha > 0$ and $\beta > 0$ are hyperparameters weighting the contributions of BCE and Dice loss, respectively. For initial experiments, set $\alpha = 1.0$ and $\beta = 1.0$, with tuning based on validation results. In addition, R2U-Net uses skip connections to concatenate encoder feature maps with decoder feature maps, transferring all features equally. To focus on relevant regions, we incorporate attention gates. Let $\boldsymbol{x}_l \in \mathbb{R}^{C\times H\times W}$ denote the encoder feature map at layer l, and $\boldsymbol{g}_l \in \mathbb{R}^{C'\times H'\times W'}$ the gating signal from the corresponding decoder layer. The attention coefficients $\alpha_l \in [0, 1]^{H\times W}$ are computed as

$$\alpha_l = \sigma\left(\boldsymbol{W}_\alpha^\top\left(\sigma\left(\boldsymbol{W}_x^\top \boldsymbol{x}_l + \boldsymbol{W}_g^\top \boldsymbol{g}_l + \boldsymbol{b}\right)\right)\right), \tag{16}$$

where σ is the sigmoid function, $\boldsymbol{W}_x$, $\boldsymbol{W}_g$, and $\boldsymbol{W}_\alpha$ are learnable weight matrices, and $\boldsymbol{b}$ is a bias term. Note that $\boldsymbol{x}_l$ and $\boldsymbol{g}_l$ may require resizing (e.g., via upsampling or downsampling) to match dimensions, followed by a 1×1 convolution to align channel sizes. The attended feature map is then

$$\boldsymbol{x}_l' = \alpha_l \odot \boldsymbol{x}_l, \tag{17}$$

where $\odot$ denotes element-wise multiplication. Finally, $\boldsymbol{x}_l'$ is concatenated with the decoder feature map. These modifications—a combined BCE-Dice loss and attention-enhanced skip connections—aim to improve R2U-Net's performance on medical image segmentation tasks by addressing class imbalance and enhancing feature relevance. Both changes are straightforward to implement in frameworks like Keras or TensorFlow, maintaining the model's efficiency.