REM-Net: Recursive Erasure Memory Network for Explanation Refinement on Commonsense Question Answering

Yinya Huang, Meng Fang, Xunlin Zhan, Qingxing Cao, Xiaodan Liang and Liang Lin

February 3, 2021

REM-Net: Recursive Erasure Memory Network for Explanation Refinement on Commonsense Question Answering

Yinya Huang, Meng Fang, Xunlin Zhan, Qingxing Cao, Xiaodan Liang, Liang Lin

Abstract

Commonsense Question Answering (commonsense QA) tasks aim to examine QA systems' capability of reasoning over context with complicated logical relationships and implicit commonsense knowledge. A desirable model should be able to provide not only correct answers but also persuasive explanations. Current works incorporate external knowledge presuming that valid explanations are included. However, the explanations are usually confounded and need further distinguishment. In this work, we propose a recursive erasure memory network (REM-Net), which learns to refine explanations for more precise interpretation while reasoning to obtain correct answers. REM-Net integrates a pre-trained knowledge graph generator, to provide possible explanations based on the commonsense question, and a recursive erasure memory module (REM), which refines the explanations. The REM module recursively erases confounding explanations to ensure that the model captures the most crucial clues. Experimental results on multiple commonsense QA benchmarks demonstrate that our REM-Net outperforms the competing methods. The case study also shows the model's ability to find more precise explanations.

Context
The seed germinates. The plant grows.
The plant flowers. Produces fruit.
The fruit releases seeds. The plant dies.
Question
Suppose less nutrients in the soil happens,
how will it affect less seeds germinates?
Answer Options
(A) More. (B) Less. (C) No effect.

Explanation Sentences

not is a good idea	is located at plant
not made of iron	is created by plant
causes starvation	is inherited from plant
is part of ecosystem	is related to soil decay
is a symbol of decay	is part of flower
has a less oxygen	is a plant
ends with die	requires soil
not capable of grow	has a no life
desires of water	desires of water
…	…

Figure 1: An example from the WIQA benchmark with some explanations (in the lower box) of the commonsense question (in the upper box). Some of the explanation sentences are confounded to the question: although semantically related to the question, the sentences in red are not beneficial to answer the question. By contrast, the sentences in blue explains the question well. Our REM-Net can successfully discover reasonable and supporting explanation sentences in blue.

1 Introduction

Commonsense question answering tasks (commonsense QA) need more complicated commonsense and logical reasoning in that the key information is mostly unexpressed and complicated. Solving these tasks requires to answer the questions by reasoning over context via mining reasonable explanations. This makes commonsense QA distinguished from the traditional machine reading comprehension (MRC) tasks, which can solely predict the answer via semantic match.

Current approaches that resort to explanations are mainly in three groups. The first group of methods (Devlin et al., 2019; Liu et al., 2019) are language models pre-trained on large-scale corpora that refer to diverse commonsense context. Those models are proved to contain certain commonsense knowledge (Tandon et al., 2019; Trinh and Le, 2018). Some of the approaches (Ye et al., 2019) further fine-tune the models to adapt to specific datasets. The second group of methods (Lv et al., 2020; Lin et al., 2019) incorporate external knowledge such as knowledge graph subgraphs and encodes the knowledge features via graph models such as GCN (Kipf and Welling, 2016). The third group of methods (Rajani et al., 2019) train models to generate explanations to facilitate the commonsense answer prediction. These approaches focus on enriching the model features with great amounts of external knowledge that are supposed to contain valid explanations to the commonsense questions. However, the quality of the incorporated explanations is not guaranteed, as some of the sentences could be invalid and confounding to the questions, but seldom of current methods develop a capability to distinguish them.

One example that shows the confoundedness of the explanations is presented in Figure 1. The explanation sentences are generated based on the commonsense question with COMET (Bosselut et al., 2019). Most of the explanation sentences are semantically related to the key phrases (i.e., "*less nutrients in the soil*" and "*less seeds germinates*") in the question, but they contribute differently to answering the question. For example, "*is part of flower*" conveys an attribute of the concept "*seeds*", but does not tell us how in fact it will affect "*less seeds germinates*". By contrast, "*causes starvation*" gives straightforward information that fills the causal gap between the key phrases "*less nutrients in the soil*" and "*less seeds germinates*". Therefore, sentences like "*is part of flower*" confounds the answering of the question, while "*causes starvation*" as an explanation is much more favorable. Our purpose in this work is to exploit a model that learns to discover the supporting explanations among the confounding ones so that to provide interpretations of answering commonsense questions.

In this paper, we study explanation refinement for commonsense QA tasks. With this purpose, we propose a model called recursive erasure memory network (REM-Net). The REM-Net consists of three main components: a query encoder, an explanation generator, and a recursive erasure memory module (REM). Specifically, the query encoder is a pre-trained language model that encodes the commonsense question. The explanation generator is a knowledge graph generator that is trained to generate commonsense knowledge triplets. The knowledge graph triplets are converted into plain sentences and provided as explanations to the question. This explanation generator module can be substituted by a retrieval-based module or simply adopting semantic embeddings from the query encoder. The recursive erasure memory module (REM) then refines the explanations by recursively erase the confounders. The REM module is a memory network that takes the question embeddings from the query encoder as the queries and the explanation embeddings from the explanation generator as memory. The query attentively visits the memory recursively to calculate scores of the extent to which each explanation sentence supports the question, the sentences that are regarded as confounding explanations are then erased.

We conduct experiments on two commonsense QA benchmarks (WIQA (Tandon et al., 2019) and CosmosQA (Huang et al., 2019)) and demonstrate that REM-Net outperforms current methods and produces reasonable refinement of the explanations. Our contributions are mainly three-fold:

- We propose a model called the recursive erasure memory network (REM-Net) towards recursively refining the explanations according to the commonsense question for better reasoning capability.
- The REM module incorporates an erasure manipulation into the memory network, so that to recursively estimate the explanation sentences and can distinguish the supporting sentences from the confounding ones. These manipulations indicate a further interpretation of how the question is being answered.
- Experimental results show that REM-Net outperforms competing methods. Besides, the case study presents the refined explanations, indicating that the refinement is reasonable since the discovered supporting sentences and confounding sentences are intuitive.

2 Related Works

Commonsense Question Answering Similar to open-domain question answering tasks (Rajpurkar et al., 2018; Kwiatkowski et al., 2019), commonsense question answering (Tandon et al., 2019; Huang et al., 2019) requires open-domain information to support the answer prediction. But different from open-domain question answering tasks that the text comprehension is straightforward and the retrieved open-domain information is direct to the questions, in commonsense question answering tasks the open-domain information is more complicated in that they play a role as explanations to bridge the understanding gap in the commonsense questions. Current works leverage the open-domain information by whether incorporating external knowledge as explanations or training the models to generate explanations. Lv et al. (2020) extracts knowledge from ConceptNet (Speer et al., 2017) and Wikipedia, and learns features with GCN (Kipf and Welling, 2016) and graph attention (Veličković et al., 2017). Zhong et al. (2019) retrieves ConceptNet (Speer et al., 2017) triplets and train two functions to measure direct and indirect connections between concepts. Rajani et al. (2019) train a GPT (Zhong et al., 2019) to generate reasonable explanations for the questions. During evaluation, the model generates explanations and predicts the multi-choice answers concurrently. Ye et al. (2019) automatically constructs a commonsense multi-choice dataset from ConceptNet triplets. However, the retrieved or generated explanations are usually not further refined, and some of them could be unnecessary or even confounding to answering the questions. The proposed model explores to refine the original explanations to discover those most supporting explanations to the commonsense questions and therefore provides stronger interpretations.

Memory Networks Memory networks (Weston et al., 2015; Bordes et al., 2015; Miller et al., 2016; Sukhbaatar et al., 2015) are proposed to solve early reasoning problems such as bAbI (Weston et al., 2016)) that requires to locate useful information for answer prediction. The sentences are stored into memory slots and later selected for the question answering. Recently, multi-head attention memory networks (Dai et al., 2019) are proposed so that takes advantage of the transformer-based networks. Our proposed model is based on multi-head attention memory network that is modified with a recursive erasure manipulation to adapt to the commonsense question answering tasks for accurate explanation refinement.

