Strong and Simple Baselines for Multimodal Utterance Embeddings: Supplementary Material

Paul Pu Liang*, **Yao Chong Lim***,
Yao-Hung Hubert Tsai, Ruslan Salakhutdinov, Louis-Philippe Morency
School of Computer Science, Carnegie Mellon University
{pliang,yaochonl,yaohungt,rsalakhu,morency}@cs.cmu.edu

1 Appendix

1.1 Proof of Theorem 1

We begin by restating the likelihood of a multimodal segment $\mathbf{s}$ under our model:

$$\mathbb{P}[\mathbf{s}|m_{\mathbf{s}}] \tag{1}$$
$$= \mathbb{P}[\mathbf{w}|m_{\mathbf{s}}]^{\alpha_{\mathbf{w}}}\mathbb{P}[\mathbf{v}|m_{\mathbf{s}}]^{\alpha_{\mathbf{v}}}\mathbb{P}[\mathbf{a}|m_{\mathbf{s}}]^{\alpha_{\mathbf{a}}} \tag{2}$$
$$= \prod_{w\in\mathbf{w}}\mathbb{P}[w|m_{\mathbf{s}}]^{\alpha_{\mathbf{w}}}\prod_{v\in\mathbf{v}}\mathbb{P}[v|m_{\mathbf{s}}]^{\alpha_{\mathbf{v}}}\prod_{a\in\mathbf{a}}\mathbb{P}[a|m_{\mathbf{s}}]^{\alpha_{\mathbf{a}}} \tag{3}$$

We define the objective function by the maximum likelihood estimator of the multimodal utterance embedding and the parameters. The estimator is obtained by solving the unknown variables that maximizes the log-likelihood of the observed multimodal utterance (i.e., $\mathbf{s}$):

$$\mathcal{L}(m_{\mathbf{s}}, W, b; \mathbf{s}) = \log \mathbb{P}[\mathbf{s}|m_{\mathbf{s}}; W, b] \tag{4}$$
$$= \sum_{w\in\mathbf{w}}\log\mathbb{P}[w|m_{\mathbf{s}}]^{\alpha_{\mathbf{w}}} + \sum_{v\in\mathbf{v}}\log\mathbb{P}[v|m_{\mathbf{s}}]^{\alpha_{\mathbf{v}}} + \sum_{a\in\mathbf{a}}\log\mathbb{P}[a|m_{\mathbf{s}}]^{\alpha_{\mathbf{a}}} \tag{5}$$

with W and b denoting all linear transformation parameters. Our goal is to solve for the optimal embedding $m_{\mathbf{s}}^* = \arg\max_{m_{\mathbf{s}}^*}\mathcal{L}(m_{\mathbf{s}}, W, b; \mathbf{s})$. We will begin by simplifying each of the terms: $\log\left(\mathbb{P}[w|m_{\mathbf{s}}]\right)^{\alpha_{\mathbf{w}}}, \log\left(\mathbb{P}[v|m_{\mathbf{s}}]\right)^{\alpha_{\mathbf{v}}}$, and $\log\left(\mathbb{P}[a|m_{\mathbf{s}}]\right)^{\alpha_{\mathbf{a}}}$.

For the language features, we follow the approach in (Arora et al., 2017). We define:

$$f_w(m_{\mathbf{s}}) \tag{6}$$
$$= \log\mathbb{P}[w|m_{\mathbf{s}}]^{\alpha_{\mathbf{w}}} \tag{7}$$
$$= \alpha_{\mathbf{w}}\log\mathbb{P}[w|m_{\mathbf{s}}] \tag{8}$$
$$= \alpha_{\mathbf{w}}\log\left[\alpha p(w) + (1-\alpha)\frac{\exp\left(\langle w, m_{\mathbf{s}}\rangle\right)}{Z_{m_{\mathbf{s}}}}\right] \tag{9}$$

By taking the gradient $\nabla_{m_{\mathbf{s}}} f_w(m_{\mathbf{s}})$ and making a Taylor approximation,

$$f_w(m_{\mathbf{s}}) \approx f_w(0) + \nabla_{m_{\mathbf{s}}} f_w(0)^\top m_{\mathbf{s}} \tag{10}$$
$$= c + \frac{\alpha_{\mathbf{w}}(1-\alpha)/(\alpha Z)}{p(w) + (1-\alpha)/(\alpha Z)}\langle w, m_{\mathbf{s}}\rangle \tag{11}$$

