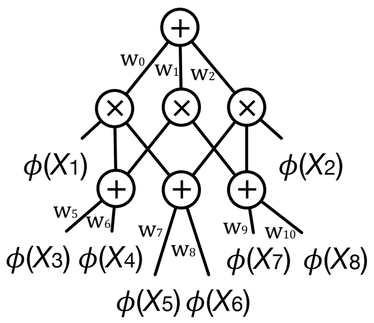
Unifying Sum-Product Networks and Submodular Fields

Abram L. Friesen and Pedro Domingos

{afriesen, pedrod}@cs.washington.edu

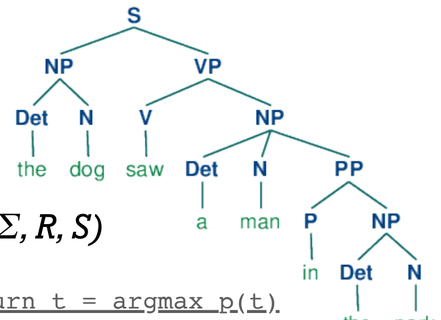
Paul G. Allen School of Computer Science and Engineering
University of Washington, Seattle WA, USA

Sum-Product Networks (SPNs)



- Deep, probabilistic
- Tractable inference
- Expressive

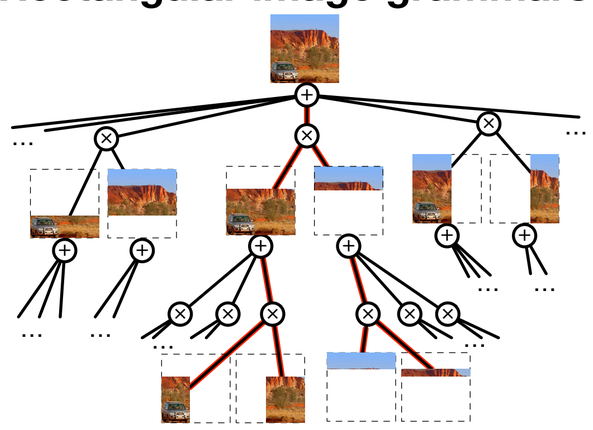
PCFGs



$G = (N, \Sigma, R, S)$

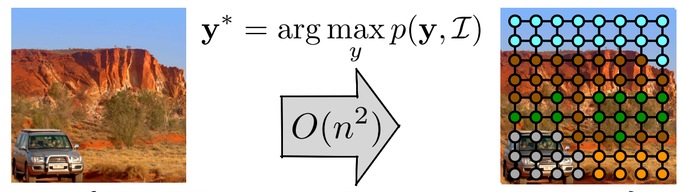
```
CYK: return t = argmax p(t)
for span length l = 1…n,
 for span start s = 1…n-l+1,
  for span split point p = 1…l-1,
   for each production v:X→YZ
    if Y[s,p] and Z[s+p,l-p] exist
     then store X[s,l]
return backtrace(S[1,n])
```

Rectangular image grammars



But intractable if we want *arbitrary region shapes*.

Submodular MRFs



$y \in \{\text{Sky}, \text{Mountain}, \text{Tree}, \text{Sand}, \text{Vehicle}\}$

$$p(\mathbf{y}, \mathcal{I}) \propto \exp(-E(\mathbf{y}, \mathcal{I}))$$

$$E(\mathbf{y}, \mathcal{I}) = \sum_{p \in \mathcal{I}} \theta_p(y_p) + \sum_{(p,q) \in \mathcal{I}} \theta_{pq}(y_p, y_q)$$

where potentials are *attractive*:

