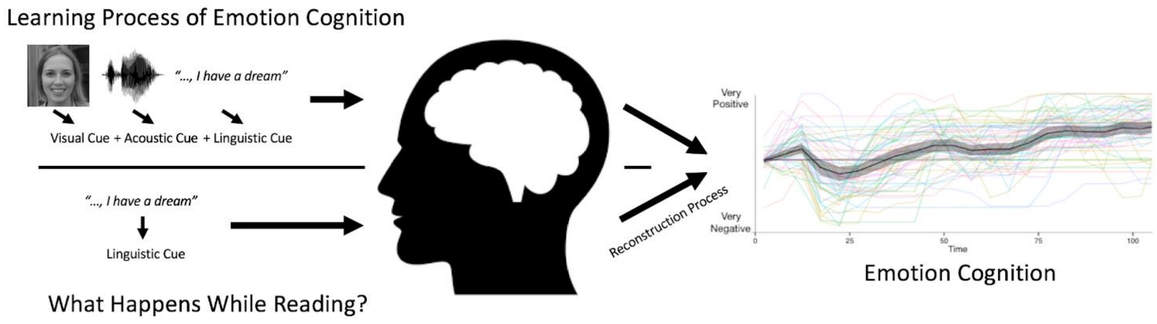
EmoNet: Reconstruction Of Emotion As People Read Using Deep Neural Network With Attention

Zhengxuan Wu, Sherine Zhang, Xuan Zhang

{wuzhengx, sherinez, kayleez}@stanford.edu, Stanford University

Problem & Task



Our **GOAL** is to build a model that makes continuous prediction of emotion valence as people read text. This involves a emotion reconstruction process of the true emotion that is expressed by the story-teller.

Analysis

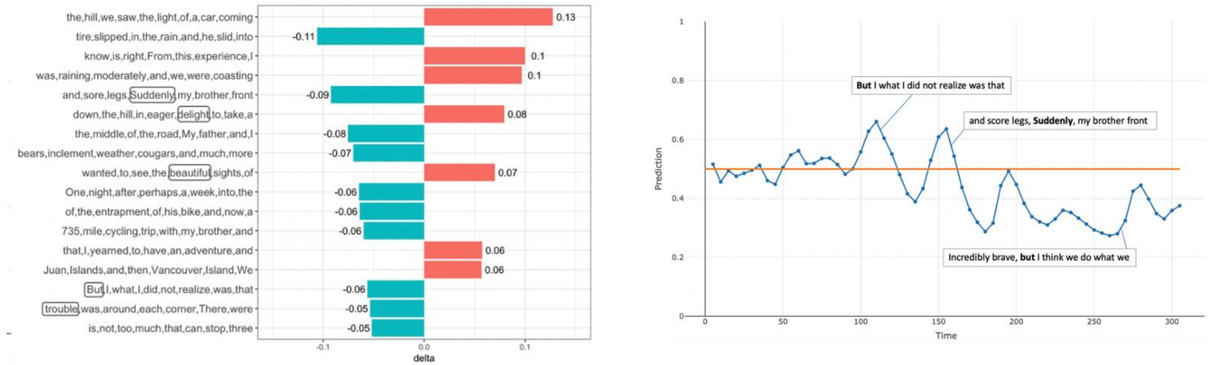


FIGURE 1: We collected a personal emotional story online, and feed into our model. The model is making emotion valence prediction of the story-teller. In our mind, we are indeed reconstructing the whole story as if we were the storyteller.

