# Multi-Task Recurrent Modular Networks

Dongkuan Xu[1], Wei Cheng[2], Xin Dong[3], Bo Zong[2], Wenchao Yu[2], Jingchao Ni[2],
Dongjin Song[4], Xuchao Zhang[2], Haifeng Chen[2], Xiang Zhang[1]

*[1]The Pennsylvania State University, [2]NEC Laboratories America, Inc., [3]Rutgers University, [4]University of Connecticut*

## Motivation

### ☑ Multi-Task Sequence Tasks

- Sequence Learning
  — E.g., sentiment classification, sequence labelling
- Multi-Task Learning
  — Advantages include computational advantage
- However, multi-task architectures applicable to recurrent models are underexplored

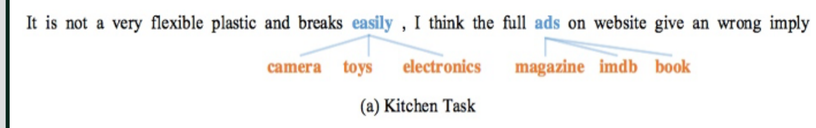
### ☑ Goal

- ***A Recurrent Module***
  — Can be integrated into any multi-task learning approach for sequential data
  — To improve model capacity, flexibility, generalization

### ☑ Challenges

- Dynamics of task relationships
- Limited knowledge of the task relatedness
- Generalization ability

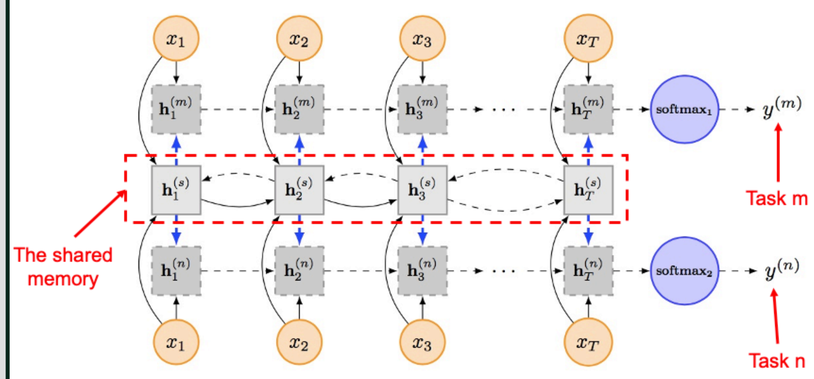
➡ Example: Dynamic task relatedness in NLP
— Sentiment prediction of 16 datasets (multi-task learning)




(a) Kitchen Task

The Illustrations of the three most relevant tasks for each word in the "Kitchen" task (AAAI'19, Liu et al.)

### ☑ Typical Architecture

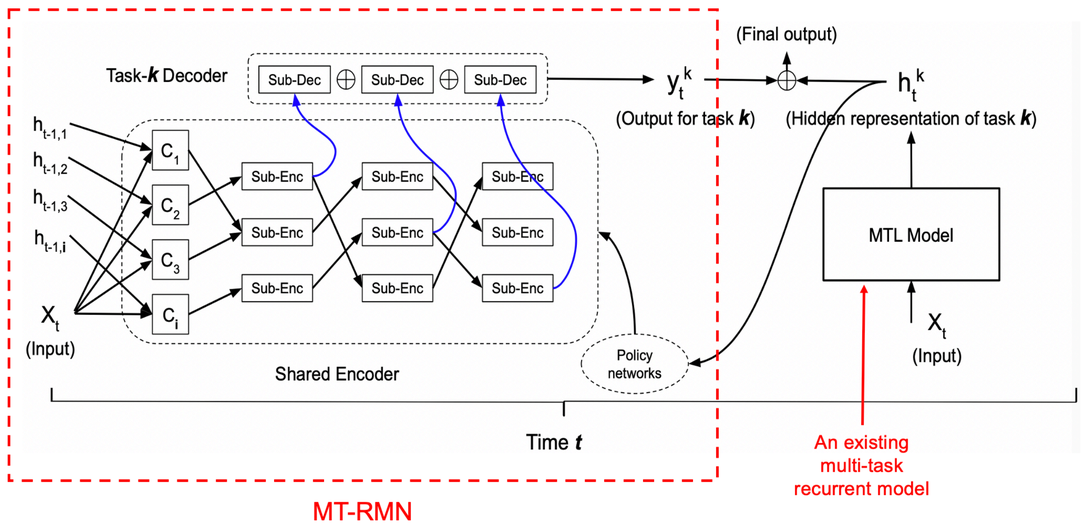



PSP-MTL (IJCAI'16)

However, existing approaches are not flexible enough to learn the dynamic relationship

## Proposed Model: MT-RMN

### ☑ Architecture

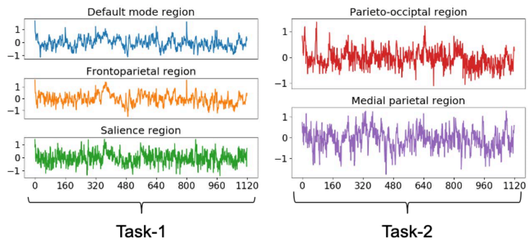



### ☑ How to Make Connection Decision

- Generate decision vector via policy network: $\boldsymbol{\beta}_{t,j}^k = \widehat{\mathcal{N}}^k(\boldsymbol{u}_{t,j}) = MLP(\boldsymbol{u}_{t,j} \oplus \boldsymbol{W}_k \boldsymbol{h}_t^k)$
- Estimate the binary decision value via Straight-Through Router: $\zeta = \arg\max_i [\alpha_i + g_i], i \in \{0, 1\}$

## Experiments

- **Experiments on Task-fMRI**
  — Task-Evoked Functional MRI Data




The task-fMRI data of twelve high-order brain regions of a participant. The regions are grouped into four groups based on their functionalities. Each group is used to construct a multi-class classification task of time series. The participant was asked to perform seven tasks successively.

