Regularizing RNNs for Caption Generation by Reconstructing The Past with The Present
Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, Wei Liu
{jschenxinpeng, forest.linma, cswjiang}@gmail.com; jian.yao@whu.edu.cn; wliu@ee.columbia.edu

Introduction
We propose a novel architecture, namely Auto-Reconstructor Network (ARNet), which coupling with the conventional encoder-decoder framework, works in an end-to-end fashion for caption generation.

Architecture
Our proposed ARNet connects two neighbouring hidden states by reconstructing the past hidden state with present one. In this paper, ARNet is realized by another LSTM.

Experiments
Hidden states visualization of the attentive encoder-decoder model (a) and the attentive encoder-decoder-ARNet model (b). The filled circles in blue represent the hidden states generated in the training mode, while the open circles in red are obtained in the inference mode.

Experiments Cont.
We define two different distance metrics based on cosine distance to evaluate the discrepancy between the hidden states from training and inference, respectively.
1. Mean centroid distance: \( d_{mc}(U, V) = \frac{1}{|U|} \sum_{u \in U} \frac{1}{|V|} \sum_{v \in V} \cos(u, v) \)
2. Point-wise distance: \( d_{pw}(U, V) = \frac{1}{|U|} \sum_{u \in U} \left| \frac{1}{|V|} \sum_{v \in V} \cos(u, v) - d(u, v) \right| \)

Experiments Cont.

Motivation
• The latent relationships between neighbouring hidden states in RNNs are not fully exploited.
• The discrepancy problem, also named as exposure bias, in RNN between training and inference for sequence prediction tasks still exists.

Our proposed ARNet connects two neighbouring hidden states by reconstructing the past hidden state with present one. In this paper, ARNet is realized by another LSTM.

Experiments
Hidden states visualization of the attentive encoder-decoder model (a) and the attentive encoder-decoder-ARNet model (b). The filled circles in blue represent the hidden states generated in the training mode, while the open circles in red are obtained in the inference mode.

We define two different distance metrics based on cosine distance to evaluate the discrepancy between the hidden states from training and inference, respectively.
1. Mean centroid distance: \( d_{mc}(U, V) = \frac{1}{|U|} \sum_{u \in U} \frac{1}{|V|} \sum_{v \in V} \cos(u, v) \)
2. Point-wise distance: \( d_{pw}(U, V) = \frac{1}{|U|} \sum_{u \in U} \left| \frac{1}{|V|} \sum_{v \in V} \cos(u, v) - d(u, v) \right| \)