GATED MULTIMODAL FUSION WITH CONTRASTIVE LEARNING FOR TURN-TAKING PREDICTION IN HUMAN-ROBOT DIALOGUE

Jiudong Yang[1†], *Peiying Wang*[1†], *Yi Zhu*[1,2], *Mingchao Feng*[1], *Meng Chen*[1], *Xiaodong He*[1]

[1]JD AI, Beijing, China, [2]LTL, University of Cambridge
cdyangjiudong3@jd.com, wangpeiying3@jd.com, yz568@cam.ac.uk
fengmingchao@jd.com, chenmeng20@jd.com, xiaodong.he@jd.com

ABSTRACT

Turn-taking, aiming to decide when the next speaker can start talking, is an essential component in building human-robot spoken dialogue systems. Previous studies indicate that multimodal cues can facilitate this challenging task. However, due to the paucity of public multimodal datasets, current methods are mostly limited to either utilizing unimodal features or simplistic multimodal ensemble models. Besides, the inherent class imbalance in real scenario, e.g. sentence ending with short pause will be mostly regarded as the end of turn, also poses great challenge to the turn-taking decision. In this paper, we first collect a large-scale annotated corpus for turn-taking with over 5,000 real human-robot dialogues in speech and text modalities. Then, a novel gated multimodal fusion mechanism is devised to utilize various information seamlessly for turn-taking prediction. More importantly, to tackle the data imbalance issue, we design a simple yet effective data augmentation method to construct negative instances without supervision and apply contrastive learning to obtain better feature representations. Extensive experiments are conducted and the results demonstrate the superiority and competitiveness of our model over several state-of-the-art baselines.

***Index Terms*—** Multimodal Fusion, Turn-taking, Barge-in, Endpointing, Spoken Dialogue System

1. INTRODUCTION

For spoken dialog systems, turn-taking is an essential component which allows participants in a dialogue to exchange control of the floor [1]. Given an utterance in a conversation, a **hold** means that the next utterance will be continued by the same speaker while a **switch** indicates that the next utterance will be uttered by the other speaker. For human-robot conversations occurred on the telephone with Interactive Voice Response (IVR) systems, turn-taking plays a critical role for user in providing natural interaction experience.

Most of previous works in turn-taking focus on the user end-of-turn detection, i.e. **endpointing**. It assumes that turn

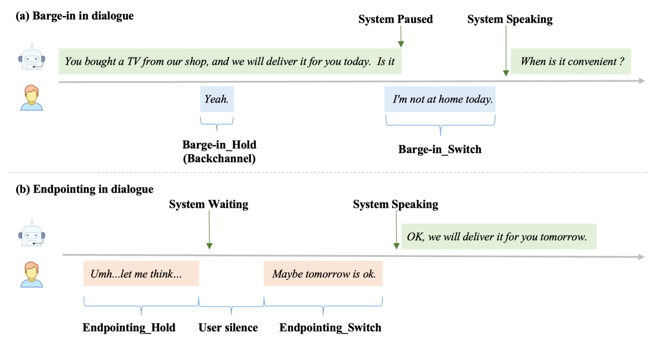

Fig. 1. Example of turn-taking in a conversation.

switch occurs when a speaker has stopped speaking and a period of silence comes out. Traditionally, a naive approach for endpointing is that when the current speaker pauses for a heuristically designed threshold [2], the system will take the turn. However, this approach is limited in its naturalness that the fixed threshold can potentially be too short (frequent interruptions) or too long (awkward pauses). To address this problem, machine learning methods have gained popularity since 1970s [3, 4, *inter alia*], and models based on inter-pausal unit (IPU), an audio segment followed by silence longer than 200 milliseconds, have mostly been studied recently because of its simplicity in practice [5, 6]. For a specific IPU, various cues across modalities, such as prosody, semantics, syntax, breathing, gesture, and eye-gaze can be extracted and integrated to determine whether this turn is yielded or not [7, 8].

