Joint Deep Learning

Dong Wang

2015/03/09

Deep learning in various principles

- Speech recognition
- Speaker recognition
- Natural language processing
- Image processing

General of deep learning

- Feed forward, convolutional, recurrent
- Initialization and training
- Regularization
- Adaptation
- Feature extraction and modeling
- Advantages:
  - Learn in discriminative fashion
  - Learn hierarchical representations
  - Learn high non-linearity
  - Learn complex relationship
  - Learn multiple conditions
  - Learn temporal, spatial, spectral information (CNN)
  - Learn sequence (RNN)

Deep learning in speech recognition

- Basic
  - Tandem/hybrid
  - Supervised/unsupervised
  - Multiconditional training
  - Multilingual training
  - Adaptation
  - Noisy training
- Hot
  - Structure learning
  - Conditional training (rate, language,speaker,noise type, SNR,music,…)
  - Applications: music removal, Tone removal,VAD
  - Visualization

Deep learning in speaker recognition

- Existing approaches:
  - Statistics for i-vector
  - Frame-based d-vector
- Promising research
  - Feature learning
  - Convolutional learning
  - K-max sequence learning
  - RNN speaker embedding
  - Conditional learning: language, phone posterior

Deep learning in NLP

- Current
  - Word and phrase embedding
  - document embedding: paragraph vector, semantic hashing, DSSM,CLSM, RNN vector
  - Knowledge embedding
  - Neural LM, translation,POS,semantic labelling, NER,QA,…
- Hot
  - Large scale RNN LM
  - Topic modeling
  - Multiple resource inference(language, source,domain)

Joint deep learning

- Deep learning is possible to learn related tasks
- Project multimodalities to the same semantic space