3 Recursive Erasure Memory Network

In this section, we introduce the proposed model, which consists of three main modules. The overview of the model is presented in Figure 2.

The initial query embedding is denoted as $\mathbf{q}^1$. The embedding of a single explanation sentence is denoted as $\mathbf{e}^1$, hence the overall explanation sentences for a question are denoted as a matrix $\mathbf{E}^1$. At recursive step t, the query embedding is similarly denoted as $\mathbf{q}^t$, the explanation matrix is denoted as $\mathbf{E}^t$, and the explanation scores are denoted as $\mathbf{s}^t$.

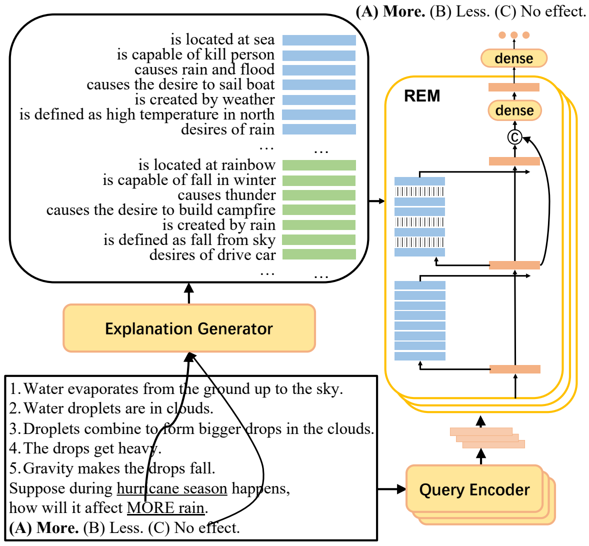

Figure 2: The proposed model REM-Net which consists of three main components. The query encoder encodes the commonsense question. The explanation generator generates explanation sentences. The recursive erasure memory module (REM) discovers the supporting explanations out of the confounding ones.

3.1 Query Encoder

Query encoder provides a primary understanding of the commonsense question, and the outputting embeddings contribute to the subsequent memory network as the initial queries.

The query encoder follows the baselines in the literature to use pre-trained Transformer-based encoder, so that the encoded query embeddings are rich in contextual information. In this paper, we use BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as the backbones. We follow the input format in Tandon et al. (2019) as "`[CLS] context [SEP] question [SEP] answer option`". The "`[CLS]`" embedding in the last BERT layer is taken as the outputting query embedding $\mathbf{q}^1 \in \mathbb{R}^h$, where h is hidden state size.

3.2 Explanation Generator

To obtain the initial possible explanations to the question, the explanation generator provides sentences that are related to the commonsense questions. Rather than retrieving subgraphs or texts from existing knowledge bases (e.g., ConceptNet, Wikipedia) using information retrieval techniques, where the retrieved explanations are restricted to the scope of the knowledge bases, we instead take advantage of the pre-trained generative model to obtain out-of-scope explanations.

The details of the explanation generator are pre-

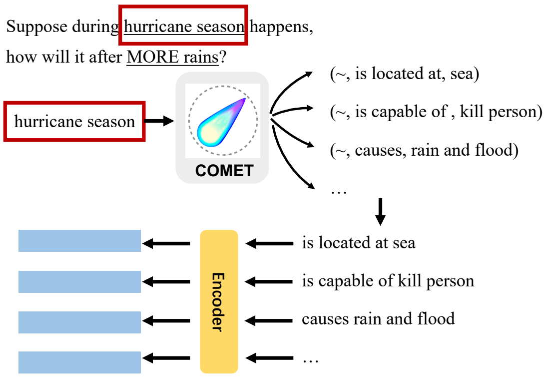

Figure 3: The explanation generator with a COMET (Bosselut et al., 2019) and an encoder. COMET takes the key phrases extracted from the commonsense questions as input and generates knowledge graph triplets. The triplets are then converted into explanation sentences with templates. The encoder is a pre-trained Transformer encoder that encodes the explanation sentences as embeddings.

sented in Figure 3, which is composed of a pre-trained knowledge graph generator and an encoder. Based on the commonsense question, key phrases are first extracted with pre-defined rules. The pre-trained knowledge graph generator COMET (Bosselut et al., 2019) treats the key phrases as head components of triplets, then generates relations and tails to form complete triplets. With the COMET templates [1], the knowledge graph triplets are then converted into sentences explain the commonsense question. A Transformer-based encoder then encodes the sentences into embeddings, which are then provided to the downstream memory network as initial memory embeddings.

Suppose there are I explanation sentences for each commonsense question, then each of them is formed into token sequence "`[CLS] explanation sentence [SEP]`". The `[CLS]` embedding in the last BERT layer is taken as the explanation embedding $\mathbf{e}_i^1 \in \mathbb{R}^h$, where the superscript 1 denotes the first recursive step, and h is hidden state size. The I explanation embeddings $\mathbf{e}_i^1, i \in \{1, ..., I\}$ are then formed into an explanation matrix $\mathbf{E}^1 \in \mathbb{R}^{I \times h}$ and fed into the memory network.

Context as Explanations To look into the model's capability of leveraging information at hand without augmenting any external knowledge, we develop a substitution of the explanation generator that solely uses the context paragraph in the original question sample as the explanation sentences. To obtain explanation embedding $\mathbf{e}_i^1 \in \mathbb{R}^h$, this substitution module directly takes the contextual token embeddings from the last multi-head attention layer in the query encoder and merges the token embeddings from the same context sentence by summation. The explanation embeddings are then formed into the explanation matrix $\mathbf{E}^1 \in \mathbb{R}^{I \times h}$, and fed into the memory network.

[1] https://mosaickg.apps.allenai.org/

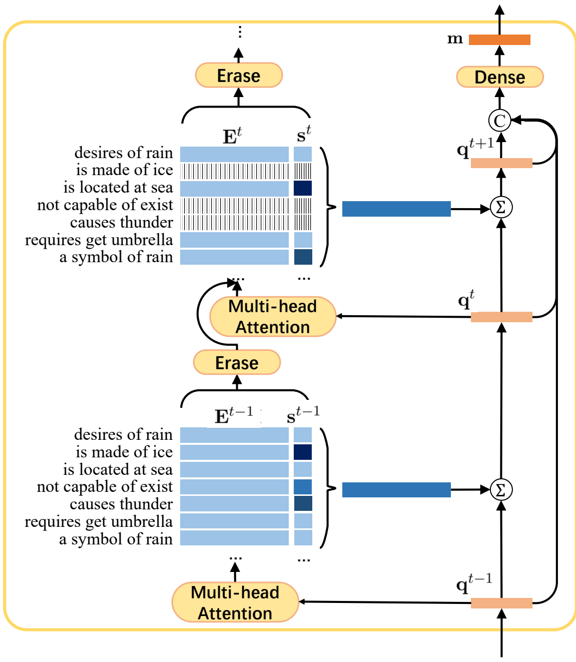

Figure 4: Details of the Recursive Erasure Memory (REM) module. The module is a memory network that conducts an erasure manipulation recursively. Multi-head attention calculates scores of each explanation embedding to estimate the extent to which the explanations support the question. The erasure manipulation is then conducted on the explanation matrix based on the scores, and pads some of the explanation embeddings.

3.3 Recursive Erasure Memory Module

The recursive erasure memory module (REM) is a multi-head attention memory network that conducts an erasure manipulation recursively to filter out unsupporting explanation sentences to the commonsense questions. The detailed architecture of the REM module is shown in Figure 4.

