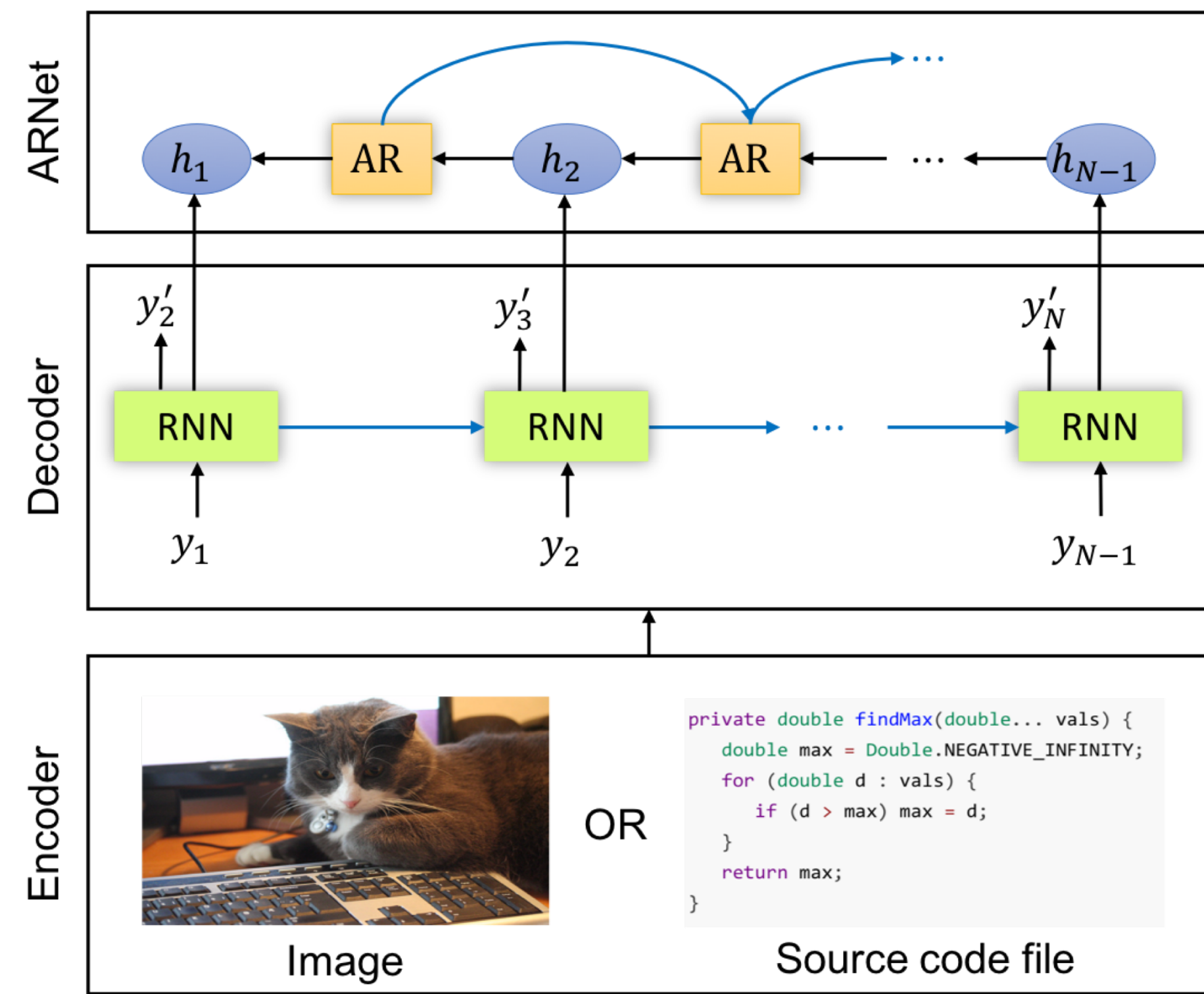


Introduction

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.

