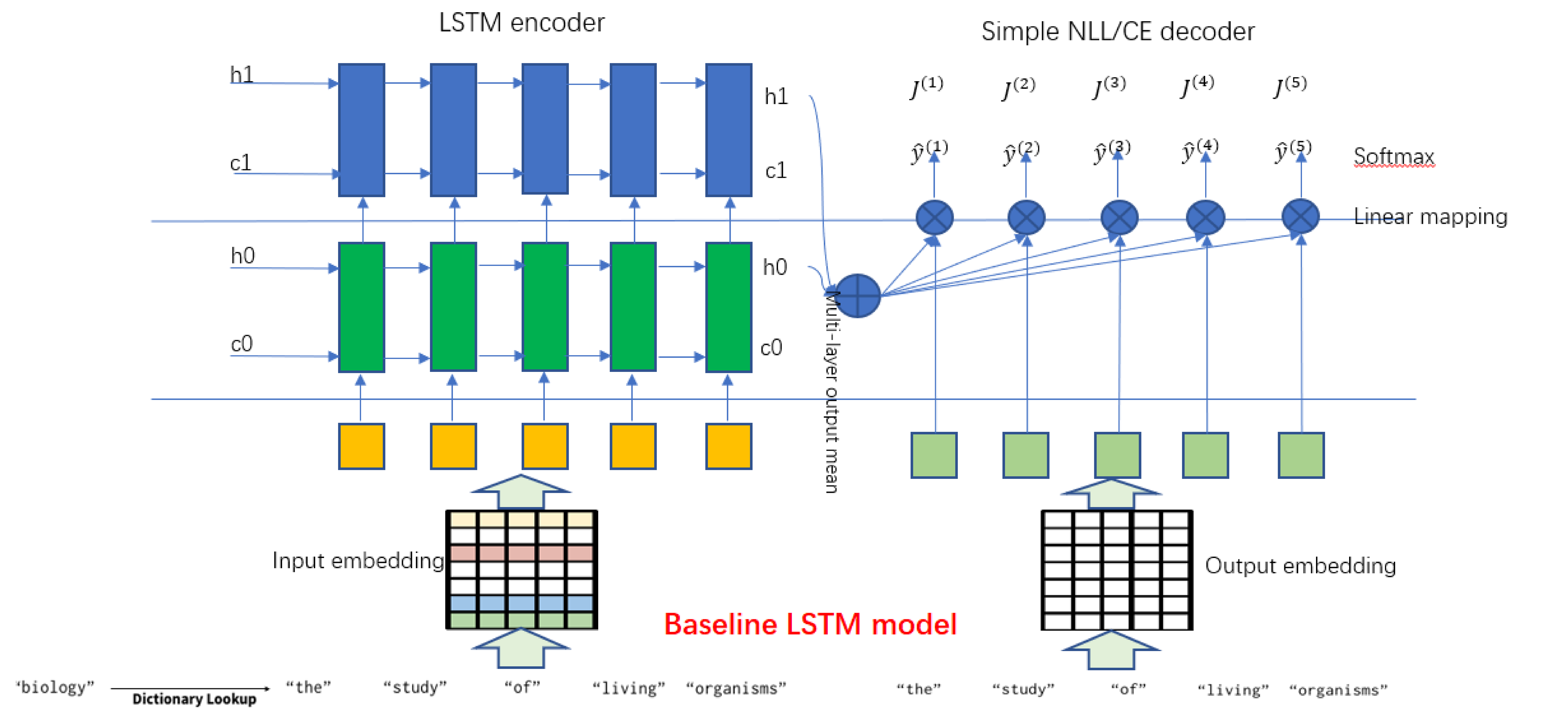
HYBRID DISTRIBUTIONAL AND DEFINITIONAL WORD VECTORS

Haiyuan Mei and Ranjani Iyer
Stanford University, Department of Computer Science

OVERVIEW

- **Motivation**: Out Of Vocabulary (OOV) problem - exploration of word definitions in downstream NLP tasks.
- **Prior methods:** Def2Vec, on-the-fly embeddings capable of capturing OOV words, and limited usage exploration. [1]
- **Approach: HybridVec** Generate word embedding from word definitions, combine it with distributed representations, and explore the possibility or improving downstream NLP tasks.
- **Evaluation:** Intrinsic word embedding benchmarks and Extrinsic NMT evaluation, shown to improve translation perplexities and capture complementary aspect of word regarding distributed representation.



MODEL

- **Baseline LSTM:** A two-layer LSTM encoder, Simple linear decoder and NLL loss, where the encoder layer hidden output denotes the final definitional word vector.
- **Seq2Seq:** A two-layer LSTM encoder with dropouts plus a two layer LSTM decoder without attention.
- **Variational AutoEncoder:** Adapted VAE with single-layer LSTM encoder and decoder with Gaussian prior regularizer[3].

$$\begin{aligned}\mathcal{L}(\theta;x) = -\mathrm{KL}(q_\theta(\vec{z}|x)||p(\vec{z})) \\ + \mathbb{E}_{q_\theta(\vec{z}|x)}[\log p_\theta(x|\vec{z})] \\ \leq \log p(x) \quad .\end{aligned}$$

TRAINING

- **Dataset: GloVe** [4]. All models are trained on pretrained 300d GloVe vectors based on a crawl of 2014 Wikipedia.
  Definitions retrieved from **WordNet**[5].
- **HybridVec Implementation:** Pytorch, Adam optimizer, Xavier initialization, hidden size 150, learning rate of 1e-4, batch size 64, 15/20 epochs.
- **Intrinsic evaluation:** Word embedding benchmarks [2]
- **NMT Dataset:** OpenNMT-py demo(10k) dataset. Only for comparasion between GloVe and HybridVec.

