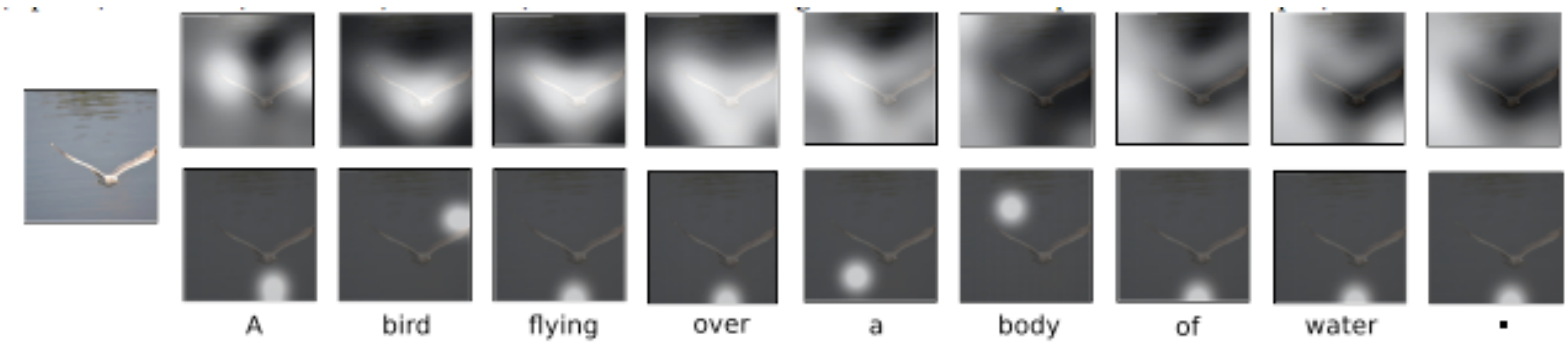
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio

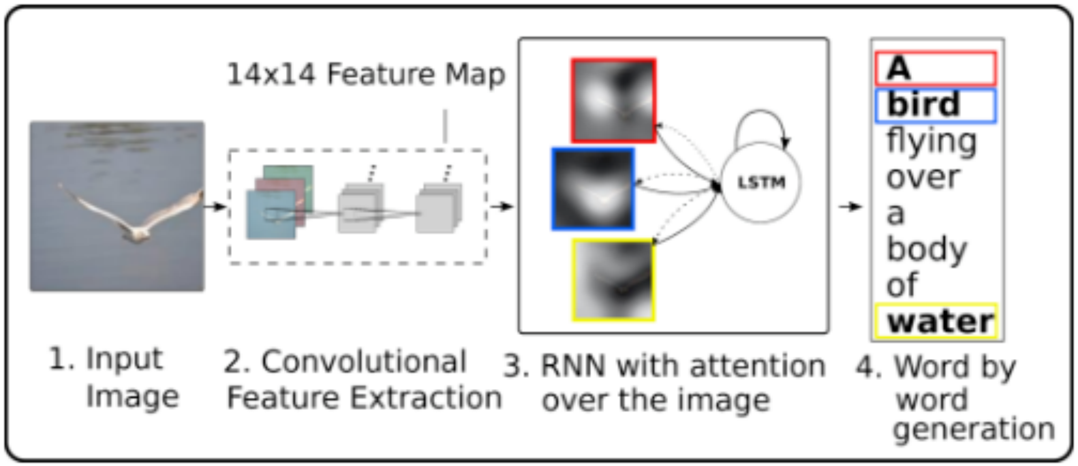
Presented by Kathy Ge

Motivation: Attention



- "attention allows for salient features to dynamically come to the forefront as needed"