$$\theta_{pq}(0,0) + \theta_{pq}(1,1) \leq \theta_{pq}(0,1) + \theta_{pq}(1,0)$$

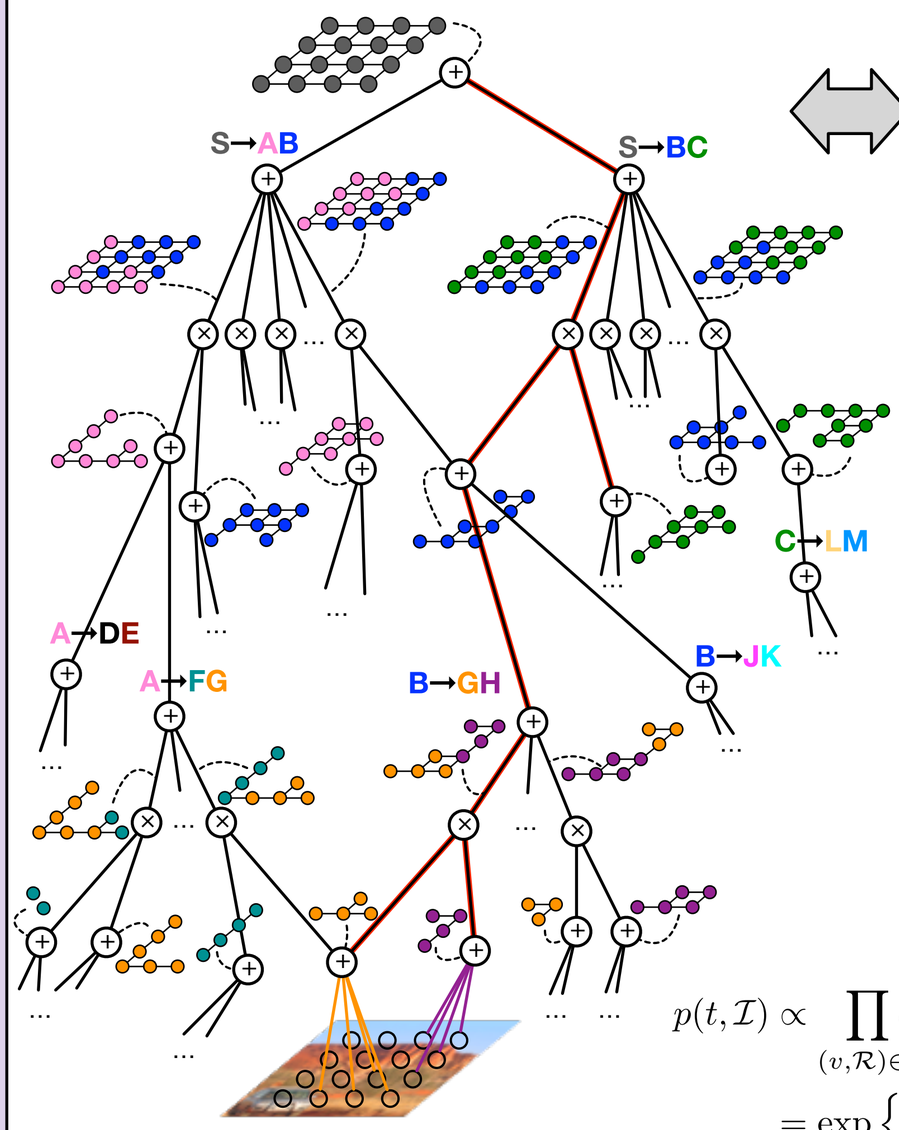
But label structure is very limited.

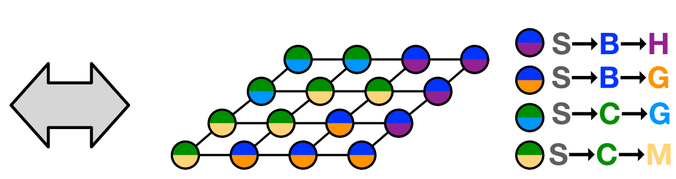
Submodular Sum-Product Networks (SSPNs)

Goal: the expressive label space of SPNs with the flexible spatial structure of MRFs.

SSPN: an SPN in which the edge weights are defined by submodular random fields.

Can think of as
a) an SPN with sum nodes with an exponential number of children, or
b) a submodular MRF with an exponential number of labels.





Key ideas:

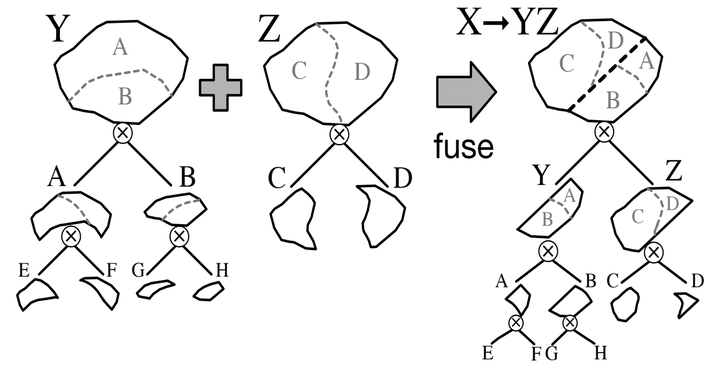
1. One MRF per production; use a different sub-MRF for each subregion.
2. Parse tree energy is sum of energies of production MRFs over their regions.
3. Conditioned on region, production, and all sub-parses, the choice of labeling is submodular.