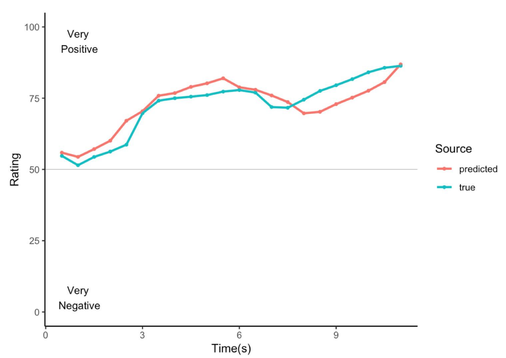
Data

DATASET

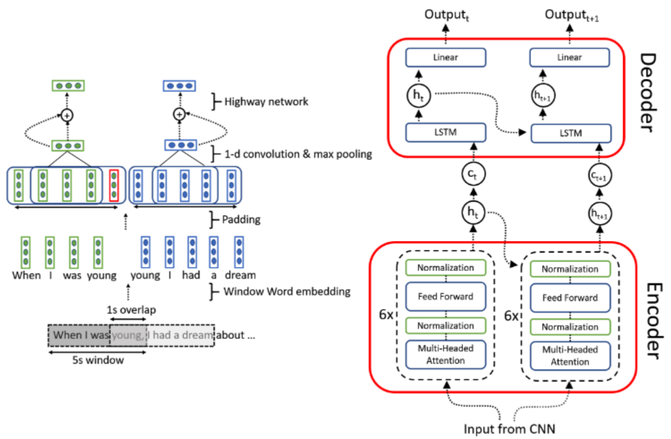
- 193 video clips in total.
- 49 participants describing personal events, e.g.
  - Positive emotions
    - prize-winning moments
    - university graduations
  - Negative emotions
    - a loved one passed away
    - a romantic break-up with significant others
- Divided into Training (60%, 117 videos), Test (20%, 38 videos) and Evaluation (20%, 38 videos) sets.

RATINGS

- True ratings are average values of human evaluations (approximately 20 evals per video).



Approach



MODELS

- CNN Input Embedding Layers To Embed Time Windows
- Applied Local Attention Scores for the Hidden States Output of LSTM encoders.
- LSTM Variants:
  - LSTM with Linear Decoder
  - LSTM with LSTM Decoder
  - LSTM with Auto-regression Decoder
  - Transformer with LSTM Decoder
- Customized Loss Function
  - Loss function MSE
  - Loss function CCC

Results

Window Size & Overlap Size	Statistics	LSTM	ED-LSTM	AR-LSTM	Transformer
		Evaluation Set			
Window = 5s, Overlap = 0s	MSE Loss	0.048	0.031	0.025	**0.014**
	Corr	0.365	0.533	0.442	**0.612**
	CCC	0.224	0.351	0.325	**0.44**
	Best CCC	0.801	0.92	0.915	**0.983**
Window = 5s, Overlap = 2.5s	MSE Loss	0.05	0.03	0.027	**0.014**
	Corr	0.346	0.53	0.443	**0.548**
	CCC	0.217	0.355	0.334	**0.434**
	Best CCC	0.8	0.929	0.921	**0.978**
Window = 10s, Overlap = 5s	MSE Loss	0.063	0.029	0.058	**0.015**
	Corr	0.367	0.518	0.398	**0.557**
	CCC	0.218	0.337	0.243	**0.397**
	Best CCC	0.773	0.964	0.769	**0.975**
Window = 5s, Overlap = 0s, with CCC Loss	CCC	-	-	-	**0.493**
*** Human Evaluated Baseline CCC with visual and audio features = **0.47**					
		Test Set			
Loss Function	**Statistics**	**LSTM**	**ED-LSTM**	**AR-LSTM**	**Transformer**
CCC Loss	CCC Mean	0.291	0.095	0.296	**0.449**
	CCC Std	0.272	0.329	0.31	**0.352**
MSE Loss	CCC Mean	0.149	0.251	0.177	**0.469**
	CCC Std	0.241	0.299	0.269	**0.301**

TABLE 1: This shows the comparison of performances of different models using our tuned parameters. The performance of Transformer is the best.

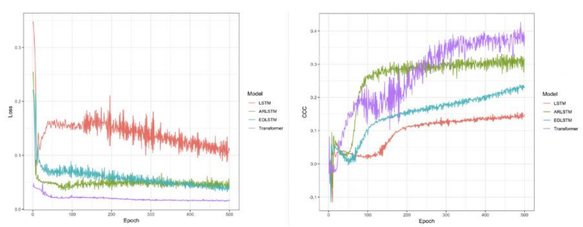


FIGURE 2: This shows the training loss and CCC score as a function of the epoch number.

Evaluation

$$\begin{aligned}\mathrm{CCC}_{XY} &\equiv 1 - \frac{E\left[(X-Y)^2\right]}{E\left[(X-Y)^2\right]|_{\text{setting } \rho_{XY}=0}} \\ &= 1 - \frac{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2 - 2\rho_{XY}\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2} \\ &= \frac{2\rho_{XY}\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2}\end{aligned}$$

Objective: We want to maximize the CCC score which represents the agreement level between the true rating curve and our predicted rating curve.

Conclusion

- **Modeling** the emotion cognition process with linguistic inputs remains a difficult yet important build block of understanding the emotion cognition process of human.
- We have outlined ways to construct valid and effective models to make continuous emotion valence predictions over large text corpus.
- We hope that this paper will inspire more researchers to accomplish more ambitious and quantitative results in the future.

References

- Tan, X., Ong, D. et al. A Multimodal LSTM for Predicting Listener Empathic Responses Over Time
- Code available at https://github.com/sherinezzzzzzz/224NProject