Although remarkable progress has been made, some issues are still present in turn-taking research. (1) There is a dearth of public multimodal dataset for turn-taking from real scenario: previous works mostly experiment on private in-house datasets [9], pure text corpus transcribed from dialogues [5, 10], and constructed dataset with Wizard-of-Oz setup [11, 12] which is difficult to extract fine-grained speech information such as timing. Moreover, most of them ignore handling user interruptions (**barge-in**), where switch occurs when a speaker starts uttering before the other speaker finishes speaking [13]. Barge-in detection is crucial when the system asks longer questions or gives longer instructions which the

† Equal contribution.

user might have heard before or can be predicted from context. Figure 1 shows a dialogue example of both endpointing and barge-in. (2) Current multimodal approaches mainly use recurrent neural network (RNN) [5, 14] to deal with the feature sequences, whereas more advanced and efficient neural models such as Transformer [15] are not fully explored. Besides, when combining the features from different modalities, only simple ensemble techniques [16] are utilized, which can not optimize all feature extractors jointly.

Motivated by above limitations, in this paper, we first collect a large-scale human-robot dialogue corpus from online conversation IVR systemin real scenario (§2). The dataset covers both endpointing and barge-in situations, and contains more than 5,000 dialogues. Then we propose a novel **G**ated **M**ultimodal **F**usion model (denoted as **GMF**) for turn-taking prediction based on IPU in spoken dialogue system. GMF contains extendable feature extractors to obtain features from speech and text modalities (§3). Specifically, the prevalent Transformer [15] and ResNet [17] blocks are employed for processing text and speech respectively, and finer-grained timing features from dialogue are also considered. Additionally, to alleviate the issue of class imbalance stemming from the characteristics of turn-taking dataset, we perform data augmentations by constructing samples for the minority class with self-supervised methods combined with constrastive learning. Extensive experiments were conducted to compare with several state-of-the-art baselines, and the results demonstrate the effectiveness of our proposed model.

2. DATASET

Our dataset is collected from a commercial conversational IVR system, where conversations take place between customer and intelligent robot over the phone. During the call, the robot tries to make an appointment with customer for the delivery time and address of purchased goods. Each dialogue session lasts about 1-2 minutes with around 5-10 turns, and all turns mentioning the name of customer are removed for anonymization. We manually transcribe all speech into text, hence both speech and text information are available.

We extract IPUs of *customer speech* from corresponding channel of IVR system. Then we group the extracted IPUs into two *disjoint* subsets of endpointing and barge-in with the following heuristics: IPU which does not overlap with any robot speech is identified as endpointing, whereas for barge-in the customer interrupts while robot is speaking, i.e. customer speech starts later and overlaps with robot speech. For both subsets, two graduate students majoring in linguistics are instructed to annotate whether the system should **switch** or **hold** for each IPU given the whole dialogue for more accurate decision. For endpointing, *switch* means that the customer has finished his/her current speech, and the robot should take the turn, whereas *hold* means that the customer has not finished and wants to continue speaking. For barge-in, *switch* represents that the customer interrupted the robot by saying something meaningful and wants the robot to stop talking, while *hold* means that the voice from customer might be background noise or backchannels (phatic response without significant information like *yeah* and *uh-huh*), and the robot should ignore it and keep speaking. See Figure 1 for annotated turn-taking labels in each case. The Fleiss kappa score of the annotation is 0.827, indicating substantial inter-annotator agreement.

Endpointing		**Barge-in**	
Switch	Hold	Switch	Hold
2451	844	1942	6312
74.4%	25.6%	23.5%	76.5%

Table 1. Statistics of our dataset.

The final dataset consists of 5,380 dialogues in total. Table 2 shows the dataset statistics in both cases. In our scenario, as the robot starts the conversation proactively and the customer usually gives short answers (e.g., confirmation), we can see that there are more *switch* instances in endpointing compared to barge-in, where most *tentative interruptions* are false barge-in coming from noise or backchannel, both of which are very common in dyadic conversations via telephone. These observations show the complexity of our dataset, and we will mitigate the class imbalance issue via data augmentation with contrastive learning in §3.