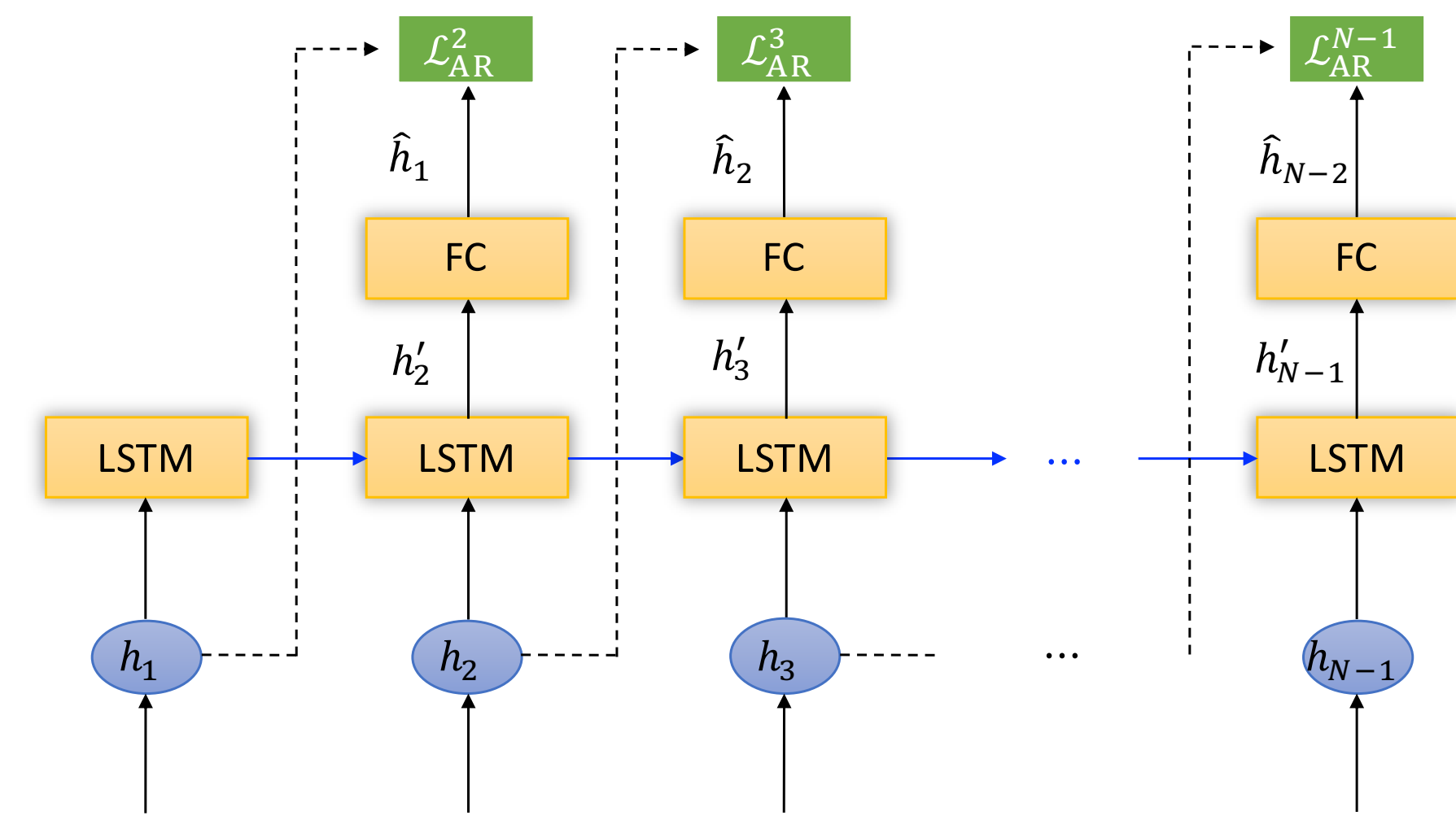


Motivation

- 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.

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.