Image Caption Generation with Attention Mechanism

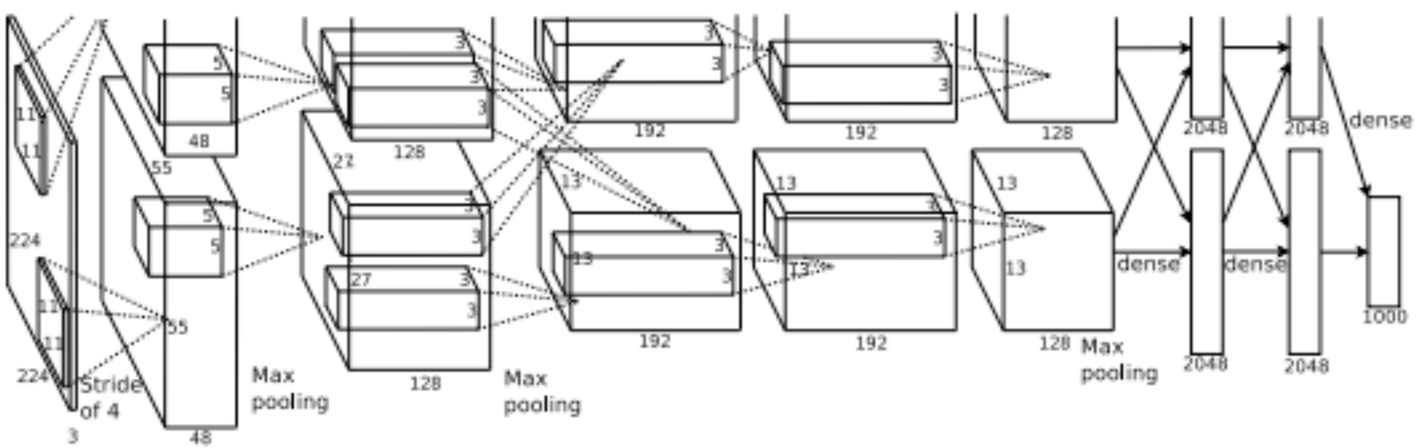


- Encoder: lower convolutional layer of a CNN
- Decoder: LSTM which generates a caption one word at a time
- Attention mechanism
  - Deterministic "soft" mechanism
  - Stochastic "hard" mechanism
- Output:

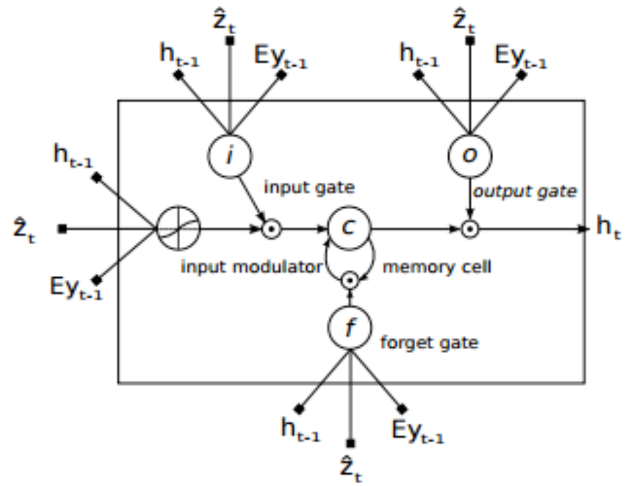
$$y = \{\mathbf{y}_1, \ldots, \mathbf{y}_C\}, \ \mathbf{y}_i \in \mathbb{R}^K$$

Encoder: CNN

- Lower convolutional layer of a CNN is used, to capture spatial information encoded in images
- annotation vector $a = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\},\ \mathbf{a}_i \in \mathbb{R}^D$



Decoder: LSTM



$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{D+m+n,n} \begin{pmatrix} \mathbf{E}\mathbf{y}_{t-1} \\ \mathbf{h}_{t-1} \\ \hat{\mathbf{z}}_t \end{pmatrix}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t).$$

- where i_t, f_t, c_t, o_t, h_t are the input, forget, memory, output, and hidden state of the LSTM at time t
- $\hat{\mathbf{z}} \in \mathbb{R}^D$ is the context vector which captures the visual information associated with a particular input location
- $\mathbf{E} \in \mathbb{R}^{m \times K}$ is the embedding matrix

Learning Stochastic "Hard" vs Deterministic "Soft" Attention

- Given an annotation vector $\mathbf{a}_i$, $i = 1, \cdots, L$ for each location i, an attention mechanism generates a positive weight α_i
- Weight of each annotation vector is computed by an attention model f_{att} using a multi-layer perceptron conditioned on previous hidden states h_{t-1}

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}.$$

- Define a function ϕ which computes the context vector z_t given the annotation vectors and corresponding weights

$$\hat{\mathbf{z}}_t = \phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right)$$