Table 1: Results (accuracy %) on two groups of *tfMRI* tasks.

| Groups | Group-1 | | | Group-2 | | |
|---|---|---|---|---|---|---|
| | Task-1 | Task-2 | Task-3 | Task-1 | Task-3 | Task-4 |
| FS-MTL | 89.7±0.6 | 93.5±1.1 | 89.1±0.3 | 73.2±0.5 | 81.2±0.6 | 81.3±1.4 |
| SP-MTL | 88.7±1.4 | 94.0±1.9 | 90.7±2.3 | 73.7±1.5 | 78.0±1.5 | 81.2±2.1 |
| DC-MTL | 89.2±0.6 | 95.6±1.2 | 90.8±1.0 | 74.5±0.6 | 80.0±0.7 | 80.8±1.3 |
| IC-MTL | 89.6±0.4 | 95.7±1.1 | 91.3±1.7 | 74.5±0.4 | 81.4±1.1 | 81.8±2.6 |
| RRNs | 88.9±3.4 | 95.4±2.2 | 90.5±2.8 | 75.4±2.2 | 83.2±3.0 | 82.6±3.8 |
| *mtl*-RMN | **90.8±2.1** | **96.7±1.6** | **92.9±2.4** | **76.1±1.8** | **84.4±2.4** | **83.5±2.2** |

Table 2: Four test-time scenarios to evaluate generalization ability and the results (accuracy%).

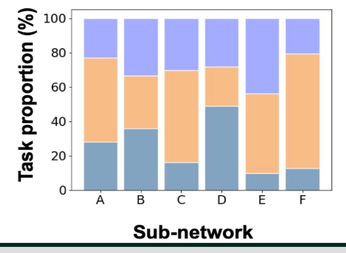
| Scenario settings | | | | | Results of different methods | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Scenarios | Training tasks | | | Test task | LSTM | $\text{RMN}_{un}$ | RRNs | IC-MTL | $\text{RMN}_{tr}$ |
| A | Task-1 | Task-2 | Task-3 | Task-4 | 82.6±1.0 | 82.9±1.2 | 83.0±2.7 | 81.2±1.7 | **83.3±2.3** |
| B | Task-1 | Task-2 | Task-4 | Task-3 | 82.3±0.6 | 84.5±1.5 | 84.8±2.5 | 82.7±1.3 | **85.6±1.7** |
| C | Task-1 | Task-3 | Task-4 | Task-2 | 93.7±0.6 | 95.3±0.8 | 95.6±1.7 | 95.5±0.7 | **97.0±1.8** |
| D | Task-2 | Task-3 | Task-4 | Task-1 | 71.7±1.4 | 73.1±1.7 | 78.4±4.0 | 77.3±2.1 | **79.3±2.6** |

- **Experiments on POS Tagging**

➡ **Task: POS Tagging of Code-Switched Sentences**
— POS Tagging: to mark up a word in a corpus to a corresponding speech tag
— Code-Switched Text: words from multiple languages

➡ **Data Sets**
— Task 1: A Hindi-English code-switch data POS tagging
— Task 2: Hindi POS tagging
— Task 3: English POS tagging

| Methods | Original | | DC-MTL | | IC-MTL | | *rrn*-RMN | | MT-RMN | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | $F_1$ | Accuracy | $F_1$ | Accuracy | $F_1$ | Accuracy | $F_1$ | Accuracy | $F_1$ |
| NS-MTL | 44.3±0.7 | 38.9±1.6 | 51.5±0.5 | 49.4±0.9 | 51.4±1.4 | 49.7±0.8 | 53.5±1.9 | 52.7±0.8 | **55.2±1.2** | **53.4±2.4** |
| Cross-stitch. | 46.0±1.9 | 12.7±0.8 | 48.5±0.9 | 16.1±1.3 | 47.7±1.7 | 14.3±0.7 | **52.1±1.0** | 19.5±0.6 | 51.7±1.9 | **21.4±3.2** |
| Sluice | 58.7±0.7 | 55.4±2.1 | 59.2±1.7 | 56.0±2.3 | 59.5±1.9 | 56.2±0.7 | 60.4±1.2 | 56.6±0.7 | **61.5±1.7** | **57.2±1.8** |
| GIRNet | 62.8±0.6 | 46.6±0.4 | 62.5±1.1 | 57.4±0.5 | 63.1±1.5 | 58.6±1.3 | 63.6±1.7 | 61.3±1.5 | **64.5±2.1** | **62.6±2.4** |




1) The proportion of tasks assigned to each sub-network
2) Different colors distinguish different tasks
3) A-F represent the six sub-networks

**Welcome!** Email: *dux19@psu.edu* Twitter: *DongkuanXu* WeChat: *xudongkuan220019*