For the visual features, we can decompose the likelihood $\mathbb{P}[v|m_{\mathbf{s}}]$ as a product of the likelihoods in each coordinate $\prod_{i=1}^{|v|}\mathbb{P}[v(i)|m_{\mathbf{s}}]$ since we assume a diagonal covariance matrix. Let $v(i) \in \mathbb{R}$ denote the ith visual feature and $W_v^\mu(i) \in \mathbb{R}^{|m_{\mathbf{s}}|}$ be the i-th column of W_v^μ.

$$\mu_v(i) = W_v^\mu(i)m_{\mathbf{s}} + b_v^\mu(i) \tag{12}$$
$$\sigma_v(i) = \exp\left(W_v^\sigma(i)m_{\mathbf{s}} + b_v^\sigma(i)\right) \tag{13}$$
$$v(i)|m_{\mathbf{s}} \sim N(\mu_v(i), \sigma_v^2(i)) \tag{14}$$
$$\mathbb{P}[v(i)|m_{\mathbf{s}}] = \frac{1}{\sqrt{2\pi}\sigma_v(i)}\exp\left(-\frac{(v(i)-\mu_v(i))^2}{2\sigma_v^2(i)}\right) \tag{15}$$

Define $f_{v(i)}(m_{\mathbf{s}})$ as follows:

$$f_{v(i)}(m_{\mathbf{s}}) \tag{16}$$
$$= \log\mathbb{P}[v(i)|m_{\mathbf{s}}]^{\alpha_{\mathbf{v}}} \tag{17}$$
$$= \alpha_{\mathbf{v}}\log\mathbb{P}[v(i)|m_{\mathbf{s}}] \tag{18}$$
$$= -\alpha_{\mathbf{v}}\log\left(\sqrt{2\pi}\sigma_v(i)\right) - \alpha_{\mathbf{v}}\frac{(v(i)-\mu_v(i))^2}{2\sigma_v^2(i)} \tag{19}$$
$$= -\alpha_{\mathbf{v}}\log\left(\sqrt{2\pi}\exp\left(W_v^\sigma(i)m + b_v^\sigma(i)\right)\right) - \alpha_{\mathbf{v}}\frac{(v(i)-W_v^\mu(i)m - b_v^\mu(i))^2}{2\exp\left(W_v^\sigma(i)m_{\mathbf{s}} + b_v^\sigma(i)\right)^2} \tag{20}$$
$$= -\alpha_{\mathbf{v}}\log\sqrt{2\pi} - \left(W_v^\sigma(i)m_{\mathbf{s}} + b_v^\sigma(i)\right) - \alpha_{\mathbf{v}}\frac{(v(i)-W_v^\mu(i)m - b_v^\mu(i))^2}{2\exp\left(2W_v^\sigma(i)m_{\mathbf{s}} + 2b_v^\sigma(i)\right)} \tag{21}$$

The gradient $\nabla_{m_\mathbf{s}} f_{v(i)}(m_\mathbf{s})$ is as follows

$$\nabla_{m_\mathbf{s}} f_{v(i)}(m_\mathbf{s}) \tag{22}$$

$$= -\alpha_\mathbf{v} W_v^\sigma(i) - \alpha_\mathbf{v} \frac{1}{4\sigma_v(i)^4}\left[2(v(i) - \mu_v(i))\right] \tag{23}$$

$$= \alpha_\mathbf{v} \frac{\left[(v(i) - \mu_v(i))W_v^\mu(i) + (v(i) - \mu_v(i))^2 W_v^\sigma(i)\right]}{\sigma_v(i)^2} - \alpha_\mathbf{v} W_v^\sigma(i) \tag{24}$$

$$= \alpha_\mathbf{v} \frac{v(i) - \mu_v(i)}{\sigma_v(i)^2} W_v^\mu(i) + \alpha_\mathbf{v}\left(\frac{(v(i) - \mu_v(i))^2}{\sigma_v(i)^2} - 1\right) W_v^\sigma(i) \tag{25}$$

By Taylor expansion, we have that

$$f_{v(i)}(m_\mathbf{s}) \tag{26}$$

$$\approx f_{v(i)}(0) + \nabla_{m_\mathbf{s}} f_{v(i)}(0)^\top m_\mathbf{s} \tag{27}$$

$$= \underbrace{-\alpha_\mathbf{v} \log\left(\sqrt{2\pi}\exp\left(b_v^\sigma(i)\right)\right) - \alpha_\mathbf{v} \frac{(v(i) - b_v^\mu(i))^2}{2\exp\left(2b_v^\sigma(i)\right)}}_{\text{constant with respect to } m_\mathbf{s}} + \alpha_\mathbf{v} \frac{v(i) - b_v^\mu(i)}{\exp\left(2b_v^\sigma(i)\right)} \langle W_v^\mu(i), m_\mathbf{s}\rangle + \alpha_\mathbf{v}\left(\frac{(v(i) - b_v^\mu(i))^2}{\exp\left(2b_v^\sigma(i)\right)} - 1\right)\langle W_v^\sigma(i), m_\mathbf{s}\rangle \tag{28}$$