- Given the previous word, previous hidden state and context vector, compute output word probability

$$p(\mathbf{y}_t|\mathbf{a}, \mathbf{y}_1^{t-1}) \propto \exp(\mathbf{L}_o(\mathbf{E}\mathbf{y}_{t-1} + \mathbf{L}_h\mathbf{h}_t + \mathbf{L}_z\hat{\mathbf{z}}_t))$$

Deterministic "Soft" Attention

- Compute expectation of context vector directly

$$\mathbb{E}_{p(s_t|a)}[\hat{\mathbf{z}}_t] = \sum_{i=1}^{L} \alpha_{t,i}\mathbf{a}_i$$

- Then can compute a soft attention weighted annotation vector

$$\phi\left(\{\mathbf{a}_i\},\{\alpha_i\}\right) = \sum_i^L \alpha_i \mathbf{a}_i$$

- This model is smooth and differentiable, can be computed using standard backpropagation

Doubly Stochastic Attention

- When training the deterministic version of the model, can introduce a doubly stochastic regularization, where

$$\sum_t \alpha_{ti} \approx 1.$$

- This encourages model to pay equal attention to every part of the image throughout the caption generation
- In experiments, improved overall BLEU score, and lead to more rich and descriptive captions
- The model is trained by minimizing the negative log likelihood with penalty

$$L_d = -\log(P(\mathbf{y}|\mathbf{x})) + \lambda \sum_i^L (1 - \sum_t^C \alpha_{ti})^2$$

Stochastic “Hard” Attention

- Let s_t represent the random variable corresponding to the location where the model decides to focus attention at the t^{th} word

$$p(s_{t,i} = 1 \mid s_{j<t}, \mathbf{a}) = \alpha_{t,i}$$
$$\hat{\mathbf{z}}_t = \sum_i s_{t,i}\mathbf{a}_i.$$

- where z_t is a random variable, and s_t are intermediate latent variables

Stochastic "Hard" Attention

- Define objection function, L_S, the variational lower bound

$$L_s = \sum_s p(s|\mathbf{a}) \log p(\mathbf{y}|s, \mathbf{a}) \leq \log \sum_s p(s|\mathbf{a}) p(\mathbf{y}|s, \mathbf{a}) = \log p(\mathbf{y}|\mathbf{a}) \qquad (1)$$

- Gradient w.r.t. parameters of model, W

$$\begin{aligned} \frac{\partial L_s}{\partial W} &= \sum_s p(s|\mathbf{a}) \left[\frac{\partial \log p(\mathbf{y}|s, \mathbf{a})}{\partial W} + \log p(\mathbf{y}|s, \mathbf{a}) \frac{\partial \log p(s|\mathbf{a})}{\partial W} \right] \\ &\approx \frac{1}{N} \sum_{n=1}^{N} \left[\frac{\partial \log p(\mathbf{y}|\tilde{s}^n, \mathbf{a})}{\partial W} + logp(\mathbf{y}|\tilde{s}^n, \mathbf{a}) \frac{\partial \log p(\tilde{s}^n|\mathbf{a})}{\partial W} \right] \end{aligned} \qquad (2)$$

where $\quad \tilde{s}_t \sim \mathbf{Multinoulli}_L(\{\alpha_i\})$

Stochastic "Hard" Attention

- Reduce estimator variance by using a moving average baseline and introducing entropy term H[s]
- Final learning rule: gradient w.r.t. parameters of model, W

$$\frac{\partial L_s}{\partial W} = \sum_s p(s|\mathbf{a}) \left[\frac{\partial \log p(\mathbf{y}|s, \mathbf{a})}{\partial W} + \log p(\mathbf{y}|s, \mathbf{a}) \frac{\partial \log p(s|\mathbf{a})}{\partial W} \right]$$
$$\approx \frac{1}{N} \sum_{n=1}^{N} \left[\frac{\partial \log p(\mathbf{y}|\tilde{s}^n, \mathbf{a})}{\partial W} + \lambda_r (\log p(\mathbf{y}|\tilde{s}^n, \mathbf{a}) - b) \frac{\partial \log p(\tilde{s}^n|\mathbf{a})}{\partial W} + \lambda_e \frac{\partial H[\tilde{s}^n]}{\partial W} \right] \tag{3}$$

where λ_r, λ_e are hyperparameters, and b is exponential decay used in calculating moving average baseline