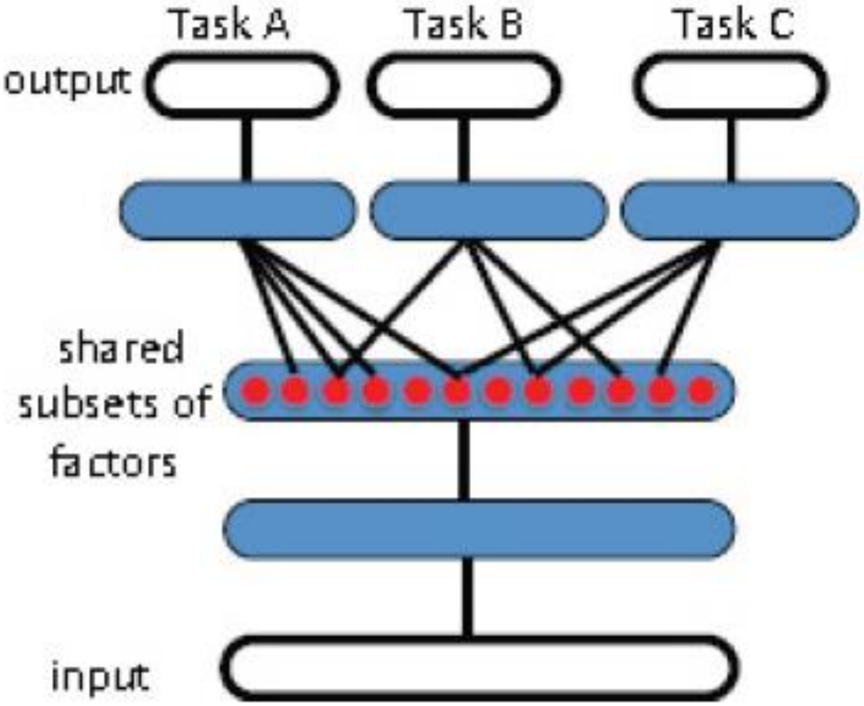


Image + text

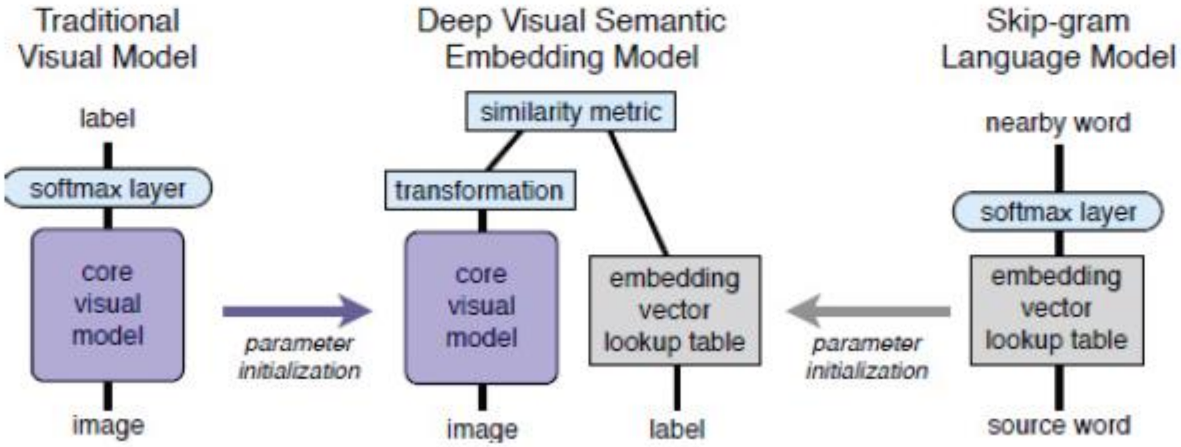
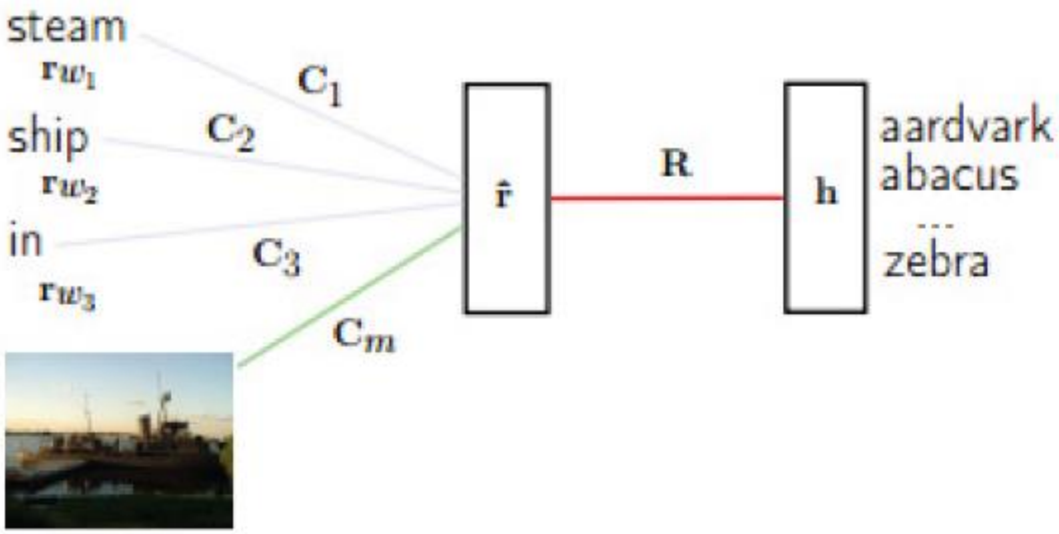
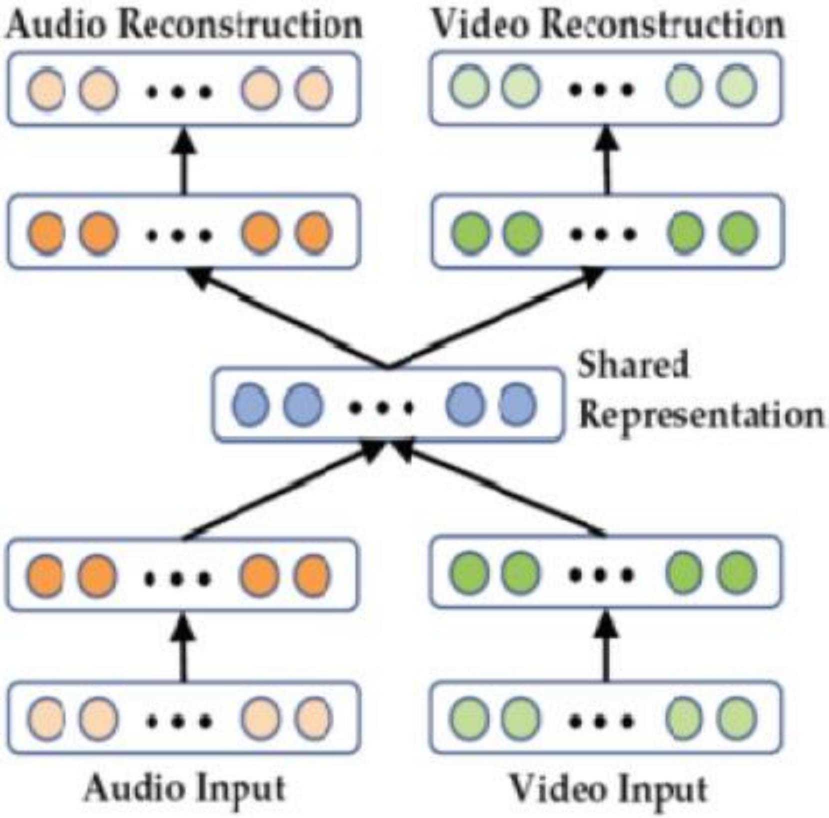