This module is launched by the initial query $\mathbf{q}^1$ from the query encoder and the initial explanation matrix $\mathbf{E}^1$ from the explanation generator (or its substitution module). It first calculates learnable scores of the explanation sentences using multi-head attention (Vaswani et al., 2017):

$$\mathbf{s}^1 = \text{MultiHead}(\mathbf{q}^1, \mathbf{E}^1, \mathbf{E}^1), \tag{1}$$

The scores $\mathbf{s}^1$ are then used for updating the query embedding $\mathbf{q}^1$ and conducting erasure manipulation on the explanation matrix $\mathbf{E}^1$. To update query embedding $\mathbf{q}^1$, the explanation matrix in the memory slots are weighted summed and merged into a single embedding and added to the original query embedding. To conduct the erasure manipulation on the explanations matrix $\mathbf{E}^1$, the explanation sentences are sorted by the scores and those with the k highest weights are padded and erased from the memory. This calculate-update-erase process is conducted recursively until termination.

Formally, at recursive step $t-1$, REM module conducts multi-head attention on the query $\mathbf{q}^{t-1} \in \mathbb{R}^h$ and the explanation matrix $\mathbf{E}^{t-1} \in \mathbb{R}^{I\times h}$, where $\mathbf{E}^{t-1}$ performs as key and value and $\mathbf{q}^{t-1}$ as query (Equation 2). Each explanation embedding is multiplied with the query embedding and it outputs the explanation scores $\mathbf{s}^{t-1} \in \mathbb{R}^I$:

$$\mathbf{s}^{t-1} = \text{MultiHead}(\mathbf{q}^{t-1}, \mathbf{E}^{t-1}, \mathbf{E}^{t-1}). \tag{2}$$

The scores are taken to weight the evidence matrix and update the query:

$$\mathbf{q}^t = \mathbf{q}^{t-1} + {\mathbf{E}^{t-1}}^\top \mathbf{s}^{t-1}. \tag{3}$$

The erasure manipulation is then conducted on the explanation matrix $\mathbf{E}^{t-1}$. The explanation embeddings are sorted by the scores, and those with the highest k scores are padded. The explanation matrix is then updated to updated to $\mathbf{E}^t$:

$$\mathbf{E}^t = \begin{bmatrix} \mathbf{e}_0^t \\ \mathbf{e}_1^t \\ \vdots \\ \mathbf{e}_I^t \end{bmatrix}, \; \mathbf{e}_i^t = \begin{cases} \mathbf{e}_i^{t-1}, \; s_i^{t-1} \geq s_{[k]}^{t-1}, \\ \\ \mathbf{0}, \; s_i^{t-1} < s_{[k]}^{t-1}, \end{cases} \tag{4}$$

where $s_{[k]}^{t-1}$ is the score ranking kth among $\mathbf{s}^{t-1}$.

Finally, queries in all recursive steps $\mathbf{q}^t, t \in \{1, ..., T\}$ are concatenated as the output of the REM module:

$$\mathbf{m} = [\mathbf{q}^1; ...; \mathbf{q}^T]\mathbf{W}_m + \mathbf{b}_m, \tag{5}$$

where $[;]$ indicates the concatenation operation, $\mathbf{m} \in \mathbb{R}^h$, $\mathbf{W}_m \in \mathbb{R}^{hT\times h}$, and $\mathbf{b}_m \in \mathbb{R}^h$.

3.4 Answer Prediction

The probabilities Pr of choosing the final answer option are:

$$Pr = \text{SoftMax}([\mathbf{m}_1; ...; \mathbf{m}_C]\mathbf{W}_p + b_p,), \tag{6}$$

where $[;]$ indicates the concatenation operation, C is the number of answer options, $\mathbf{p} \in \mathbb{R}^C$, $\mathbf{W}_p \in \mathbb{R}^{h\times 1}$, $b_p \in \mathbb{R}$.

4 Experiments

In this section, we conduct experiments to demonstrate the effectiveness of our proposed model and exhibit the refinement of the explanations.

4.1 Datasets

We experiment with two popular commonsense QA benchmarks, WIQA (Tandon et al., 2019) and CosmosQA (Huang et al., 2019).

- **CosmosQA** (Huang et al., 2019) includes questions of daily life scenarios, such as cultural norms, counterfactual reasoning, situational fact, and temporal event. The scenarios are plentiful and the questions are also diverse. The questions are in a multi-choice format.
- **WIQA** (Tandon et al., 2019) is a benchmark of counterfactual "what-if" questions. The context paragraphs provide descriptions of natural phenomenons, which are manually written based on specifically defined "influence graphs". The questions are split into three types ("in-para", "out-of-para", "no-effect") depending on whether the questions are derived from the original "influence graphs". For "out-of-para" and "no-effect" questions, the context paragraphs are irrelevant to the questions, so that they are unable to provide meaningful explanations.

4.2 Baselines

We compare our model with different groups of competitive models.

- **Group 1:** Baselines without pre-training. Most of the approaches within this group are taken from the benchmark papers. For WIQA, the Majority method (Tandon et al., 2019) predicts the most frequent answer option in the training set. The Polarity method (Tandon et al., 2019) predicts answers according to the way that the comparative words sentences collocates. Adaboost (Freund and Schapire, 1995) uses bag-of-words features from the questions. Decomp-Attn (Parikh et al., 2016) is a decomposable attention model that reformulates the dataset into an inference task. For CosmosQA, Sliding Window (Richardson et al., 2013) considers the similarity between the context paragraph and the answer options. Stanford Attentive Reader (Chen et al., 2016), Gated-Attention Reader (Dhingra et al., 2017)

Group	Method	Dev	Test
Group 1	Sliding Window (Richardson et al., 2013)	25.0	24.9
	Stanford Attentive Reader (Chen et al., 2016)	45.3	44.4
	Gated-Attention Reader (Dhingra et al., 2017)	46.9	46.2
	Co-Matching (Wang et al., 2018b)	45.9	44.7
Group 2	Commonsense-Rc (Wang et al., 2018a)	47.6	48.2
	GPT-FT (Radford et al., 2018)	54.0	54.4
	DMCN (Zhang et al., 2020)	67.1	67.6
	BERT-Large (Devlin et al., 2019)	66.2	67.1
	BERT-Large (ensemble)	67.1	67.5
	BERT-Large Multiway (Huang et al., 2019)	68.3	68.4
Group 3	MemN2N (Sukhbaatar et al., 2015)	30.6	31.0
	BERT-Large + explanations	67.1	67.2
	RoBERTa-Large + explanations	80.8	81.3
Ours	REM-Net-Large$_{text}$	67.9	68.5
	REM-Net-Large	69.5	70.1
	REM-Net-RoBERTa-Large$_{text}$	80.8	81.8
	REM-Net-RoBERTa-Large	**81.2**	**82.0**
Human perf.		-	94.0

Table 1: Result comparisons (%) on the CosmosQA development set and test set. Our models are compared with three groups of baselines.

and Co-Matching (Wang et al., 2018b) are reading comprehension systems that performs attention mechanism differently.

- **Group 2:** Pre-trained models without external explanations. Commonsense-RC (Wang et al., 2018a) is an LSTM-based model pre-trained on RACE (Lai et al., 2017). Transformer-based pre-trained language models such as GPT (Radford et al., 2018), BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are proved to contain some commonsense knowledge (Trinh and Le, 2018) since they are trained on large-scale corpora.
- **Group 3:** Models with external explanations. End-to-end memory networks (Sukhbaatar et al., 2015) are LSTM-based memory networks working on the external explanations stored in the memory slots. "BERT-Base + explanations", "BERT-Large + explanations" and "RoBERTa-Large + explanations" simply augment the question input by concatenating the external explanations.

4.3 Experimental Settings

We introduce the detailed experimental settings, including the settings used to generate the raw explanation sentences and the implementation details of our model.

4.3.1 Generating Raw Explanations

The explanation generator generates raw explanations based on the key phrases from the commonsense questions. For WIQA, in which the questions and answer options exhibit some regular patterns, in that the question consists of a cause clause (that starts with "*suppose*") and an effect clause (that starts with "*how will it affect*"), we extract key phrases out of both clauses. The cause key phrase and the effect key phrase are separately used to generate the explanation sentences. For CosmosQA, in which the questions and answer options are varied, we use the TAGME toolkit [2] for the extraction.

