"One model to learn them all"

Paper review

Boreiko Valentyn

Department of Mathematics

Deep Learning Meetup, 2017

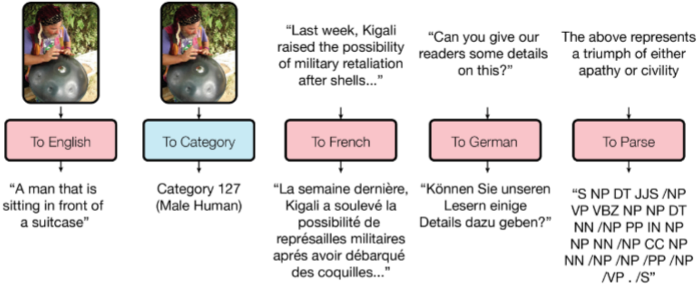
Multi modal Learning(MML)

Multimodal learning is a good model to represent the joint representations of different modalities. Features of MML are:

- Multiple modalities are learned together.
- Input is in one modality, otput is in another one.
- One modality complements another one during learning (transfer learning).

Can we learn unified deep learning model to solve tasks across multiple domains?

Yes! Solutions: learning joint image-word embeddings, as well as *embedding images and sentences into a common space*. Other approaches to multimodal learning include *deep Boltzmann machines*, *autoencoders*, *recurrent neural networks*, and others.



To English

"A man that is sitting in front of a suitcase"

To Category

Category 127 (Male Human)

"Last week, Kigali raised the possibility of military retaliation after shells..."

To French

"La semaine dernière, Kigali a soulevé la possibilité de représailles militaires aprés avoir débarqué des coquilles..."

"Can you give our readers some details on this?"

To German

"Können Sie unseren Lesern einige Details dazu geben?"

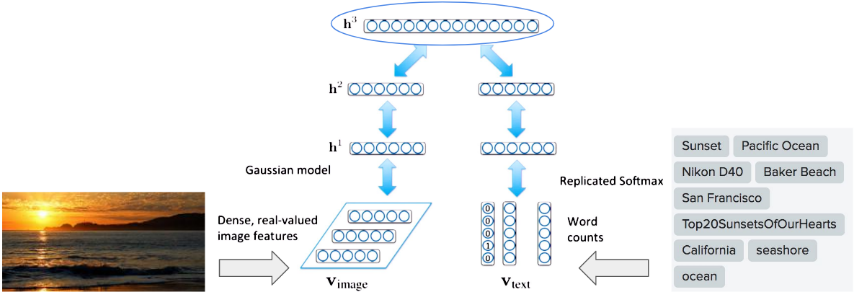
The above represents a triumph of either apathy or civility

To Parse

"S NP DT JJS /NP VP VBZ NP NP DT NN /NP PP IN NP NP NN /NP CC NP NN /NP /NP /PP /NP /VP . /S"

Some examples - Flickr, SoundNet

One possible application - Image captioning - Flickr.



Possile solution with Deep Boltzman Machine, which (if we talk about Restricted Boltzman Machine) is two layer *undirected, probabilistic* graphical model with *visible* **x** and *hidden* **h** units, with joint probability:

$$P_{W,\mathbf{b},\mathbf{c}}(\mathbf{x},\mathbf{h}) = \frac{e^{-E_{W,\mathbf{b},\mathbf{c}}(\mathbf{x},\mathbf{h})}}{\sum_{\mathbf{x}_j}\sum_{\mathbf{h}_j} e^{-E_{W,\mathbf{b},\mathbf{c}}(\mathbf{x}_j,\mathbf{h}_j)}}.$$

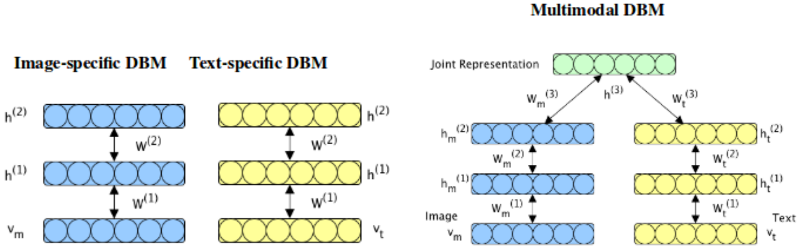
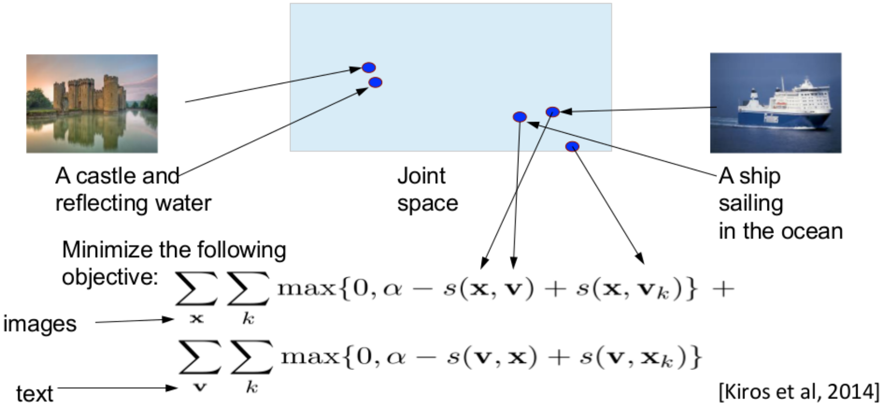