3. APPROACH

3.1. Gated Multimodal Fusion Model

Previous studies [7, 8] have shown that turn-taking cues across different modalities can be complementary. The combination of several cues can lead to more accurate predictions of the speaker's intentions. Inspired by this, we propose a novel model (denoted as **GMF**) to fuse various multimodal features, which is illustrated in Figure 2. Three different encoders are devised to encode text, speech, and categorical or continuous features correspondingly, which intend to catch the semantic, acoustic and timing features respectively. Then a gated multimodal fusion block is devised to fuse the above representations seamlessly. Finally, the output of the fusion layer is fed into the *sigmoid* function for prediction.

Semantic features. Intuitively, the verbal aspect of spoken language, such as the words spoken and the semantic and pragmatic information that can be derived from those, should be very important for indicating turn shifts [18, 16]. The completion of a syntactic unit is a basic requirement for considering the turn as "finished". Considering the powerful ability of the Transformer block [15] in text representation learning, we apply it to encode both the context and current utterances:

$$\mathbf{r}^s = Transformer_{Encoder}(\mathbf{e}) \tag{1}$$

where **e** is the input embedding: the sum of token, position and segment embeddings. $\mathbf{r}^s$ is the text representation.

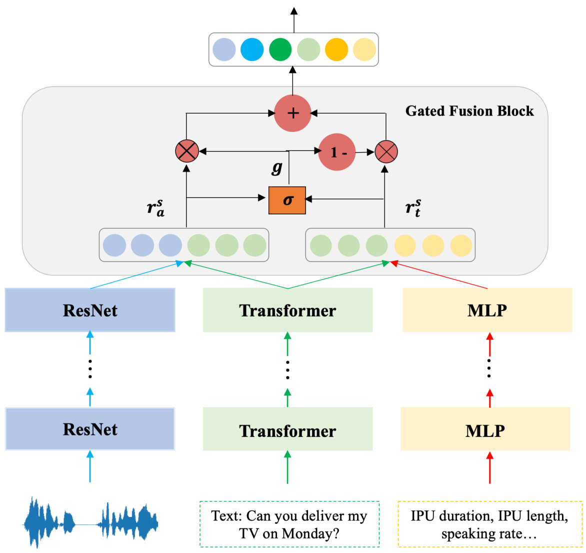

Fig. 2. Architecture of our proposed model GMF.

Acoustic features. The role of acoustic features has been the subject of much interest in turn-taking prediction [19]. In this work, the acoustic features consist of prosodic features (e.g., *energy*, *pitch*), and speech features (e.g., *zero-crossing-rate*, *filterbank*). OpenSmile toolkit [20] is used to extract above features. Following previous work [14], we extract the features in the last 2 seconds of each IPU segment with a frame shift size of 50 milliseconds. For each frame, we concatenate the above mentioned features to a single vector (30 dimensions), which will be viewed as frame representation. Inspired by the successful use of ResNet architecture [17] in audio tasks [21], we feed the sequential frame representations $\mathbf{f}$ into 18 ResNet layers to obtain the acoustic representation $\mathbf{r}^a$ as follows.

$$\mathbf{r}^a = ResNet_{Encoder}(\mathbf{f}) \tag{2}$$

Timing features. Based on the analysis in Section 2 and previous research [22, 16], timing features can also be good indicators for turn-taking prediction. Here, we extract following four timing features, including *time duration of IPU*, *text length of IPU*, *time interval with last turn*, and *speaking rate*. All features are discretized and randomly initialized with dense vectors, then several Multilayer Perceptron (MLP) layers are applied to map timing features $\mathbf{t}$ into timing representation $\mathbf{r}^t$.