4.3.2 Implementation Details

We use BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as the backbones. The sequence length for the query encoder is 128, which is sufficient to include the "`[CLS] context [SEP] question [SEP] answer option`" sequence ($> 88\%$). For the explanation generator, the sequence length is set to 30, enabling it to include each explanation sentence ($> 99\%$).

When training WIQA, since there are two groups of explanation sentences (the cause group and the effect group), we adopt two parallel REM modules to separately refine the cause explanations and the effect explanations. The best number of recursive steps is 2. The upper bound of erased explanation sentences at each recursive step is set to 50. The model is optimized by Adam (Kingma and Ba, 2015) with a learning rate of 1×10^{-5}. Warmup steps are set to 1000. We train 25 epochs with batch size 8. For CosmosQA, we use a single REM

[2] https://tagme.d4science.org/tagme/

Group	Method	In-para	Out-of-para	No-effect	Total
Group 1	Majority (Tandon et al., 2019)*	45.46	49.47	0.55	30.66
	Polarity (Tandon et al., 2019)*	**76.31**	53.59	0.27	39.43
	Adaboost (Freund and Schapire, 1995)*	49.41	36.61	48.42	43.93
	Decomp-Att (Tandon et al., 2019)*	56.31	48.56	73.42	59.48
Group 2	BERT-Base (no para)	66.60	64.29	74.90	69.13
	BERT-Base	70.57	58.54	91.08	74.26
	BERT-Base (ensemble)	71.51	61.82	90.72	75.61
	BERT-Large	73.40	63.88	90.52	76.69
	BERT-Large (ensemble)	71.51	62.73	90.04	75.69
Group 3	MemN2N (Sukhbaatar et al., 2015)	38.50	38.01	39.52	38.85
	BERT-Base + explanations	70.57	61.00	90.72	75.12
	BERT-Large + explanations	73.40	63.88	90.52	76.69
Ours	REM-Net-Base$_{text}$	72.45	62.48	90.19	75.82
	REM-Net-Base	73.58	63.05	**91.71**	76.89
	REM-Net-Large$_{text}$	72.08	67.08	89.48	77.32
	REM-Net-Large	75.67	**67.98**	87.65	**77.56**
Human perf.		-	-	-	96.33

Table 2: Result comparisons (%) on the WIQA test set, including accuracies on three separate question types ("in-para", "out-of-para", "no-effect"), and the overall test set. The baselines labeled with * are directly taken from Tandon et al. (2019), where the test set is slightly different.

Context
After 15 years of paying premiums to Allstate , I have finally started the process of shopping for a new insurance company . I ca n't say I ' ve been unhappy with Allstate but it 's time to see if they are truly giving me a good deal or not . A couple things have caused me to do this .
Question and Options
Why is it a good idea to shop for insurance regularly ?
(A) Sometimes your current insurance will be too complacent with you .
(B) None of the above choices .
(C) You need to keep your insurance provider on their toes.
(D) It helps make sure that you are getting the best deal possible .
Erased Explanations
As a result, he/she feels sad.
As a result, he/she feels good.
As a result, he/she feels annoyed.
As a result, he/she feels satisfied.
As a result, he/she feels happy.
Reserved Explanations
Before, he/she needed have the information.
Because he/she wanted to have good quality of products.
He/she is seen as cautious.
He/she is seen as smart.
He/she is seen as responsible.

Table 3: Example of explanation refinement by the REM module. The question is taken from the CosmosQA development set, and the explanations are generated by the explanation generator. The erased and reserved explanations are presented.

module to refine the explanations. The best number of recursive step is 2. The upper bound of erased explanation sentences at each recursive step is set to 10. The model is optimized using the Adam optimizer with a learning rate of 5×10^{-6} and warmup steps of 1500. The model is trained with 10 epochs and a batch size of 4.

4.4 Experimental Results

The experimental results on CosmosQA and WIQA are presented in Table 1 and Table 2, respectively. Our proposed REM-Net is compared with three groups of baselines, as explained in Section 4.2. We report the results of two variants of our model, REM-Net$_{text}$ and REM-Net. REM-Net is the complete version of our model with the complete explanation generator used to produce raw explanations. The REM-Net$_{text}$ uses the variant of the explanation generator explained in Section 3.2 that takes only the context paragraphs as explanations. It is shown that our models outperform competitive baselines, which demonstrates that our models are effective. Moreover, the comparison with MemN2N (Sukhbaatar et al., 2015) and BERT, which incorporate the same explanation sentences, indicates that the performance boost of our model is beyond the augmentation of additional information. The recursive erasure manipulation on the explanations is beneficial.

4.5 Case Study

We showcase several examples to demonstrate REM-Net and REM-Net$_{text}$'s capability of explanation refinement. Table 3 presents an example of CosmosQA in which REM-Net refines the gen-

Context
The oil needs to be pumped from the ground.
After it is pumped it then is transported to a factory.
In the factory the oil is processed and turned into fuel.
Once the fuel is refined it is then sent to a truck.
By truck the fuel is sent to the gas station.
Question and Options
Suppose more oil is processed happens, how will it affect MORE oil arriving at gas stations ?
(A) More. (B) Less. (C) No effect.
Erased Explanations
The oil needs to be pumped from the ground.
After it is pumped it then is transported to a factory.
In the factory the oil is processed and turned into fuel.
Reserved Explanations
Once the fuel is refined it is then sent to a truck.
By truck the fuel is sent to the gas station.

Table 4: Example of explanation refinement by the REM module. The question is taken from the WIQA test set, and explanations are merely the sentences in the context paragraph. The erased sentences and retained sentences are presented.

erated explanations. The question concerns the reason for buying insurance regularly. The context paragraph tells a story about the narrator deciding to change his/her insurance products, but the reason for his/her decision is not provided. The generated explanations supply such reasons, hence benefits the understanding of the question. The erased explanations such as "*As a result, he/she feels sad*" or "*As a result, he/she feels happy*" are intuitively confounding to the question, since changing the insurance products are normally someone's rational decision. On the contrary, sentences like "*Because he/she wanted to have good quality of products*" support the question well, as they provide reasonable explanations. It is intuitive that the reserved explanations by REM-Net explain the reason better than the erased explanations. Table 4 provides an example of WIQA in which REM-Net$_{text}$ refines the context paragraph sentences. The example concerns the process of fuel production. REM-Net$_{text}$ erases the sentences talking about how the oil is turned into fuel and retains the explanations of how the oil being transported, which is reasonable for the question.

4.6 Ablation Study

To further investigate the benefits of each component of the proposed REM-Net, we conduct ablation studies on the explanation generator module and the REM module, the results of which are presented in Table 5 and Table 6. The performances of REM-Net are generally better than those of REM-Net$_{text}$. This is due to the augmented information

	Dev	Test
REM-Net-Large$_{text}$	67.87	68.53
w/o E	67.57	68.45
w/o E, w/o R	67.37	67.08
REM-Net-Large	**69.49**	**70.07**
w/o E	68.44	68.58
w/o E, w/o R	68.27	68.53

Table 5: Ablation studies on REM-Net-Large that are conducted on CosmosQA. E denotes the erasure manipulation, while R refers to the recursion mechanism.

	In-para	Out-of-para	No-effect	Total
REM-Net-Base$_{text}$	72.45	62.48	90.19	75.82
w/o E	71.32	61.41	90.04	75.12
w/o E, w/o R	70.94	60.18	91.31	75.09
REM-Net-Base	**73.58**	**63.05**	**91.71**	**76.89**
w/o E	72.64	62.97	**91.71**	76.69
w/o E, w/o R	71.89	60.34	91.55	75.42

Table 6: Ablation studies on REM-Net-Base that are conducted on WIQA. E signifies the erasure manipulation, while R indicates to the recursion mechanism.

provided by the explanation generator. Moreover, removing the erasure manipulation from the REM module leads to a performance decline. This indicates that excluding those confounding explanation sentences benefits the results. Further removing the recursive mechanism, which means the REM module calculates the explanation scores only once, brings a further performance drop. This indicates that recursively estimating the explanation sentences refines the understanding of the question and provides better interpretation. Therefore, erasure manipulation, the recursive mechanism, and the generated explanations all contribute to the benefits provided by our model.