Figure 2: **Left:** Image-specific two-layer DBM that uses a Gaussian model to model the distribution over real-valued image features. **Middle:** Text-specific two-layer DBM that uses a Replicated Softmax model to model its distribution over the word count vectors. **Right:** A Multimodal DBM that models the joint distribution over image and text inputs.

Another possible solution via embedding of the images and sentences into a common space.

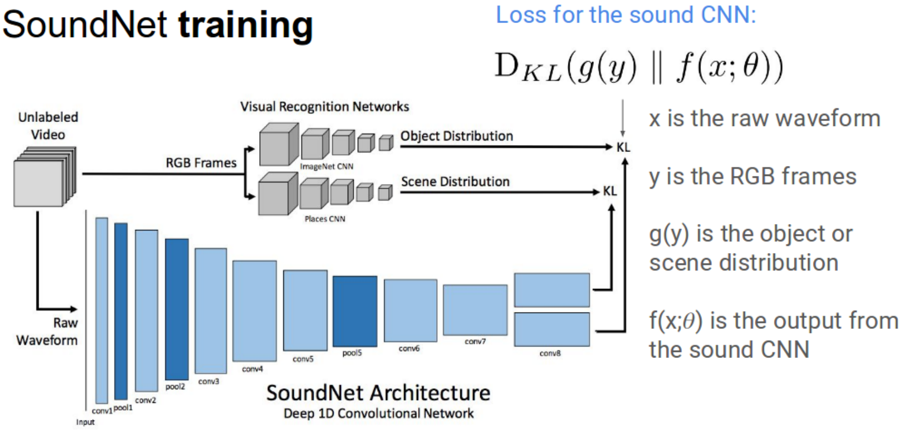


Minimize the following objective:

$$\sum_{\mathbf{x}} \sum_{k} \max\{0, \alpha - s(\mathbf{x}, \mathbf{v}) + s(\mathbf{x}, \mathbf{v}_k)\} + \sum_{\mathbf{v}} \sum_{k} \max\{0, \alpha - s(\mathbf{v}, \mathbf{x}) + s(\mathbf{v}, \mathbf{x}_k)\}$$

[Kiros et al, 2014]

Another usecase - learning sound representation from unlabeled video - SoundNet.



SoundNet **training**

Loss for the sound CNN:

$$\mathrm{D}_{KL}(g(y) \,\|\, f(x;\theta))$$

x is the raw waveform

y is the RGB frames

g(y) is the object or scene distribution

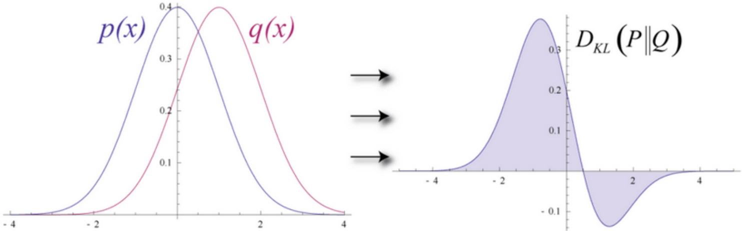
f(x;θ) is the output from the sound CNN

video-showcases can be found her - http://soundnet.csail.mit.edu/

Kullback Leibler Divergence

KL-Divergence measures the non-overlapping, or diverging, areas under the two curves.

$$\sum_i P(i) log(\frac{P(i)}{Q(i)}) - \textit{for discrete } P \textit{ and } Q$$



Model's task

The model is trained simultaneosly on these 8 corpora:
(1) WSJ speech corpus
(2) ImageNet dataset
(3) COCO image captioning dataset
(4) WSJ parsing dataset
(5) WMT English-German translation corpus
(6) The reverse of the above: German-English translation
(7) WMT English-French translation corpus
(8) The reverse of the above: German-French translation.