$$= c + \alpha_\mathbf{v} \frac{v(i) - b_v^\mu(i)}{\exp\left(2b_v^\sigma(i)\right)} \langle W_v^\mu(i), m_\mathbf{s}\rangle + \alpha_\mathbf{v}\left(\frac{(v(i) - b_v^\mu(i))^2}{\exp\left(2b_v^\sigma(i)\right)} - 1\right)\langle W_v^\sigma(i), m_\mathbf{s}\rangle \tag{29}$$

By our symmetric paramterization of the acoustic features, we have that:

$$f_{a(i)}(m_\mathbf{s}) \tag{30}$$

$$\approx c + \alpha_\mathbf{a} \frac{a(i) - b_a^\mu(i)}{\exp\left(2b_a^\sigma(i)\right)} \langle W_a^\mu(i), m_\mathbf{s}\rangle + \alpha_\mathbf{a}\left(\frac{(a(i) - b_a^\mu(i))^2}{\exp\left(2b_a^\sigma(i)\right)} - 1\right)\langle W_a^\sigma(i), m_\mathbf{s}\rangle \tag{31}$$

Rewriting this in matrix form, we obtain that

$$f_w(m_\mathbf{s}) = c + \psi_w \langle w, m_\mathbf{s}\rangle \tag{32}$$

$$f_v(m_\mathbf{s}) = \sum_{i \in |v|} f_{v(i)}(m_\mathbf{s}) \tag{33}$$

$$= c + \left\langle W_v^{\mu\top}(v - b_v^\mu)\psi_v^{(1)}, m_\mathbf{s}\right\rangle + \left\langle W_v^{\sigma\top}(v - b_v^\mu) \otimes (v - b_v^\mu)\psi_v^{(2)}, m_\mathbf{s}\right\rangle \tag{34}$$

$$f_a(m_\mathbf{s}) = \sum_{i \in |a|} f_{a(i)}(m_\mathbf{s}) \tag{35}$$

$$= c + \left\langle W_a^{\mu\top}(a - b_a^\mu)\psi_a^{(1)}, m_\mathbf{s}\right\rangle + \left\langle W_a^{\sigma\top}(a - b_a^\mu) \otimes (a - b_a^\mu)\psi_a^{(2)}, m_\mathbf{s}\right\rangle \tag{36}$$

where $\otimes$ denotes Hadamard (element-wise) product and the weights ψ's are given as follows:

$$\psi_w = \frac{\alpha_\mathbf{w}(1 - \alpha)/(\alpha Z)}{p(w) + (1 - \alpha)/(\alpha Z)} \tag{37}$$

$$\psi_v^{(1)} = \mathrm{diag}\left(\frac{\alpha_\mathbf{v}}{\exp\left(2b_v^\sigma\right)}\right) \tag{38}$$

$$\psi_v^{(2)} = \mathrm{diag}\left(\frac{\alpha_\mathbf{v}}{\exp\left(2b_v^\sigma\right)} - \alpha_\mathbf{v}\right) \tag{39}$$

$$\psi_a^{(1)} = \mathrm{diag}\left(\frac{\alpha_\mathbf{a}}{\exp\left(2b_a^\sigma\right)}\right) \tag{40}$$

$$\psi_a^{(2)} = \mathrm{diag}\left(\frac{\alpha_\mathbf{a}}{\exp\left(2b_a^\sigma\right)} - \alpha_\mathbf{a}\right) \tag{41}$$

Observe that $W_v^{\sigma\top}(v - b_v^\mu)$ is a composition of a shift $-b_v^\mu$ and a linear transformation $W_v^{\sigma\top}$ of the visual features into the multimodal embedding space. Note that $\mathbb{E}[v|m_\mathbf{s}] = b_v^\mu$. In other words, this shifts the visual features towards 0 in expectation before transforming them into the multimodal embedding space. Our choice of a Gaussian likelihood for the visual and acoustic features introduces a squared term $W_v^{\sigma\top}(v - b_v^\mu) \otimes (v - b_v^\mu)$ to account for the ℓ_2 distance present in the Gaussian pdf. Secondly, regarding the weights ψ's, note that: 1) the weights for a modality are proportional to the global hyperparameters α assigned to that modality, and 2) the weights ψ_w are inversely proportional to $p(w)$ (rare words carry more weight). The weights ψ_v's and ψ_a's scales each feature dimension inversely by their magnitude.