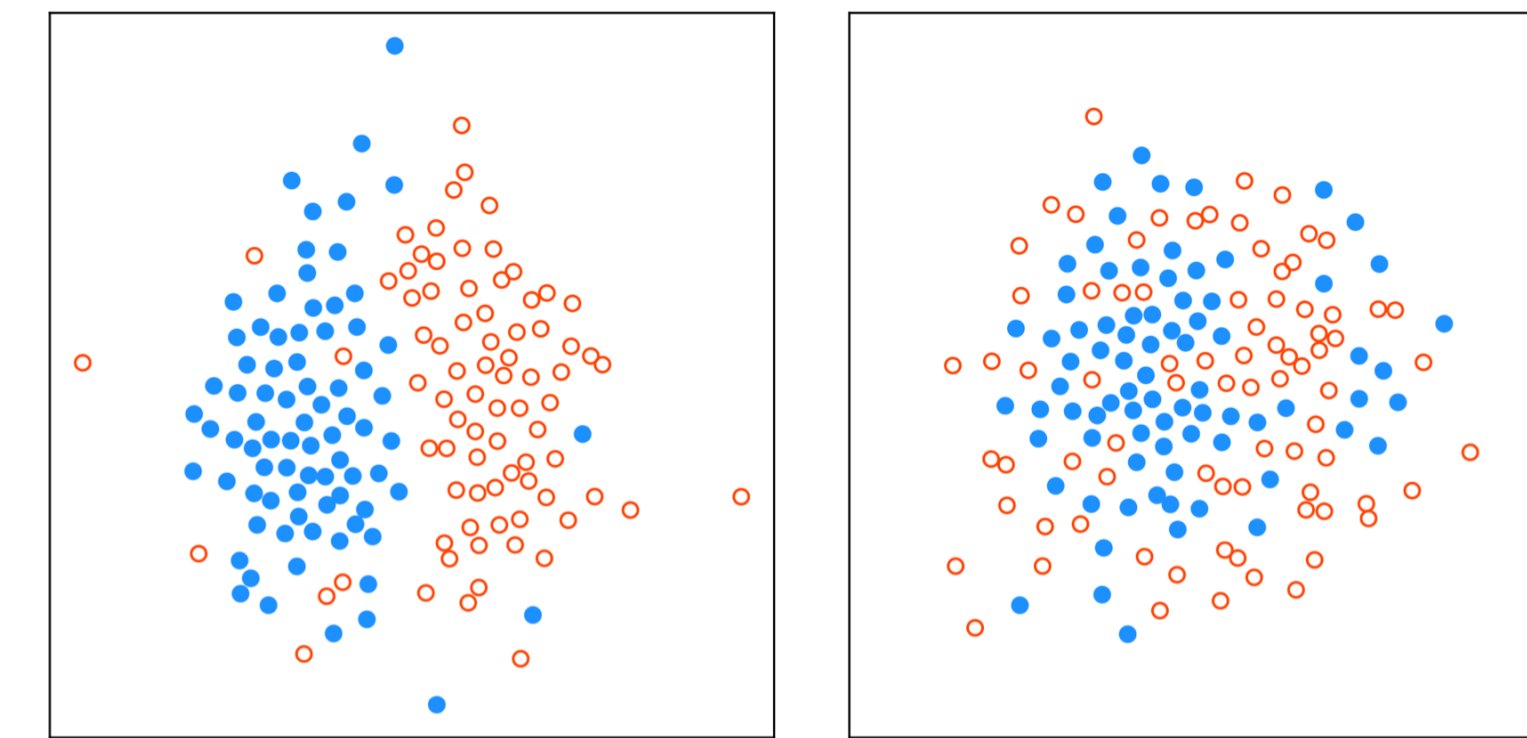
Experiments

- Image Captioning on MSCOCO

Model Name	B-1	B-2	B-3	B-4	M	R	C	S
NIC (Vinyals et al. 2015)	0.666	0.451	0.304	0.203	-	-	-	-
m-RNN (Mao et al. 2015)	0.670	0.490	0.350	0.250	-	-	-	-
Soft Attention (Xu et al. 2015)	0.707	0.492	0.344	0.243	0.239	-	-	-
Hard Attention (Xu et al. 2015)	0.718	0.504	0.357	0.250	0.230	-	-	-
Semantic Attention (You et al. 2016)	0.709	0.537	0.402	0.304	0.243	-	-	-
ReviewNet (Yang et al. 2016)	-	-	-	0.290	0.237	-	0.886	-
LSTM-A5 (Yao et al. 2017)	0.730	0.565	0.429	0.325	0.251	0.538	0.986	-
Encoder-Decoder	0.718	0.547	0.412	0.311	0.251	0.530	0.961	0.179
Encoder-Decoder + Zoneout	0.708	0.537	0.403	0.304	0.249	0.525	0.941	0.176
Encoder-Decoder + Scheduled Sampling	0.718	0.548	0.414	0.315	0.252	0.531	0.975	0.180
Encoder-Decoder + ARNet	0.730	0.562	0.425	0.321	0.252	0.535	0.988	0.182
Att-Encoder-Decoder	0.727	0.557	0.421	0.318	0.259	0.537	0.996	0.185
Att-Encoder-Decoder + Zoneout	0.720	0.549	0.415	0.314	0.251	0.532	0.975	0.181
Att-Encoder-Decoder + Scheduled Sampling	0.731	0.563	0.426	0.322	0.256	0.538	1.006	0.187
Att-Encoder-Decoder + ARNet	0.740	0.576	0.440	0.335	0.261	0.546	1.034	0.190

Experiments Cont.

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.

- Mean centroid distance: $d_{mc}(\mathbf{U}, \mathbf{V}) = d(\frac{1}{B} \sum_i^B u_{i_i}, \frac{1}{B} \sum_j^B v_{j_j})$
- Point-wise distance: $d_{pw}(\mathbf{U}, \mathbf{V}) = \frac{1}{B} (\sum_{i=1}^B d(u_{i_i}, v_{i_i}))$