Model's architecture

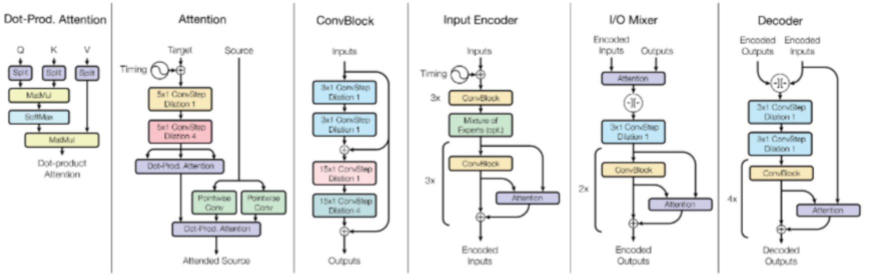
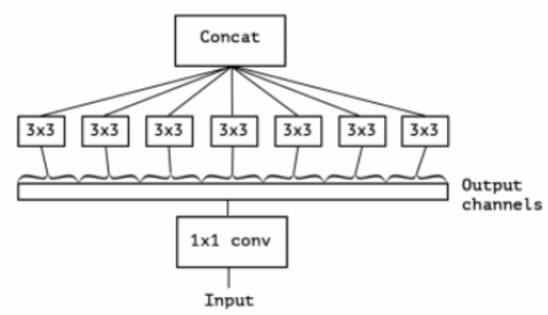


Figure 3: Architecture of the MultiModel; see text for details.

Convolution blocks with depthwise separable convolution

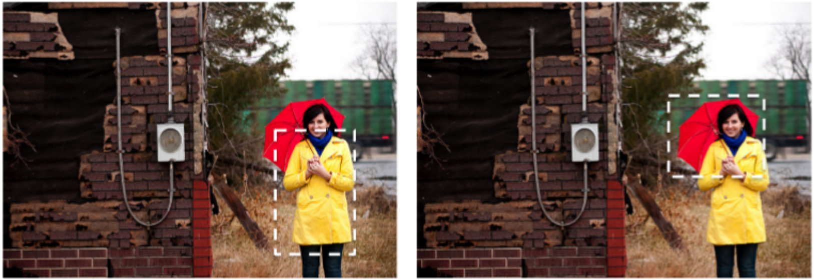
Special type of convolution (depthwise separable with batch normalization) is used, that saves computational time and makes the model more stable.

$$ConvStep_{d,s,f}(W,x) = LN(SepConv_{d,s,f}(W, ReLU(x))).$$



Though for Xception Net, Chollet uses, depthwise separable layers which perform 3x3 convolutions for each channel and then 1x1 convolutions on the output from 3x3 convolutions (opposite order of operations depicted in image above)

Attention



What is the color of the coat?

Traditional VQA: analyze the whole image -> analyze question -> give answer: ~~brown~~

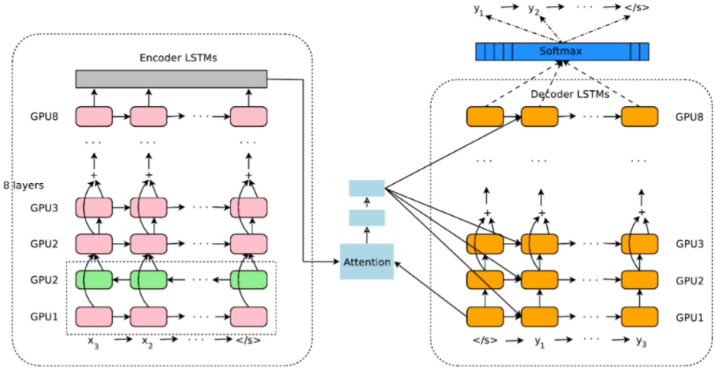
Attention based VQA: find coat -> judge the color of coat -> give answer: yellow

What is the color of the umbrella?

Traditional VQA: analyze the whole image -> analyze question -> give answer: ~~green~~

Attention based VQA: find umbrella -> judge the color of umbrella -> give answer: red

New output is a weighted sum over the encoded inputs.



$$s_t = AttentionFunction(\mathbf{y}_{i-1}, \mathbf{x}_t) \quad \forall t, \quad 1 \le t \le M$$

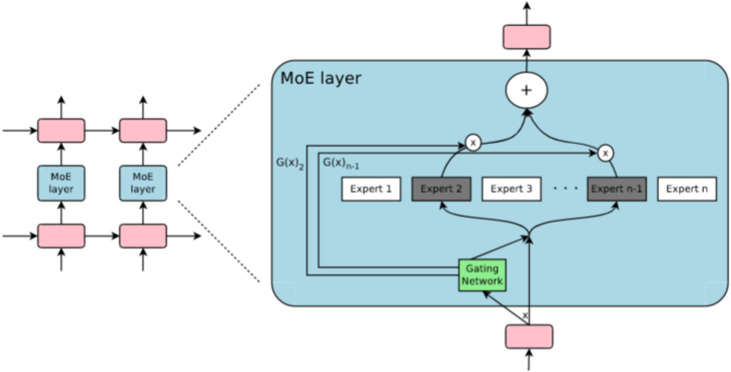
$$p_t = \exp(s_t) / \sum_{t=1}^{M} \exp(s_t) \qquad \forall t, \quad 1 \le t \le M$$