$$p(t, \mathcal{I}) \propto \prod_{(v,\mathcal{R}) \in t} \phi(v|\text{head}(v)) \cdot \phi(\mathbf{y}^v, \mathcal{I}|v, \mathcal{R})$$

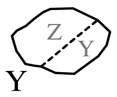
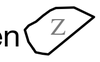
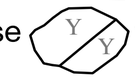
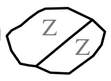
$$= \exp\left\{-\sum_{(v,\mathcal{R}) \in t} q_v + E_v(\mathbf{y}^v, \mathcal{I}, \mathcal{R})\right\}$$

$$= \exp\left\{-\sum_{(v,\mathcal{R}) \in t} q_v + \sum_{p \in \mathcal{R}} \theta_p^v(y_p^v) + \sum_{(p,q) \in \mathcal{R}} \theta_{pq}^v(y_p^v, y_q^v)\right\}$$

Convergent, approximate MAP inference



Fuse parse trees for Y and Z to get a parse for v:X→YZ with a single graph cut.

To improve parse of 
1. (re)parse  as Y
2. (re)parse  as Y given 
3. (re)parse  as Z
4. (re)parse  as Z given 
5. fuse  with 

Iteratively fuse upwards ⇒ a move-making algorithm.

```
InferSSPN: output t ≈ argmax p(t,I)
while not converged,
 for each symbol X in rev. topo. order,
  for each instance of X with region R,
   for each production v:X→YZ,
    get best parses of R as Y and Z
    fuse these to get a parse of R as v
    fuse these over I\R to get a parse of I as v
   set lowest energy v as parse of R as X
return best parse of I as S
```

Weight learning via hard gradient descent

$$\frac{\partial}{\partial w_i} \log p_{\mathbf{w}}(\mathbf{y}|\mathcal{I}) \approx n_i(t^*_{\mathbf{y}|\mathcal{I}}) - n_i(t^*_{\mathcal{I}}),$$

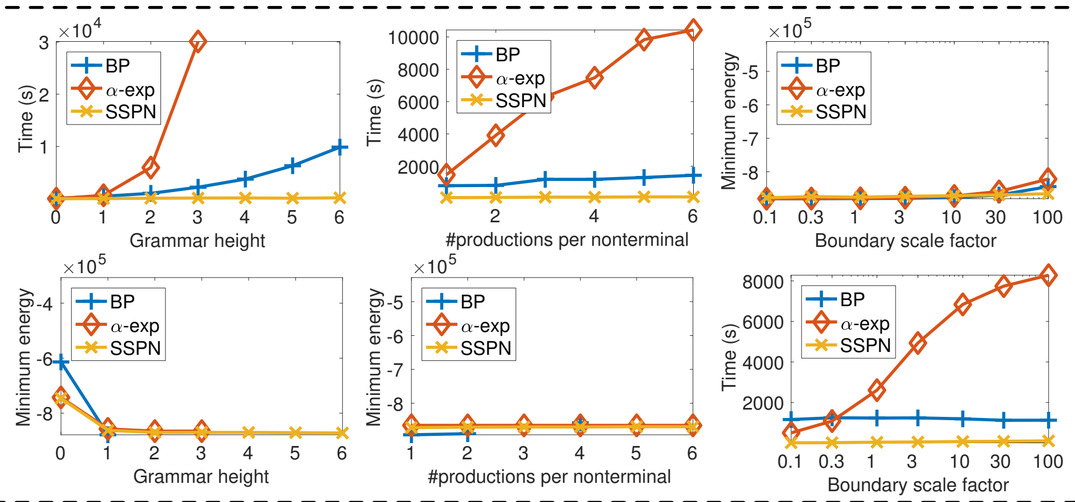
$$\text{where } t^*_{\cdot} = \arg\min_{t \in \mathcal{T}_{\mathbf{w}}(\cdot)} E(t, \mathcal{I})$$

Analytical and Empirical Results

Inference

Programmatically generated grammars.

α-expansion is within a constant factor of global minimum.



Theorem 1 *InferSSPN converges to a local minimum in a finite number of iterations, where each iteration has time complexity* $O(|G|c(n)n) \approx O(|G|n)$.

Model evaluation

Induced grammars by over-segmenting subsets of SBD.

Mean pixel accuracy on SBD test	
DeepLab	87.46
DeepLab + MRF	87.60
SSPN (+ oracle)	**90.03**