Model Name	d_{mc}	d_{pw}
Encoder-Decoder	0.747	0.719
Encoder-Decoder-ARNet	0.514	0.561
Att-Encoder-Decoder	0.773	0.760
Att-Encoder-Decoder-ARNet	0.491	0.595

Experiments Cont.

- Code Captioning on HabeasCorpus

Model Name	B-1	B-2	B-3	B-4	M	R
ReviewNet (Yang et al. 2016)	0.192	0.105	0.074	0.057	0.085	0.200
Encoder-Decoder	0.183	0.093	0.063	0.047	0.080	0.188
Encoder-Decoder + Zoneout	0.182	0.080	0.063	0.047	0.080	0.181
Encoder-Decoder + Scheduled Sampling	0.186	0.098	0.067	0.051	0.082	0.194
Encoder-Decoder + ARNet	0.196	0.107	0.075	0.058	0.089	0.213
Att-Encoder-Decoder	0.228	0.140	0.106	0.088	0.105	0.256
Att-Encoder-Decoder + Zoneout	0.227	0.140	0.105	0.086	0.090	0.220
Att-Encoder-Decoder + Scheduled Sampling	0.229	0.142	0.108	0.089	0.107	0.270
Att-Encoder-Decoder + ARNet	0.255	0.173	0.139	0.120	0.123	0.289

Model Name	d_{mc}	d_{pw}
Encoder-Decoder	0.643	0.722
Encoder-Decoder-ARNet	0.641	0.699
Att-Encoder-Decoder	0.594	0.712
Att-Encoder-Decoder-ARNet	0.322	0.465

- Permuted Sequential MNIST

Model Name	Test Accuracy
LSTM + Dropout (Le et al. 2015)	0.925
LSTM + Zoneout (Krueger et al. 2016)	0.931
Unregularized LSTM	0.914
LSTM + ARNet	0.933

Conclusions

- ARNet can regularize the transition dynamics and mitigate the discrepancy of RNN for sequence prediction tasks.
- Besides the caption generation tasks, ARNet can help model long term dependencies. Experiments on permuted sequential MNIST task demonstrate the superiority of our proposed ARNet.