$$\mathbf{r}^t = MLP_{Encoder}(\mathbf{t}) \tag{3}$$

Gated multimodal fusion. After we obtain three representations from different modalities, we need to fuse them into the final representation for class prediction. Inspired by the flow control in recurrent architectures like GRU or LSTM, we devise a novel gated multimodal fusion block to control the contribution of different modalities. Considering semantic features play a vital role in turn-taking prediction, we first fuse $\mathbf{r}^s$ with $\mathbf{r}^a$ and $\mathbf{r}^t$ independently using fully-connected layers, resulting in $\mathbf{r}_a^s = FC(\mathbf{r}^s, \mathbf{r}^a)$ and $\mathbf{r}_t^s = FC(\mathbf{r}^s, \mathbf{r}^t)$. Then we combine them further as follows:

$$g = \sigma(\mathbf{W}_g \cdot [\mathbf{r}_a^s, \mathbf{r}_t^s]) \tag{4}$$

$$\mathbf{r} = g \cdot \mathbf{r}_a^s + (1-g) \cdot \mathbf{r}_t^s \tag{5}$$

$$\hat{\mathbf{y}} = \sigma(\mathbf{W}_f \mathbf{r} + b) \tag{6}$$

where g is the gating vector, $\sigma(\cdot)$ is the *sigmoid* function, and $\hat{\mathbf{y}}$ is the predicted label. $\mathbf{W}_g$ and $\mathbf{W}_f$ are weight matrices. The model is optimized by minimizing cross entropy loss $\mathcal{L}_{ce}$:

$$\mathcal{L}_{ce} = -y \cdot log(\hat{y}) + (1-y) \cdot log(1-\hat{y}) \tag{7}$$

3.2. Data Augmentation via Contrastive Learning

One inevitable obstacle in turn-taking prediction is the class imbalance issue. As observed and analyzed in Section 2, both endpointing and barge-in suffer from the imbalanced class distribution, which would damage the performance of classifier. To alleviate this issue, we perform data augmentations by constructing samples for the minority class with self-supervised methods and leveraging constrastive learning.

Recently, self-supervised contrastive learning (CL) has made remarkable progress in various fields [23, 24]. The basic idea is to pull together an anchor and a **positive** sample in embedding space, and to push apart the anchor from many **negative** samples. Positive examples are often obtained from data augmentations of the anchor (a.k.a views), and negative examples randomly chosen from the minibatch during training.

Inspired by [24], we take dropout [25] as minimal data augmentation to generate the positive pair for each sample. We randomly drop elements in the fused representation by a specific probability and set their values to zero. Besides, we also perform data augmentations for the minority class in our dataset. Specifically, for endpointing, we construct hundreds of turn-holding samples by corrupting the turn-switching samples into incomplete ones. We truncate the complete utterance (with more than 10 characters) by removing the last 30% of words for both the speech and text. Then the utterance becomes semantically incomplete and the label is assigned as *hold*. For barge-in, we first collect normal question and answer utterance pairs (i.e. system asks question and user answers the question) from dialogues. Then we move the answer utterance ahead and make it overlap with the system's question utterance in the time axis. By this way, we obtain hundreds of turn-switching samples for the barge-in scenario. Finally, we apply the contrastive loss as follows:

$$\mathcal{L}_{cl} = -log \frac{e^{sim(\mathbf{x},\mathbf{x}^+)/\tau}}{\sum_{i=1}^{N} e^{sim(\mathbf{x},\mathbf{x}^-)/\tau}} \tag{8}$$

where τ is a temperature hyperparameter and $sim(\cdot)$ is the cosine similarity function. As mentioned above, the positive pair is obtained by feeding the fused representation of each sample into dropout twice. The negative samples are the examples

Model	Endpointing		Barge-in	
	Acc	Macro-F1	Acc	Macro-F1
Random	0.490	0.467	0.512	0.465
MajVot$_{cls}$	0.744	0.425	0.765	0.432
LSTM$_{ens}$ [14]	0.752	0.646	0.789	0.642
MoE [16]	0.778	0.643	0.835	0.734
GMF	0.819	0.736	0.869	0.814
GMF w/ CL	**0.829**	**0.761**	**0.873**	**0.826**
w/o semantic	0.767	0.658	0.838	0.740
w/o context	0.783	0.699	0.852	0.786
w/o acoustic	0.788	0.708	0.820	0.732
w/o timing	0.791	0.707	0.853	0.791

Table 2. Turn-taking performance of different models.