5 Conclusion

In this paper, we propose a novel REM-Net that demonstrates superior reasoning capability in commonsense QA tasks while providing recursively refined commonsense explanations. REM-Net integrates an explanation generator and a REM module. The explanation generator provides possible explanations to the commonsense question, after which the REM module conducts a recursive erasure manipulation in order to refine the explanations. Experimental results demonstrate the effectiveness of REM-Net on commonsense QA tasks. Case study provides further evidence that REM-Net refines the explanations in a reasonable way by erasing the confounding explanations and discovering the supporting explanations to the questions.

References

Dharma P. Agrawal, Brij Bhooshan Gupta, Haoxiang Wang, Xiaojun Chang, Shingo Yamaguchi, and Gregorio Martínez Pérez. 2018. Guest editorial deep learning models for industry informatics. *IEEE Trans. Ind. Informatics*, 14(7):3166–3169.

Xiangpin Bai, Lei Zhu, Cheng Liang, Jingjing Li, Xiushan Nie, and Xiaojun Chang. 2020. Multi-view feature selection via nonnegative structured graph learning. *Neurocomputing*, 387:110–122.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *ArXiv, abs/1506.02075*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proc. of ACL*.

Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G. Hauptmann. 2018a. RCAA: relational context-aware agents for person search. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*, pages 86–102.

Xiaojun Chang, Xiaodan Liang, Yan Yan, and Liqiang Nie. 2020. Guest editorial: Image/video understanding and analysis. *Pattern Recognit. Lett.*, 130:1–3.

Xiaojun Chang, Wenhe Liu, Po-Yao Huang, Changlin Li, Fengda Zhu, Mingfei Han, Mingjie Li, Mengyuan Ma, Siyi Hu, Guoliang Kang, Junwei Liang, Liangke Gui, Lijun Yu, Yijun Qian, Jing Wen, and Alexander G. Hauptmann. 2019. Mmvg-inf-etrol@trecvid 2019: Activities in extended video. In *2019 TREC Video Retrieval Evaluation, TRECVID 2019, Gaithersburg, MD, USA, November 12-13, 2019*.

Xiaojun Chang, Zhigang Ma, Ming Lin, Yi Yang, and Alexander G. Hauptmann. 2017a. Feature interaction augmented sparse learning for fast kinect motion detection. *IEEE Trans. Image Process.*, 26(8):3911–3920.

Xiaojun Chang, Zhigang Ma, Yi Yang, Zhiqiang Zeng, and Alexander G. Hauptmann. 2017b. Bi-level semantic representation analysis for multimedia event detection. *IEEE Trans. Cybern.*, 47(5):1180–1197.

Xiaojun Chang, Feiping Nie, Zhigang Ma, Yi Yang, and Xiaofang Zhou. 2015a. A convex formulation for spectral shrunk clustering. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2532–2538.

Xiaojun Chang, Feiping Nie, Sen Wang, Yi Yang, Xiaofang Zhou, and Chengqi Zhang. 2016a. Compound rank-k projections for bilinear analysis. *IEEE Trans. Neural Networks Learn. Syst.*, 27(7):1502–1513.

Xiaojun Chang, Feiping Nie, Yi Yang, and Heng Huang. 2014a. A convex formulation for semi-supervised multi-label feature selection. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pages 1171–1177.

Xiaojun Chang, Feiping Nie, Yi Yang, Chengqi Zhang, and Heng Huang. 2016b. Convex sparse PCA for unsupervised feature learning. *ACM Trans. Knowl. Discov. Data*, 11(1):3:1–3:16.

Xiaojun Chang, Haoquan Shen, Sen Wang, Jiajun Liu, and Xue Li. 2014b. Semi-supervised feature analysis for multimedia annotation by mining label correlation. In *Advances in Knowledge Discovery and Data Mining - 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part II*, pages 74–85.

Xiaojun Chang, Yan Yan, and Liqiang Nie. 2018b. Guest editorial: Semantic concept discovery in MM data. *Multim. Tools Appl.*, 77(3):2945–2946.

Xiaojun Chang and Yi Yang. 2017. Semisupervised feature analysis by mining correlations among multiple tasks. *IEEE Trans. Neural Networks Learn. Syst.*, 28(10):2294–2305.

Xiaojun Chang, Yi Yang, Alexander G. Hauptmann, Eric P. Xing, and Yaoliang Yu. 2015b. Semantic concept discovery for large-scale zero-shot event detection. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 2234–2240.

Xiaojun Chang, Yi Yang, Guodong Long, Chengqi Zhang, and Alexander G. Hauptmann. 2016c. Dynamic concept composition for zero-example event detection. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3464–3470.

Xiaojun Chang, Yi Yang, Eric P. Xing, and Yaoliang Yu. 2015c. Complex event detection using semantic saliency and nearly-isotonic SVM. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1348–1357.

Xiaojun Chang, Yaoliang Yu, and Yi Yang. 2017c. Robust top-*k* multiclass SVM for visual category recognition. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 75–83.

Xiaojun Chang, Yaoliang Yu, Yi Yang, and Alexander G. Hauptmann. 2015d. Searching persuasively: Joint event detection and evidence recounting with limited supervision. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference,* *MM '15, Brisbane, Australia, October 26 - 30, 2015*, pages 581–590.

Xiaojun Chang, Yaoliang Yu, Yi Yang, and Eric P. Xing. 2016d. They are not equally reliable: Semantic event search using differentiated concept classifiers. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1884–1893.

Xiaojun Chang, Yaoliang Yu, Yi Yang, and Eric P. Xing. 2017d. Semantic pooling for complex event analysis in untrimmed videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(8):1617–1632.

Yan-shuo Chang, Feiping Nie, Zhihui Li, Xiaojun Chang, and Heng Huang. 2017e. Refined spectral clustering via embedded label propagation. *Neural Comput.*, 29(12).

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Proc. of ACL*.

Kaixuan Chen, Lina Yao, Dalin Zhang, Xiaojun Chang, Guodong Long, and Sen Wang. 2019. Distributionally robust semi-supervised learning for people-centric sensing. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3321–3328.

Kaixuan Chen, Lina Yao, Dalin Zhang, Xianzhi Wang, Xiaojun Chang, and Feiping Nie. 2020. A semisupervised recurrent convolutional attention model for human activity recognition. *IEEE Trans. Neural Networks Learn. Syst.*, 31(5):1747–1756.

Xiaojun Chen, Guowen Yuan, Wenting Wang, Feiping Nie, Xiaojun Chang, and Joshua Zhexue Huang. 2018. Local adaptive projection framework for feature selection of labeled and unlabeled data. *IEEE Trans. Neural Networks Learn. Syst.*, 29(12):6362–6373.

De Cheng, Xiaojun Chang, Li Liu, Alexander G. Hauptmann, Yihong Gong, and Nanning Zheng. 2017. Discriminative dictionary learning with ranking metric embedded for person re-identification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 964–970.

De Cheng, Yihong Gong, Xiaojun Chang, Weiwei Shi, Alexander G. Hauptmann, and Nanning Zheng. 2018. Deep feature learning via structured graph laplacian embedding for person re-identification. *Pattern Recognit.*, 82:94–104.

Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Tom Drummond, Hongdong Li, and Zongyuan Ge. 2020a. Hierarchical neural architecture search for deep stereo matching. *CoRR*, abs/2010.13501.

Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. 2020b. Hierarchical neural architecture search for deep stereo matching. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose Catherine Kanjirathinkal, and Mohan S. Kankanhalli. 2019. MMALFM: explainable recommendation by leveraging reviews and images. *ACM Trans. Inf. Syst.*, 37(2):16:1–16:28.

