self-supervised contrastive video-speech representation learning for ultrasound

1. Method

positive pair (V_T, S_T)

negative pair $(V_T, S_{T'})$, $T' = T + \delta$

(1) binary classification loss

$$
\begin{tikzcd}
v_t \arrow[r, "g_v"] & g_v(v_t) \arrow[dr] & \\
 & & f(v_t, s_t) = g_v(v_t) \oplus g_s(\eta(s_t)) \arrow[d, "\text{binary classification}"] \\
\eta(s_t) \arrow[r, "g_s"] & g_s(\eta(s_t)) \arrow[ur] & \mathcal{L}_{cls}
\end{tikzcd}
$$

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i}^{C} c_i^n log(f(v_t, s_t)_i^n),$$

(2) cross-modal contrastive learning

projected embeddings:

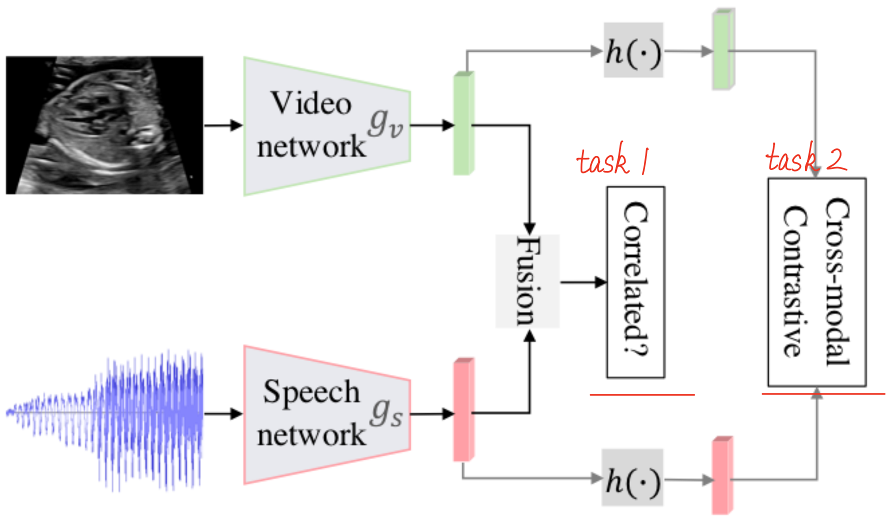
$y_v = h(g_v(v_t))$, $y_s = h(g_s(\eta(s_t)))$

cross-modal contrastive objective:

embedding of positive pair : similar

embedding of negative pair : repel

$$\mathcal{L}_{cont} = -log\frac{e^{sim(y_v, y_s)} - e^{sim(y_v, y_{s'})}}{\sum_{k=1}^{N}\mathbb{1}_{[k\neq v]}e^{sim(y_v, y_k)}},$$



$L = \alpha L_{cls} + \beta L_{cont}$

2. Experiments and implementation

g_s, g_v : ResNeXt-50 with Squeeze-and Excitation module and dilated convolutions.
same architecture, optimized seperately

gradient clipping

$\eta(s_t)$: preprocess of speech data
2D log-spectrogram representation of size 256×256
short-time Fourier transform (STFT) with 256 frequency bands,
10ms window length and 5ms hop length