Method	Endpointing		Barge-in	
	Acc	Macro-F1	Acc	Macro-F1
Concatenation	0.812	0.723	0.867	0.801
Summation	0.809	0.724	0.864	0.799
Multiplication	0.806	0.724	0.865	0.801
MFB [26]	0.813	0.728	0.861	0.798
GMF	**0.819**	**0.736**	**0.869**	**0.814**

Table 3. Turn-taking performance of different fusion methods.

from different classes in the same minibatch, including our augmented samples with self-supervised methods. To sum up, the total objective of our model is to minimize the following integrated loss:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{cl} \tag{9}$$

4. EXPERIMENT

Baselines. We compare **GMF** with the following baselines: (1) **Random**: The class is predicted randomly. (2) **MajVot**$_{cls}$: The class is predicted by majority voting based on class distribution of the training set. (3) **LSTM**$_{ens}$ [14]: It utilizes prosodic features, speech features, and linguistic features as input feature set, then three individual LSTMs are trained to catch the corresponding features, and finally a linear layer is applied to ensemble the three outputs of LSTMs. (4) **MoE** [16]: Mixture of experts that linearly interpolates four separate classifiers with SVM based on prosodic, timing, lexical & syntactic, and semantic features.

Experimental Setup. We conduct 10-fold cross validation using our dataset and report the average results. The 300-dimension Glove word embeddings are used to initialize the embedding layer of Transformer and LSTM. The number of Transformer, ResNet, and MLP layers are 3, 18, 3 respectively. The CNN kernel is set to 3 * 3 with stride of 1 in the frequency axis. The dimensions of Transformer, ResNet, and MLP are all set to 128. For all baselines, the hyper-parameters are kept consistent with the original paper. The classification *accuracy* and *Macro-F1* are used as evaluation metrics.

Main Results. Table 2 shows the results on **endpointing** and **barge-in** datasets conducted separately. It's observed that, our proposed model outperforms all baselines on both datasets significantly (Sign Test, with p-value<0.05). Especially, **GMF** outperforms state-of-the-art approach **MoE** by absolute **9.3%** and **8%** on Macro-F1 score. Considering the input features are basically the same for **LSTM**$_{ens}$, **MoE** and **GMF**, it indicates that **GMF** can extract more distinguished features and fuse them more effectively. Besides, compared with **LSTM**$_{ens}$, as both Transformer and ResNet can be easily parallelized during training, **GMF** is also more efficient. As to the class imbalance issue, after we apply the data augmentation with contrastive learning (**GMF w/ CL**), it's observed that further gains up to 2.5% Macro-F1 scores can be obtained, which demonstrates the effectiveness of contrastive learning.

Ablation Study. To investigate the contribution of different components, we also conduct ablation study by removing each modality of features from GMF separately. The second part of Table 2 shows that the performance degrades correspondingly, which proves that different multimodal features are complementary to each other. We also try to remove the dialogue context from the semantic feature and use the current utterance instead, it's observed that the performance is also damaged, which illustrates the necessity of dialogue context.

More Multimodal Fusion Approaches. Besides the gated multimodal fusion, we also verify more multimodal fusion methods, including simple fusion methods such as concatenation ($\mathbf{r} = [\mathbf{r}^a; \mathbf{r}^s; \mathbf{r}^t]$), summation ($\mathbf{r} = \mathbf{r}^a + \mathbf{r}^s + \mathbf{r}^t$), multiplication ($\mathbf{r} = \mathbf{r}^a \circ \mathbf{r}^s \circ \mathbf{r}^t$) and more advanced information fusion approach i.e. multimodal factorized bilinear pooling (MFB) [26], which is widely used in visual question answering (VQA) task. Table 4 demonstrates that the **GMF** outperforms other fusion techniques (Sign Test, with p-value<0.05), which indicates the advantage of gated multimodal fusion.