Zehui Dai, Wei Dai, Zhenhua Liu, Fengyun Rao, Huajie Chen, Guangpeng Zhang, Yadong Ding, and Jiyang Liu. 2019. Multi-task multi-head attention memory network for fine-grained sentiment analysis. In *Proc. of NLPCC*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proc. of ACL*.

Ali Emami, Adam Trischler, Kaheer Suleman, Noelia De La Cruz, and Jackie Chi Kit Cheung. 2018. A knowledge hunting framework for common sense reasoning. In *Proc. of EMNLP*.

Hehe Fan, Xiaojun Chang, De Cheng, Yi Yang, Dong Xu, and Alexander G. Hauptmann. 2017a. Complex event detection by identifying reliable shots from untrimmed videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 736–744.

Mingyu Fan, Xiaojun Chang, and Dacheng Tao. 2017b. Structure regularized unsupervised discriminant feature analysis. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1870–1876.

Mingyu Fan, Xiaojun Chang, Xiaoqin Zhang, Di Wang, and Liang Du. 2017c. Top-k supervise feature selection via ADMM for integer programming. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1646–1653.

Yoav Freund and Robert E Schapire. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer.

Zongyuan Ge, Dwarikanath Mahapatra, Xiaojun Chang, Zetao Chen, Lianhua Chi, and Huimin Lu. 2020. Improving multi-label chest x-ray disease diagnosis by exploiting disease and health labels dependencies. *Multim. Tools Appl.*, 79(21-22):14889–14902.

Chen Gong, Xiaojun Chang, Meng Fang, and Jian Yang. 2018. Teaching semi-supervised classifier via generalized distillation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 2156–2162.

Chen Gong, Dacheng Tao, Xiaojun Chang, and Jian Yang. 2019. Ensemble teaching for hybrid label propagation. *IEEE Trans. Cybern.*, 49(2):388–402.

Liangke Gui, Xiaodan Liang, Xiaojun Chang, and Alexander G. Hauptmann. 2018. Adaptive context-aware reinforced agent for handwritten text recognition. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 207.

Junwei Han, Le Yang, Dingwen Zhang, Xiaojun Chang, and Xiaodan Liang. 2018. Reinforcement cutting-agent learning for video object segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9080–9089.

Longfei Han, Dingwen Zhang, Dong Huang, Xiaojun Chang, Jun Ren, Senlin Luo, and Junwei Han. 2017. Self-paced mixture of regressions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1816–1822.

Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao. 2020. Mining inter-video proposal relations for video object detection. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*, pages 431–446.

Siyi Hu and Xiaojun Chang. 2020. Multi-view drone-based geo-localization via style and spatial alignment. *CoRR*, abs/2006.13681.

Siyi Hu, Fengda Zhu, Xiaojun Chang, and Xiaodan Liang. 2021. Updet: Universal multi-agent reinforcement learning via policy decoupling with transformers. *CoRR*, abs/2101.08001.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proc. of EMNLP*.

Po-Yao Huang, Xiaojun Chang, and Alexander G. Hauptmann. 2019a. Multi-head attention with diversity for learning grounded multilingual multimodal representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1461–1467.

Po-Yao Huang, Xiaojun Chang, Alexander G. Hauptmann, and Eduard H. Hovy. 2020a. Forward and backward multimodal NMT for improved monolingual and multilingual cross-modal retrieval. In *Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, June 8-11, 2020*, pages 53–62.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander G. Hauptmann. 2020b. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8226–8237.

Po-Yao Huang, Guoliang Kang, Wenhe Liu, Xiaojun Chang, and Alexander G. Hauptmann. 2019b. Annotation efficient cross-modal retrieval with adversarial attentive alignment. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 1758–1767.

Po-Yao Huang, Vaibhav, Xiaojun Chang, and Alexander G. Hauptmann. 2019c. Improving what cross-modal retrieval models learn through object-oriented inter- and intra-modal attention networks. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10-13, 2019*, pages 244–252.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *ArXiv, abs/1609.02907*.

Tassilo Klein and Moin Nabi. 2019. Attention is (not) all you need for commonsense reasoning. In *Proc. of ACL*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Proc. of ACL*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proc. of EMNLP*.

Changlin Li, Jiefeng Peng, Liuchun Yuan, Guangrun Wang, Xiaodan Liang, Liang Lin, and Xiaojun

Chang. 2020a. Block-wisely supervised neural architecture search with knowledge distillation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1986–1995.

Mingjie Li, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. 2020b. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *CoRR*, abs/2006.03744.

Zhihui Li, Xiaojun Chang, Lina Yao, Shirui Pan, Zongyuan Ge, and Huaxiang Zhang. 2020c. Grounding visual concepts for zero-shot event detection and event captioning. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 297–305.

Zhihui Li, Wenhe Liu, Xiaojun Chang, Lina Yao, Mahesh Prakash, and Huaxiang Zhang. 2019a. Domain-aware unsupervised cross-dataset person reidentification. In *Advanced Data Mining and Applications - 15th International Conference, ADMA 2019, Dalian, China, November 21-23, 2019, Proceedings*, pages 406–420.

Zhihui Li, Feiping Nie, Xiaojun Chang, Zhigang Ma, and Yi Yang. 2018a. Balanced clustering via exclusive lasso: A pragmatic approach. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3596–3603.

Zhihui Li, Feiping Nie, Xiaojun Chang, Liqiang Nie, Huaxiang Zhang, and Yi Yang. 2018b. Rank-constrained spectral clustering with flexible embedding. *IEEE Trans. Neural Networks Learn. Syst.*, 29(12):6073–6082.

Zhihui Li, Feiping Nie, Xiaojun Chang, and Yi Yang. 2017. Beyond trace ratio: Weighted harmonic mean of trace ratios for multiclass discriminant analysis. *IEEE Trans. Knowl. Data Eng.*, 29(10):2100–2110.

Zhihui Li, Feiping Nie, Xiaojun Chang, Yi Yang, Chengqi Zhang, and Nicu Sebe. 2018c. Dynamic affinity graph construction for spectral clustering using multiple features. *IEEE Trans. Neural Networks Learn. Syst.*, 29(12):6323–6332.

Zhihui Li, Lina Yao, Xiaojun Chang, Kun Zhan, Jiande Sun, and Huaxiang Zhang. 2019b. Zero-shot event detection via event-adaptive concept relevance mining. *Pattern Recognit.*, 88:595–603.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Proc. of EMNLP*.

Chong Liu, Xiaojun Chang, and Yi-Dong Shen. 2020a. Unity style transfer for person re-identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6886–6895.

Huan Liu, Qinghua Zheng, Minnan Luo, Xiaojun Chang, Caixia Yan, and Lina Yao. 2020b. Memory transformation networks for weakly supervised visual classification. *Knowl. Based Syst.*, 210:106432.

Huan Liu, Qinghua Zheng, Minnan Luo, Dingwen Zhang, Xiaojun Chang, and Cheng Deng. 2017a. How unlabeled web videos help complex event detection? In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4040–4046.

Wenhe Liu, Xiaojun Chang, Ling Chen, Dinh Phung, Xiaoqin Zhang, Yi Yang, and Alexander G. Hauptmann. 2020c. Pair-based uncertainty and diversity promoting early active learning for person reidentification. *ACM Trans. Intell. Syst. Technol.*, 11(2):21:1–21:15.

Wenhe Liu, Xiaojun Chang, Ling Chen, and Yi Yang. 2017b. Early active learning with pairwise constraint for person re-identification. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part I*, pages 103–118.

Wenhe Liu, Xiaojun Chang, Ling Chen, and Yi Yang. 2018a. Semi-supervised bayesian attribute learning for person re-identification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7162–7169.

Wenhe Liu, Xiaojun Chang, Yan Yan, Yi Yang, and Alexander G. Hauptmann. 2018b. Few-shot text and image classification via analogical transfer learning. *ACM Trans. Intell. Syst. Technol.*, 9(6):71:1–71:20.