RESULTS

- Word embeddings benchmarks for GloVe, LSTM Baseline and Seq2Seq model. LSTM baseline model is roughly at the level of distributional method; Seq2Seq model shows very limited evidence of such capability[2]:

	BLESS	ESSLI_1a	MEN	MTurk	RG65	SL999	WS353
Glove	0.82	0.75	0.737465	0.633182	0.769525	0.3705004	0.543326
Baseline glove	0.55	0.659091	0.51071	0.4226407	0.656402	0.3678366	0.449105
Baseline rand	0.52	0.613636	0.447908	0.3181051	0.6444908	0.3288122	0.35609
S2S enc mean	0.275	0.522727	0.106169	0.1370724	0.0890822	-0.018433	0.051959

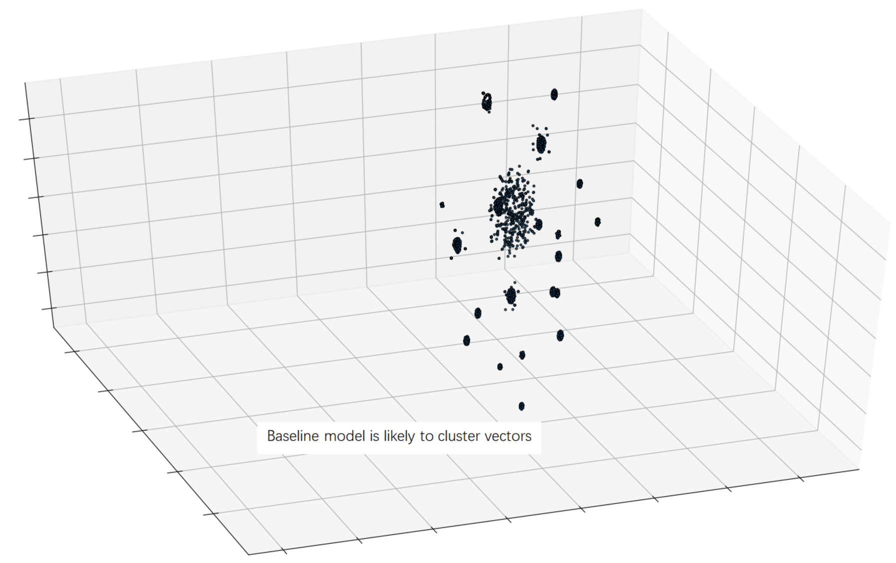
- GloVe: WEB benchmark for GloVe vectors
- Baseline glove: WEB benchmark for LSTM baseline model initialized from GloVe
- Baseline rand: WEB benchmark for LSTM baseline model initialized randomly
- s2s enc mean: WEB benchmark for Seq2seq model with encoder output mean as the def vec.

- OpenNMT compare performance improvements using LSTM baseline vector and GloVe:

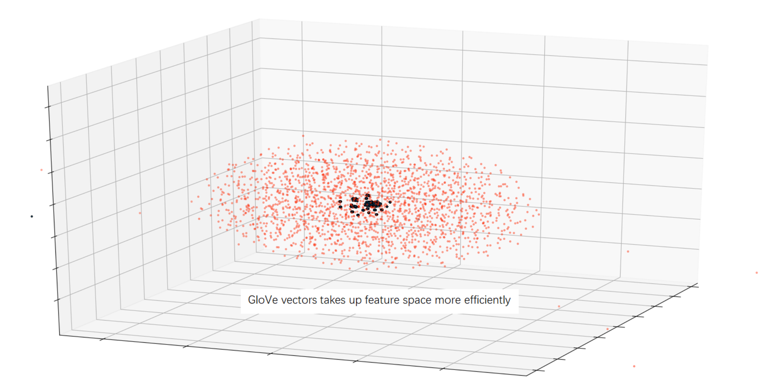
	No pretrained	LSTM Baseline	GloVe	description
Train PPL	7.47	6.69	4.84	10k nmp demo sentence trained 1 epoch, with/o pretrained word vectors. Glove has most positive impact, LSTM baseline also exhibites positive impact
Train ACC	56.29	58.4	64.32	
BLEU	0.0093	0.0137	0.0199	10K nmt training demo 10 epochs, eval on 3k nmt val sentences,similar result as above perplexity and accuracy

ANALYSIS

- LSTM baseline vectors tend to cluster in feature space. Need to train from a broader source.
- Glove makes use of feature space more efficiently, grasp more sutle meaning of words.



3D tSNE: LSTM Baseline vectors is likely to cluster



GloVe uses feature space more efficiently

FUTURE WORK

- Additional plans for model: greater regularization, inputting multiple definitions, inputting sentence structure, try other embeddings.
- Continue the exploration of combining different word vectors for downstream NLP tasks.

REFERENCES

[1] Andrey Kurenkov and Duan. Duan. Def2vec: Learningword vectors from definitions. Stanford, CA, 2016.
[2] Stanislaw Jastrzebski. Word embeddings benchmark. https://github.com/kudkudak/word-embeddings-benchmarks, 2015.
[3] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *SIGNLL Conference on Computational Natural Language Learning*, 2016.
[4] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
[5] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.