5. CONCLUSION

In this paper, we focus on fusing multimodal information seamlessly to facilitate turn-taking prediction. A novel gated multimodal fusion model equipped with constrastive learning is proposed and applied on both endpointing and barge-in situations. Extensive experiments demonstrate the superiority of our model against several strong baselines. We also contribute a large-scale human-robot dialogue corpus. In the future, we will focus on exploring more turn-taking phenomena, such as backchannel and filler words. Furthermore, we will also explore more modal features (e.g., eye-gaze and gestures) to enhance our model.

6. ACKNOWLEDGEMENT

This work is supported by the National Key R&D Program of China under Grant No.2018YFB2100802.

7. REFERENCES

[1] Angelo Cafaro, Nadine Glas, and Catherine Pelachaud, "The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions," in *AAMAS*, 2016.

[2] Gabriel Skantze, "Turn-taking in conversational systems and human-robot interaction: A review," *Computer Speech & Language*, 2020.

[3] Starkey Duncan, "Some signals and rules for taking speaking turns in conversations.," *Journal of personality and social psychology*, 1972.

[4] Hatim Khouzaimi, Romain Laroche, and Fabrice Lefevre, "Optimising turn-taking strategies with reinforcement learning," in *SIGDIAL*, 2015, pp. 315–324.

[5] Matthew Roddy, Gabriel Skantze, and Naomi Harte, "Investigating speech features for continuous turn-taking prediction using lstms," *arXiv*, 2018.

[6] Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro, "A neural turn-taking model without rnn.," in *INTERSPEECH*, 2019, pp. 4150–4154.

[7] Ryo Masumura, Taichi Asami, Hirokazu Masataki, Ryo Ishii, and Ryuichiro Higashinaka, "Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks.," in *INTERSPEECH*, 2017, pp. 1661–1665.

[8] Katharina J Rohlfing, Giuseppe Leonardi, Iris Nomikou, Joanna Rączaszek-Leonardi, and Eyke Hüllermeier, "Multimodal turn-taking: Motivations, methodological challenges, and novel approaches," *TCDS*, 2019.

[9] Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara, "Prediction of turn-taking using multitask learning with prediction of backchannels and fillers," in *Interspeech*, 2018.

[10] Erik Ekstedt and Gabriel Skantze, "Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog," *arXiv*, 2020.

[11] Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost, "Improving end-of-turn detection in spoken dialogues by detecting speaker intentions as a secondary task," in *ICASSP*, 2018.

[12] Andrei C. Coman, Koichiro Yoshino, Yukitoshi Murase, Satoshi Nakamura, and Giuseppe Riccardi, "An incremental turn-taking model for task-oriented dialog systems," in *INTERSPEECH*, 2019, pp. 4155–4159.

[13] Gabriel Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks," in *SIGDIAL*, 2017, pp. 220–230.

[14] Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara, "Turn-taking prediction based on detection of transition relevance place.," in *INTERSPEECH*, 2019.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[16] Seyedeh Zahra Razavi, Benjamin Kane, and Lenhart K Schubert, "Investigating linguistic and semantic features for turn-taking prediction in open-domain human-computer conversation.," in *INTERSPEECH*, 2019.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[18] Antoine Raux and Maxine Eskenazi, "Optimizing endpointing thresholds using dialogue features in a spoken dialogue system," in *SIGDIAL*, 2008, pp. 1–10.

[19] Sara Bögels and Francisco Torreira, "Listeners use intonational phrase boundaries to project turn ends in spoken interaction," *Journal of Phonetics*, pp. 46–57, 2015.

[20] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *ACMMM*, 2010.

[21] Logan Ford, Hao Tang, François Grondin, and James R Glass, "A deep residual network for large-scale acoustic scene analysis.," in *INTERSPEECH*, 2019.

[22] Agustín Gravano and Julia Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, pp. 601–634, 2011.

[23] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[24] Tianyu Gao, Xingcheng Yao, and Danqi Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv*, 2021.

[25] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, pp. 1929–1958, 2014.

[26] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *ICCV*, 2017.