Wenhe Liu, Guoliang Kang, Po-Yao Huang, Xiaojun Chang, Lijun Yu, Yijun Qian, Junwei Liang, Liangke Gui, Jing Wen, Peng Chen, and Alexander G. Hauptmann. 2020d. Argus: Efficient activity detection system for extended video analysis. In *IEEE Winter Applications of Computer Vision Workshops, WACV Workshops 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 126–133.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv, abs/1907.11692*.

Minnan Luo, Xiaojun Chang, Zhihui Li, Liqiang Nie, Alexander G. Hauptmann, and Qinghua Zheng. 2017a. Simple to complex cross-modal learning to rank. *Comput. Vis. Image Underst.*, 163:67–77.

Minnan Luo, Xiaojun Chang, Liqiang Nie, Yi Yang, Alexander G. Hauptmann, and Qinghua Zheng. 2018a. An adaptive semisupervised feature analysis for video semantic recognition. *IEEE Trans. Cybern.*, 48(2):648–660.

Minnan Luo, Feiping Nie, Xiaojun Chang, Yi Yang, Alexander G. Hauptmann, and Qinghua Zheng. 2016. Avoiding optimal mean robust PCA/2DPCA with non-greedy l_1-norm maximization. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 1802–1808.

Minnan Luo, Feiping Nie, Xiaojun Chang, Yi Yang, Alexander G. Hauptmann, and Qinghua Zheng. 2017b. Avoiding optimal mean $l_{2,1}$-norm maximization-based robust PCA for reconstruction. *Neural Comput.*, 29(4):1124–1150.

Minnan Luo, Feiping Nie, Xiaojun Chang, Yi Yang, Alexander G. Hauptmann, and Qinghua Zheng. 2018b. Adaptive unsupervised feature selection with structure regularization. *IEEE Trans. Neural Networks Learn. Syst.*, 29(4):944–956.

Minnan Luo, Caixia Yan, Qinghua Zheng, Xiaojun Chang, Ling Chen, and Feiping Nie. 2019. Discrete multi-graph clustering. *IEEE Trans. Image Process.*, 28(9):4701–4712.

Minnan Luo, Lingling Zhang, Feiping Nie, Xiaojun Chang, Buyue Qian, and Qinghua Zheng. 2017c. Adaptive semi-supervised learning with discriminative least squares regression. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2421–2427.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. *Proc. of AAAI*.

Zhigang Ma, Xiaojun Chang, Zhongwen Xu, Nicu Sebe, and Alexander G. Hauptmann. 2018. Joint attributes and event analysis for multimedia event detection. *IEEE Trans. Neural Networks Learn. Syst.*, 29(7):2921–2930.

Zhigang Ma, Xiaojun Chang, Yi Yang, Nicu Sebe, and Alexander G. Hauptmann. 2017. The many shades of negativity. *IEEE Trans. Multim.*, 19(7):1558–1568.

Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proc. of EMNLP*.

Deepak Ranjan Nayak, Ratnakar Dash, Xiaojun Chang, Banshidhar Majhi, and Sambit Bakshi. 2020. Automated diagnosis of pathological brain using fast curvelet entropy features. *IEEE Trans. Sustain. Comput.*, 5(3):416–427.

Liqiang Nie, Luming Zhang, Lei Meng, Xuemeng Song, Xiaojun Chang, and Xuelong Li. 2017a. Modeling disease progression via multisource multitask learners: A case study with alzheimer's disease. *IEEE Trans. Neural Networks Learn. Syst.*, 28(7):1508–1519.

Liqiang Nie, Luming Zhang, Yan Yan, Xiaojun Chang, Maofu Liu, and Ling Shaoling. 2017b. Multiview physician-specific attributes fusion for health seeking. *IEEE Trans. Cybern.*, 47(11):3680–3691.

Ram Prasad Padhy, Xiaojun Chang, Suman Kumar Choudhury, Pankaj Kumar Sa, and Sambit Bakshi. 2019. Multi-stage cascaded deconvolution for depth map and surface normal prediction from single image. *Pattern Recognit. Lett.*, 127:165–173.

Ankur P. Parikh, Oscar Tackstrom, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proc. of EMNLP*.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proc. of ACL*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *Proc. of ACL*.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. 2020a. A comprehensive survey of neural architecture search: Challenges and solutions. *CoRR*, abs/2006.02903.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Mahesh Prakash, Feiping Nie, Xin Wang, and Xiaojiang Chen. 2020b. Structured optimal graph-based clustering with flexible embedding. *IEEE Trans. Neural Networks Learn. Syst.*, 31(10):3801–3813.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proc. of EMNLP*.

Imad Rida, Sambit Bakshi, Xiaojun Chang, and Hugo Proença. 2019. Forensic shoe-print identification: A brief survey. *CoRR*, abs/1901.01431.

Maarten Sap, Ronan Le Bras, Emily Allaway, Hannah Rashkin, Chandra Bhagavatula, Nicholas Lourie, Brendan Roof, Noah Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *Proc. of AAAI*.

Amit Kumar Singh, Zhihan Lv, Huimin Lu, and Xiaojun Chang. 2020. Guest editorial: Recent trends in multimedia data-hiding: a reliable mean for secure communications. *J. Ambient Intell. Humaniz. Comput.*, 11(5):1795–1797.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proc. of AAAI*.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. *Proc. of NIPS*.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. Wiqa: A dataset for "what if…" reasoning over procedural text. In *Proc. of EMNLP*.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *ArXiv, abs/1806.02847*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NIPS*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks. *Proc. of ICLR*.

Fei Wang, Lei Zhu, Cheng Liang, Jingjing Li, Xiaojun Chang, and Ke Lu. 2020a. Robust optimal graph clustering. *Neurocomputing*, 378:153–165.

Hanmo Wang, Xiaojun Chang, Lei Shi, Yi Yang, and Yi-Dong Shen. 2018. Uncertainty sampling for action recognition via maximizing expected average precision. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 964–970.

Liang Wang, Meng Sun, Wei Zhao, Kewei Shen, and Jingming Liu. 2018a. Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. In *Proc. of SemEval*.

Rong Wang, Feiping Nie, Richang Hong, Xiaojun Chang, Xiaojun Yang, and Weizhong Yu. 2017a. Fast and orthogonal locality preserving projections for dimensionality reduction. *IEEE Trans. Image Process.*, 26(10):5019–5030.

Sen Wang, Xiaojun Chang, Xue Li, Guodong Long, Lina Yao, and Quan Z. Sheng. 2016a. Diagnosis code assignment using sparsity-based disease correlation embedding. *IEEE Trans. Knowl. Data Eng.*, 28(12):3191–3202.

Sen Wang, Xue Li, Xiaojun Chang, Lina Yao, Quan Z. Sheng, and Guodong Long. 2017b. Learning multiple diagnosis codes for ICU patients with local disease correlation mining. *ACM Trans. Knowl. Discov. Data*, 11(3):31:1–31:21.

Sen Wang, Feiping Nie, Xiaojun Chang, Xue Li, Quan Z. Sheng, and Lina Yao. 2016b. Uncovering locally discriminative structure for feature analysis. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*, pages 281–295.

Sen Wang, Feiping Nie, Xiaojun Chang, Lina Yao, Xue Li, and Quan Z. Sheng. 2015. Unsupervised feature analysis with class margin optimization. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I*, pages 383–398.

Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang. 2018b. A co-matching model for multi-choice reading comprehension. In *Proc. of ACL*.

Weitao Wang, Ruyang Liu, Meng Wang, Sen Wang, Xiaojun Chang, and Yang Chen. 2020b. Memory-based network for scene graph with unbalanced relations. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 2400–2408.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael J Witbrock. 2019. Improving natural language inference using external knowledge in the science questions domain. *Proc. of AAAI*, 33.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *Proc. of ICLR*.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. In *Proc. of ICLR*.

Man Wu, Shirui Pan, Chuan Zhou, Xiaojun Chang, and Xingquan Zhu. 2020a. Unsupervised domain adaptive graph convolutional networks. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1457–1467.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. 2020b. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 753–763.