- At each point, $\phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right)$ returns a sampled a_i at every point in time based on a multinomial distribution parametrized by α

- Similar to REINFORCE rule

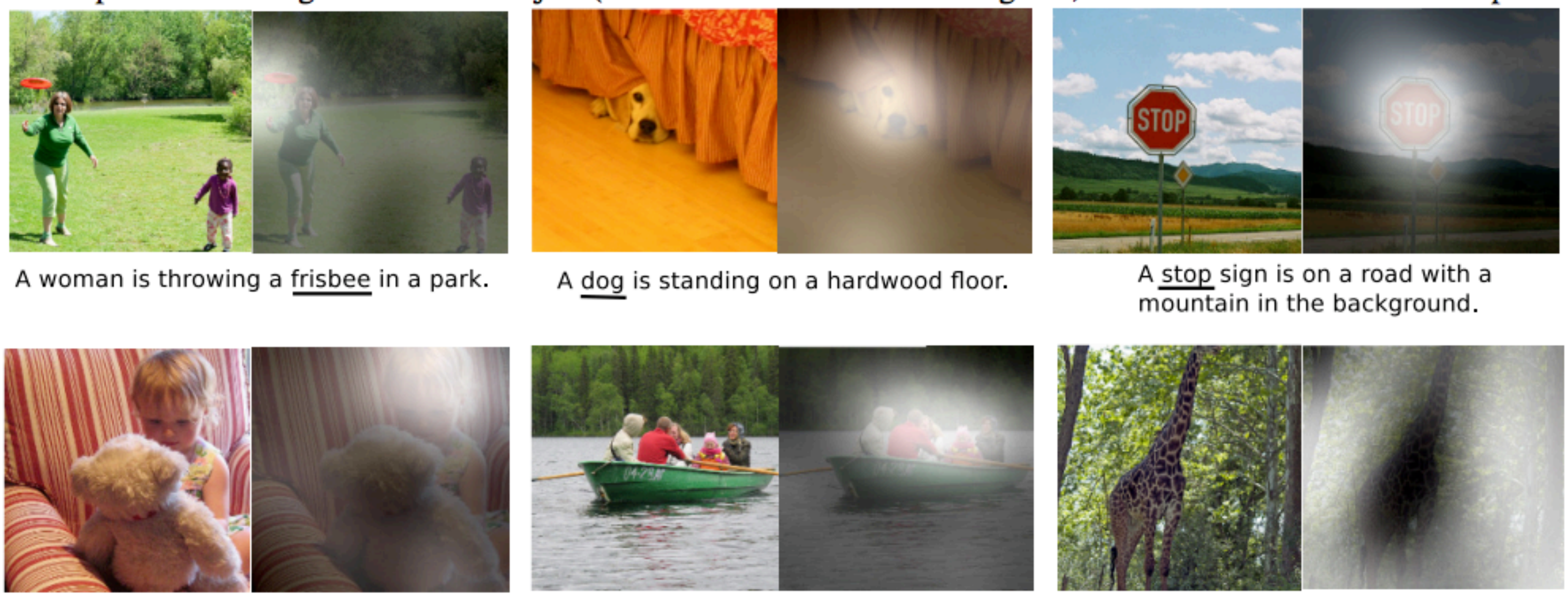
Experiments

- Evaluated performance on Flickr8K, Flickr30K, and MS COCO
- Optimized using RMSProp for Flickr8K and Adam for Flickr30K/MS COCO
- Used Oxford VGGnet pretrained on ImageNet
- Quantitative results measured using BLEU and METEOR metrics

		BLEU				
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

Qualitative Results

Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)



A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Mistakes

Figure 5. Examples of mistakes where we can use attention to gain intuition into what the model saw.

A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

A person is standing on a beach with a surfboard.