Traditional Visual Model
label
softmax layer
core visual model
image
parameter initialization
Deep Visual Semantic Embedding Model
similarity metric
transformation
core visual model
image
embedding vector lookup table
label
parameter initialization
Skip-gram Language Model
nearby word
softmax layer
embedding vector lookup table
source word

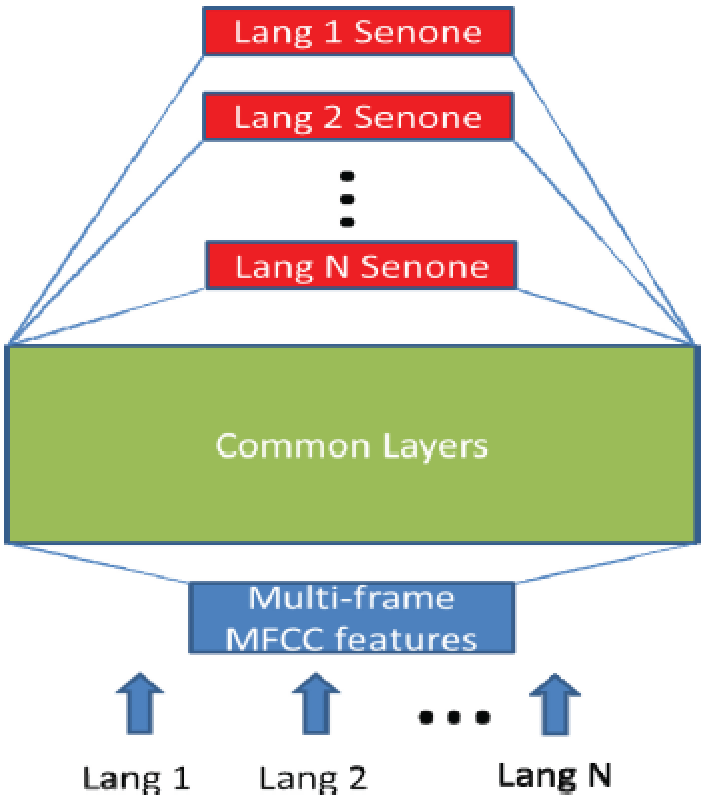
Image + text


steam
$\mathbf{r}_{w_1}$
$\mathbf{C}_1$
ship
$\mathbf{C}_2$
$\mathbf{r}_{w_2}$
in
$\mathbf{C}_3$
$\mathbf{r}_{w_3}$
$\mathbf{C}_m$
$\hat{\mathbf{r}}$
$\mathbf{R}$
$\mathbf{h}$
aardvark
abacus
...
zebra

Speech + video



Multilingual ASR



Basic ideas

- Multiple supervision
- Multiple resource
- Task relation

Speech + speaker?

- Is it possible to learn ASR and SID together?
  - Seems fine, they are related.
  - However difficult, as they do not share the same semantic space.
  - In fact they are in the opposite direction
- How about conditional training?
  - A partial supervision
  - An iterative solution

Multilingual LM?

- Multilingual text has been used to train translation model
- Is it possible to borrow multilingual resource to train LM?

Conclusions

- Deep learning need to be explored more thoroughly in ASR and SID.
- Conditional training is promising.
- Joint learning for speech and speaker recognition is possible.
- Learning multilingual LM is possible.