Finally, we know that our objective function (4) decomposes as

$$\mathcal{L}(m_\mathbf{s}, W, b; \mathbf{s}) = \sum_{w \in \mathbf{w}} f_w(m_\mathbf{s}) + \sum_{v \in \mathbf{v}} f_v(m_\mathbf{s}) + \sum_{a \in \mathbf{a}} f_a(m_\mathbf{s}) \tag{42}$$

We now use the fact that $\max_{x:\|x\|_2^2=1}$ constant + $\langle x, g\rangle = g/\|g\|$. If we assume that $m_{\mathbf{s}}^*$ lies on the unit sphere, the maximum likelihood estimate for $m_{\mathbf{s}}$ is approximately:

$$
\begin{aligned}
& m_{\mathbf{s}}^* \\
&= \sum_{w\in\mathbf{w}} \psi_w w + \sum_{v\in\mathbf{v}} \left(W_v^{\mu\top} \tilde{v}^{(1)} \psi_v^{(1)} + W_v^{\sigma\top} \tilde{v}^{(2)} \psi_v^{(2)} \right) \\
&+ \sum_{a\in\mathbf{a}} \left(W_a^{\mu\top} \tilde{a}^{(1)} \psi_a^{(1)} + W_a^{\sigma\top} \tilde{a}^{(2)} \psi_a^{(2)} \right). \qquad (43)
\end{aligned}
$$

where we have rewritten the shifted (and squared) visual and acoustic terms as

$$\tilde{v}^{(1)} = v - b_v^{\mu} \qquad (44)$$

$$\tilde{v}^{(2)} = (v - b_v^{\mu}) \otimes (v - b_v^{\mu}) \qquad (45)$$

$$\tilde{a}^{(1)} = a - b_a^{\mu} \qquad (46)$$

$$\tilde{a}^{(2)} = (a - b_a^{\mu}) \otimes (a - b_a^{\mu}) \qquad (47)$$

which concludes the proof.

1.2 Multimodal Features

Here we present extra details on feature extraction for the language, visual and acoustic modalities.

Language: We used 300 dimensional GloVe word embeddings trained on 840 billion tokens from the Common Crawl dataset (Pennington et al., 2014). These word embeddings were used to embed a sequence of individual words from video segment transcripts into a sequence of word vectors that represent spoken text.

Visual: The library Facet (iMotions, 2017) is used to extract a set of visual features including facial action units, facial landmarks, head pose, gaze tracking and HOG features (Zhu et al., 2006). These visual features are extracted from the full video segment at 30Hz to form a sequence of facial gesture measures throughout time.

Acoustic: The software COVAREP (Degottex et al., 2014) is used to extract acoustic features including 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced segmenting features (Drugman and Alwan, 2011), glottal source parameters (Childers and Lee, 1991; Drugman et al., 2012; Alku, 1992; Alku et al., 1997, 2002), peak slope parameters and maxima dispersion quotients (Kane and Gobl, 2013). These visual features are extracted from the full audio clip of each segment at 100Hz to form a sequence that represent variations in tone of voice over an audio segment.

1.3 Multimodal Alignment

We perform forced alignment using P2FA (Yuan and Liberman, 2008) to obtain the exact utterance time-stamp of each word. This allows us to align the three modalities together. Since words are considered the basic units of language we use the interval duration of each word utterance as one time-step. We acquire the aligned video and audio features by computing the expectation of their modality feature values over the word utterance time interval (Zadeh et al., 2018).

References

Paavo Alku. 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, 11(2-3):109–118.

Paavo Alku, Tom Bäckström, and Erkki Vilkman. 2002. Normalized amplitude quotient for parametrization of the glottal flow. *the Journal of the Acoustical Society of America*, 112(2):701–710.

Paavo Alku, Helmer Strik, and Erkki Vilkman. 1997. Parabolic spectral parametera new method for quantification of the glottal flow. *Speech Communication*, 22(1):67–79.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*, ICLR.

Donald G Childers and CK Lee. 1991. Vocal quality factors: Analysis, synthesis, and perception. *the Journal of the Acoustical Society of America*, 90(5):2394–2410.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarepa collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964. IEEE.

Thomas Drugman and Abeer Alwan. 2011. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Interspeech*, pages 1973–1976.

Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, and Thierry Dutoit. 2012. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):994–1006.

iMotions. 2017. Facial expression analysis.

John Kane and Christer Gobl. 2013. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1170–1179.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and Shai Avidan. 2006. Fast human detection using a cascade of histograms of oriented gradients. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1491–1498. IEEE.