Shicheng Xu, Huan Li, Xiaojun Chang, Shoou-I Yu, Xingzhong Du, Xuanchong Li, Lu Jiang, Zexi Mao, Zhen-Zhong Lan, Susanne Burger, and Alexander G. Hauptmann. 2015. Incremental multimodal query construction for video search. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, June 23-26, 2015*, pages 675–678.

Xiaowei Xue, Feiping Nie, Sen Wang, Xiaojun Chang, Bela Stantic, and Min Yao. 2017. Multi-view correlated feature learning by uncovering shared component. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 2810–2816.

Caixia Yan, Xiaojun Chang, Minnan Luo, Qinghua Zheng, Xiaoqin Zhang, Zhihui Li, and Feiping Nie. 2020a. Self-weighted robust LDA for multiclass classification with edge classes. *CoRR*, abs/2009.12362.

Caixia Yan, Qinghua Zheng, Xiaojun Chang, Minnan Luo, Chung-Hsing Yeh, and Alexander G. Hauptmann. 2020b. Semantics-preserving graph propagation for zero-shot object detection. *IEEE Trans. Image Process.*, 29:8163–8176.

Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G. Hauptmann. 2015. Multi-class active learning by uncertainty sampling with diversity maximization. *Int. J. Comput. Vis.*, 113(2):113–127.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proc. of NeurIPS*.

Zhi-Xiu Ye, Qian Chen, Wen Wang, and Zhen-Hua Ling. 2019. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. *ArXiv, abs/1908.06725*.

En Yu, Wenhe Liu, Guoliang Kang, Xiaojun Chang, Jiande Sun, and Alexander G. Hauptmann. 2019a. Inf@trecvid 2019: Instance search task. In *2019 TREC Video Retrieval Evaluation, TRECVID 2019, Gaithersburg, MD, USA, November 12-13, 2019*.

En Yu, Jiande Sun, Jing Li, Xiaojun Chang, Xian-Hua Han, and Alexander G. Hauptmann. 2019b. Adaptive semi-supervised feature selection for cross-modal retrieval. *IEEE Trans. Multim.*, 21(5):1276–1288.

En Yu, Jiande Sun, Li Wang, Xiaojun Chang, Huaxiang Zhang, and Alexander G. Hauptmann. 2019c. Cross-modal transfer hashing based on coherent projection. In *IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2019, Shanghai, China, July 8-12, 2019*, pages 477–482.

Shoou-I Yu, Lu Jiang, Zhongwen Xu, Zhenzhong Lan, Shicheng Xu, Xiaojun Chang, Xuanchong Li, Zexi Mao, Chuang Gan, Yajie Miao, Xingzhong Du, Yang Cai, Lara J. Martin, Nikolas Wolfe, Anurag Kumar, Huan Li, Ming Lin, Zhigang Ma, Yi Yang, Deyu Meng, Shiguang Shan, Pinar Duygulu Sahin, Susanne Burger, Florian Metze, Rita Singh, Bhiksha Raj, Teruko Mitamura, Richard M. Stern, and Alexander G. Hauptmann. 2015. CMU informedia@trecvid 2015: MED/SIN/LNK/SED. In *2015 TREC Video Retrieval Evaluation, TRECVID 2015, Gaithersburg, MD, USA, November 16-18, 2015*.

Zhen Yu, Jennifer Nguyen, Xiaojun Chang, John Kelly, Catriona McLean, Lei Zhang, Victoria Mar, and Zongyuan Ge. 2020. Melanoma diagnosis with spatio-temporal feature learning on sequential dermoscopic images. *CoRR*, abs/2006.10950.

Di Yuan, Xiaojun Chang, and Zhenyu He. 2020. Accurate bounding-box regression with distance-iou loss for visual tracking. *CoRR*, abs/2007.01864.

Di Yuan, Xiaojun Chang, Po-Yao Huang, Qiao Liu, and Zhenyu He. 2021. Self-supervised deep correlation tracking. *IEEE Trans. Image Process.*, 30:976–985.

Kun Zhan, Xiaojun Chang, Junpeng Guan, Ling Chen, Zhigang Ma, and Yi Yang. 2019. Adaptive structure discovery for multimedia analysis using multiple features. *IEEE Trans. Cybern.*, 49(5):1826–1834.

Dalin Zhang, Lina Yao, Kaixuan Chen, Sen Wang, Xiaojun Chang, and Yunhao Liu. 2020a. Making sense of spatio-temporal preserving representations for eeg-based human intention recognition. *IEEE Trans. Cybern.*, 50(7):3033–3044.

Dingwen Zhang, Junwei Han, Lu Jiang, Senmao Ye, and Xiaojun Chang. 2017. Revealing event saliency in unconstrained video collection. *IEEE Trans. Image Process.*, 26(4):1746–1758.

Jiaqi Zhang, Meng Wang, Qinchi Li, Sen Wang, Xiaojun Chang, and Beilun Wang. 2020b. Quadratic sparse gaussian graphical model estimation method for massive variables. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 2964–2972.

Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Mahesh Prakash, and Alexander G. Hauptmann. 2020c. Few-shot activity recognition with cross-modal memory network. *Pattern Recognit.*, 108:107348.

Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Sen Wang, Zongyuan Ge, and Alexander G. Hauptmann. 2020d. ZSTAD: zero-shot temporal activity detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 876–885.

Lingling Zhang, Jun Liu, Minnan Luo, Xiaojun Chang, and Qinghua Zheng. 2018. Deep semisupervised zero-shot learning with maximum mean discrepancy. *Neural Comput.*, 30(5).

Lingling Zhang, Jun Liu, Minnan Luo, Xiaojun Chang, Qinghua Zheng, and Alexander G. Hauptmann. 2019. Scheduled sampling for one-shot learning via matching network. *Pattern Recognit.*, 96.

Lingling Zhang, Minnan Luo, Jun Liu, Xiaojun Chang, Yi Yang, and Alexander G. Hauptmann. 2020e. Deep top-k ranking for image-sentence matching. *IEEE Trans. Multim.*, 22(3):775–785.

Miao Zhang, Huiqi Li, Shirui Pan, Xiaojun Chang, Zongyuan Ge, and Steven W. Su. 2020f. Differentiable neural architecture search in equivalent space with exploration enhancement. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Miao Zhang, Huiqi Li, Shirui Pan, Xiaojun Chang, and Steven W. Su. 2020g. Overcoming multi-model forgetting in one-shot NAS with diversity maximization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 7806–7815.

Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2020. Dcmn+: Dual co-matching network for multi-choice reading comprehension. *Proc. of AAAI*.

Zhicheng Zhao, Xuanchong Li, Xingzhong Du, Qi Chen, Yanyun Zhao, Fei Su, Xiaojun Chang, and Alexander G. Hauptmann. 2018. A unified framework with a benchmark dataset for surveillance event detection. *Neurocomputing*, 278:62–74.

Wanjun Zhong, Duyu Tang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2019. Improving question answering by commonsense-based pre-training. In *Proc. of NLPCC*.

Runwu Zhou, Xiaojun Chang, Lei Shi, Yi-Dong Shen, Yi Yang, and Feiping Nie. 2020. Person reidentification via multi-feature fusion with adaptive graph learning. *IEEE Trans. Neural Networks Learn. Syst.*, 31(5):1592–1601.

Fengda Zhu, Xiaojun Chang, Runhao Zeng, and Mingkui Tan. 2019. Continual reinforcement learning with diversity exploration and adversarial self-correction. *CoRR*, abs/1906.09205.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. 2020a. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10009–10019.

Lei Zhu, Zi Huang, Xiaojun Chang, Jingkuan Song, and Heng Tao Shen. 2017. Exploring consistent preferences: Discrete hashing with pair-exemplar for scalable landmark search. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 726–734.

Yi Zhu, Fengda Zhu, Zhaohuan Zhan, Bingqian Lin, Jianbin Jiao, Xiaojun Chang, and Xiaodan Liang. 2020b. Vision-dialog navigation by exploring cross-modal memory. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10727–10736.