$$\mathbf{a}_i = \sum_{t=1}^{M} p_t . \mathbf{x}_t$$

Mixture-of-experts

Let $G(x)$ and $E_i(x), i \in 1, .., n$ be the output of the gating network and of the $i - th$ expert network resp. for a given input x. The otput y of the MoE module is then:

$$y = \sum_{i=1}^{n} G(x) E_i(x).$$



Modality blocks

They are basicly different embedings into a common space:

$$
\begin{aligned}
c1(x,F) &= ConvStep_{f=F}(W^{3\times 3}, x) \\
c2(x,F) &= ConvStep_{f=F}(W^{3\times 3}, c1(x,F)) \\
p1(x,F) &= MaxPool_2([3\times 3], c2(x,F)) \\
ConvRes(x,F) &= p1(x,F) + ConvStep_{s=2}(W^{1\times 1}, x),
\end{aligned}
$$

Image modality

$$
\begin{aligned}
h1(x) &= ConvStep_{s=2,f=32}(W^{3\times3}, x) \\
h2(x) &= ConvStep_{f=64}(W^{3\times3}, h1(x)) \\
r1(x) &= ConvRes(h2(x), 128) \\
r2(x) &= ConvRes(r1(x), 256) \\
ImageModality_{\mathrm{in}}(x) &= ConvRes(r2(x), d)
\end{aligned}
$$

Categorical and image modality

The same works for categorical and language modality:

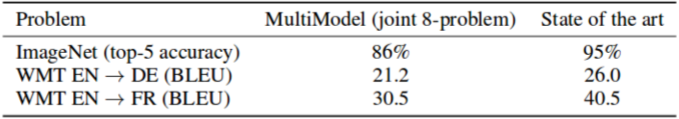
$$
\begin{aligned}
skip(x) &= ConvStep_{s=2}(W^{3\times 3}_{skip}, x) \\
h1(x) &= ConvStep(W^{3\times 3}_{h1}, x) \\
h2(x) &= ConvStep(W^{3\times 3}_{h2}, h1(x)) \\
h3(x) &= skip(x) + MaxPool_2([3 \times 3], h2(x)) \\
h4(x) &= ConvStep_{f=1536}(W^{3\times 3}_{h4}, h3(x)) \\
h5(x) &= ConvStep_{f=2048}(W^{3\times 3}, h4(x)) \\
h6(x) &= GlobalAvgPool(ReLU(h5(x)))
\end{aligned}
$$

$$CategoricalModality_{\text{out}}(x) = PointwiseConv(W^{classes}, h6(x))$$

$$
\begin{aligned}
LanguageModality_{\text{in}}(x, W_E) &= W_E \cdot x \\
LanguageModality_{\text{out}}(x, W_S) &= Softmax(W_S \cdot x)
\end{aligned}
$$

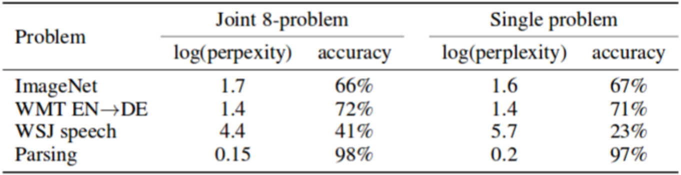
Results

And here are the results of the model.
They are not perfect, but this model is trained on 8 different tasks simultaneosly - what is already an praiseworthy achievement.



Problem	MultiModel (joint 8-problem)	State of the art
ImageNet (top-5 accuracy)	86%	95%
WMT EN → DE (BLEU)	21.2	26.0
WMT EN → FR (BLEU)	30.5	40.5

Table 1: Comparing MultiModel to state-of-the-art from [28] and [21].



Problem	Joint 8-problem		Single problem	
	log(perpexity)	accuracy	log(perplexity)	accuracy
ImageNet	1.7	66%	1.6	67%
WMT EN→DE	1.4	72%	1.4	71%
WSJ speech	4.4	41%	5.7	23%
Parsing	0.15	98%	0.2	97%

Table 2: Comparison of the MultiModel trained jointly on 8 tasks and separately on each task.

Summary

- Multi model learning can be solved with different approaches (embeddings into a common space), DBM, ...
- The Google's architecture combined encoder-decoder with special convolution, attention and mixture-of-experts blocks.
- Outlook
    - One can further learn more about transfer learning.
    - As well as one-shot learning and memory networks.