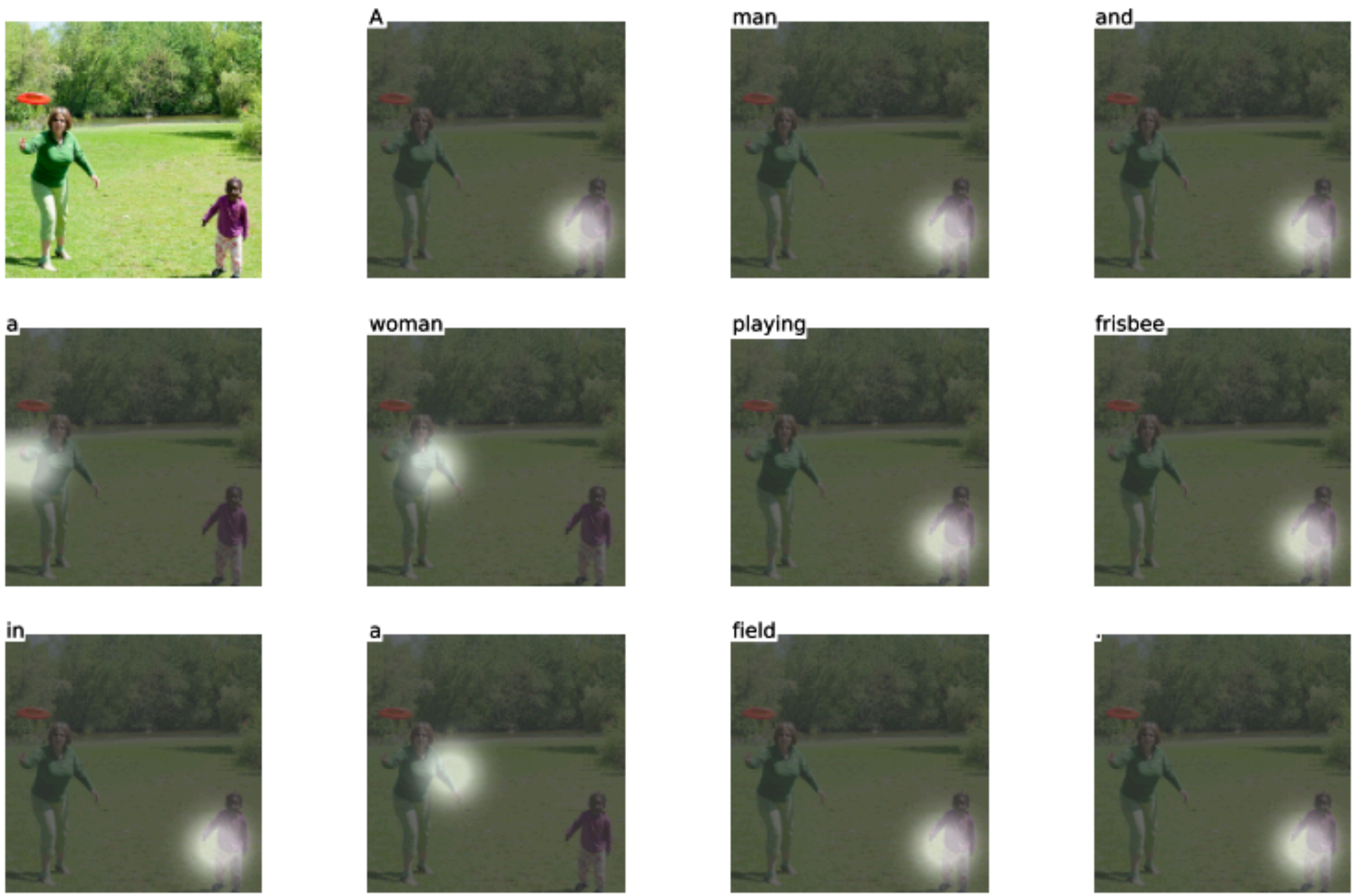
A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.

"Soft" attention model

A woman is throwing a frisbee in a park.

“Hard” attention model



A man and a woman playing frisbee in a field.

“Soft” attention model

A woman holding a clock in her hand.

“Hard” attention model

A woman is holding a donut in his hand.

Conclusion

- Xu et al. introduce an attention based model that is able describe the contents of an image
- The model is able to fix its gaze on salient objects while generating words in the caption sequence
- They compare the use of a stochastic “hard” attention mechanism by maximizing a variational lower bound and a deterministic “soft” attention mechanism using standard backpropagation
- Learned attention model can give interpretability to model generation process, and through qualitative analysis can show that alignments of words to locations in an image correspond well to human intuition

Thanks!
Any questions?