

Multimodal Word Discovery with Image Regions and Phone Sequences

ECE 590 SIP
Liming Wang
March 4, 2020

Outline

- Motivation
- Prior Works
- Problem Formulation
- Model Description
- Experiments
- Image-to-Speech Discovery (in-progress work)
- Conclusion

Motivation

- Most practical speech technologies nowadays are **supervised** ASR systems
- **Majority** of the spoken languages in the world lack the resources to train state-of-the-art ASR
- Unsupervised word discovery is challenging (Kamper et. al. 2016, 2017, Ondel et. al. 2018, 2019)
- Existing multimodal word discovery systems are data-hungry (Harwath et. al. 2018, Azuh et. al. 2019)

Statistical Noun Acquisition

- Connection between low-resource word discovery and first language noun acquisition



Image concepts: canoe jacket **male_child** man

Caption:

AH0 **M**AE1 N AH0 N D AH0 **L** IH1 **T** AH0 **L**
B OY1 IH0 N B L UW1 L AY1 F JH **AE1** K
IH0 T S AA1 R R OW1 IH0 NG AH0 Y EH1 L
OW0 K AH0 N UW1



Image concepts: hands **male_child** mouth

Caption:

AH0 **S** M AO1 **L** **B** OY1 P AH1 T IH0 NG S
AH1 M TH IH0 NG IH0 N HH IH1 Z **M** AW1
TH W IH1 DH B OW1 TH HH **AE1** N D Z



Image concepts: boat ledge **male_child**
new_jersey

Caption:

AH0 Y AH1 NG **B** OY1 W EH1 R IH0 NG AH0
JH ER1 Z IY0 L
UH1 K S D AW1 N OW1 V ER0 AH0 L EH1 JH
AE1 T **M** EH1 N
IY0 B OW1 T S IH0 N DH AH0 W AO1 T ER0

male_child:

1. 'AH0 S M AO1 L B OY1'
 2. 'L IH1 T AH0 L B OY1'
 3. 'AH0 Y AH1 NG B OY1'
- ...

Outline

- Motivation
- Prior Works
- Problem Formulation
- Model Description
- Experiments
- Image-to-Speech Discovery (in-progress work)
- Conclusion

Image-to-Phone Word Discovery

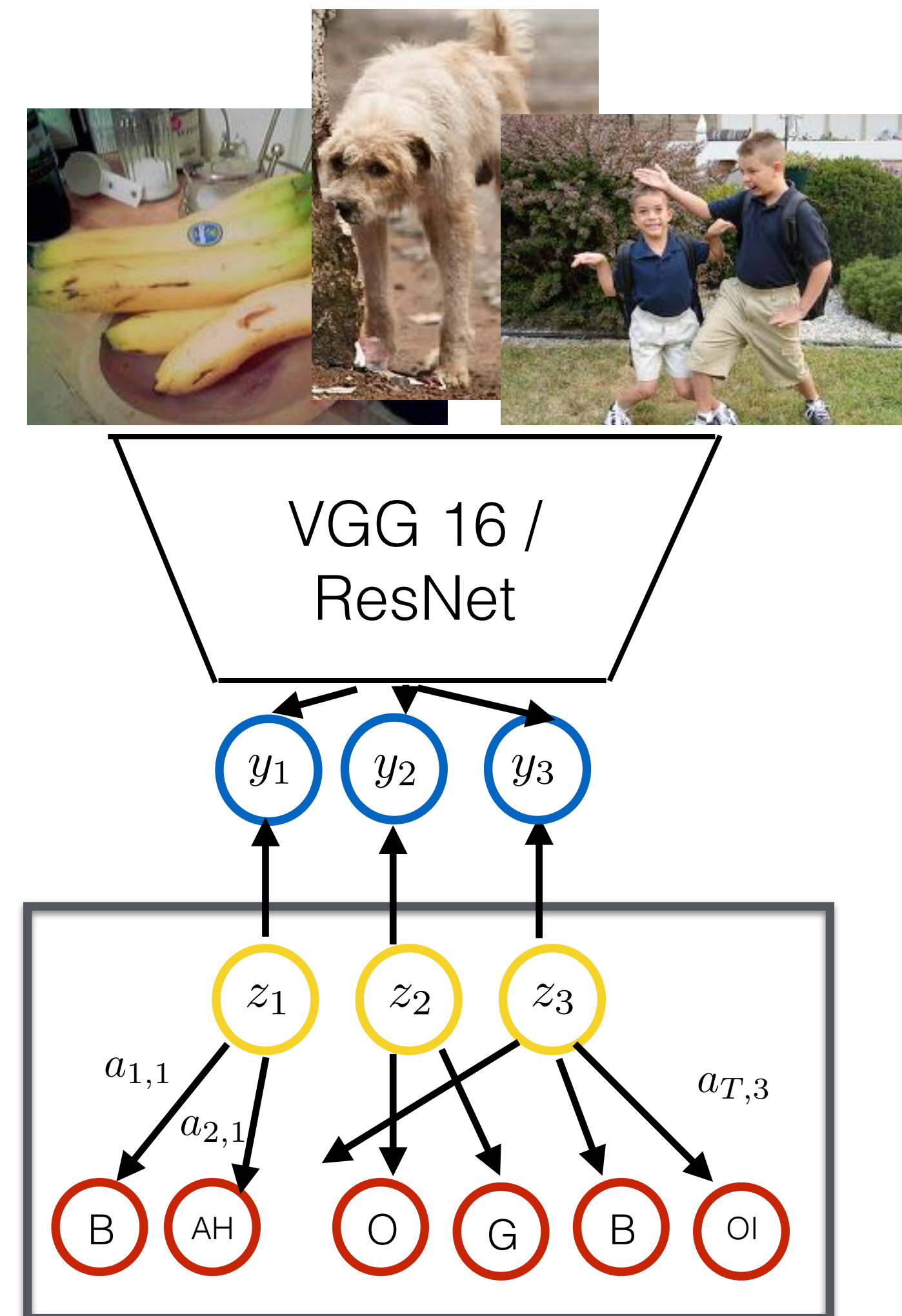
- Inputs: phone labels $x_1 x_2 \dots x_T$ and image features for each region in the scene $y_1 y_2 \dots y_N$
- Outputs: concept clusters assignment of the regions $\mathbf{z} = z_1 z_2 \dots z_N$ and the alignment between phones and regions $A \in \{0, 1\}^{T \times N}, i_t := \arg \max_i a_{ti}$

$$p(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{z}, \mathbf{A}} p(\mathbf{z}) p(\mathbf{A}) p(\mathbf{x} | \mathbf{z}, \mathbf{A}) p(\mathbf{y} | \mathbf{z})$$

Concept Probability Alignment Probability Translation Probability Visual Observation Probability

Correspondence Assumptions

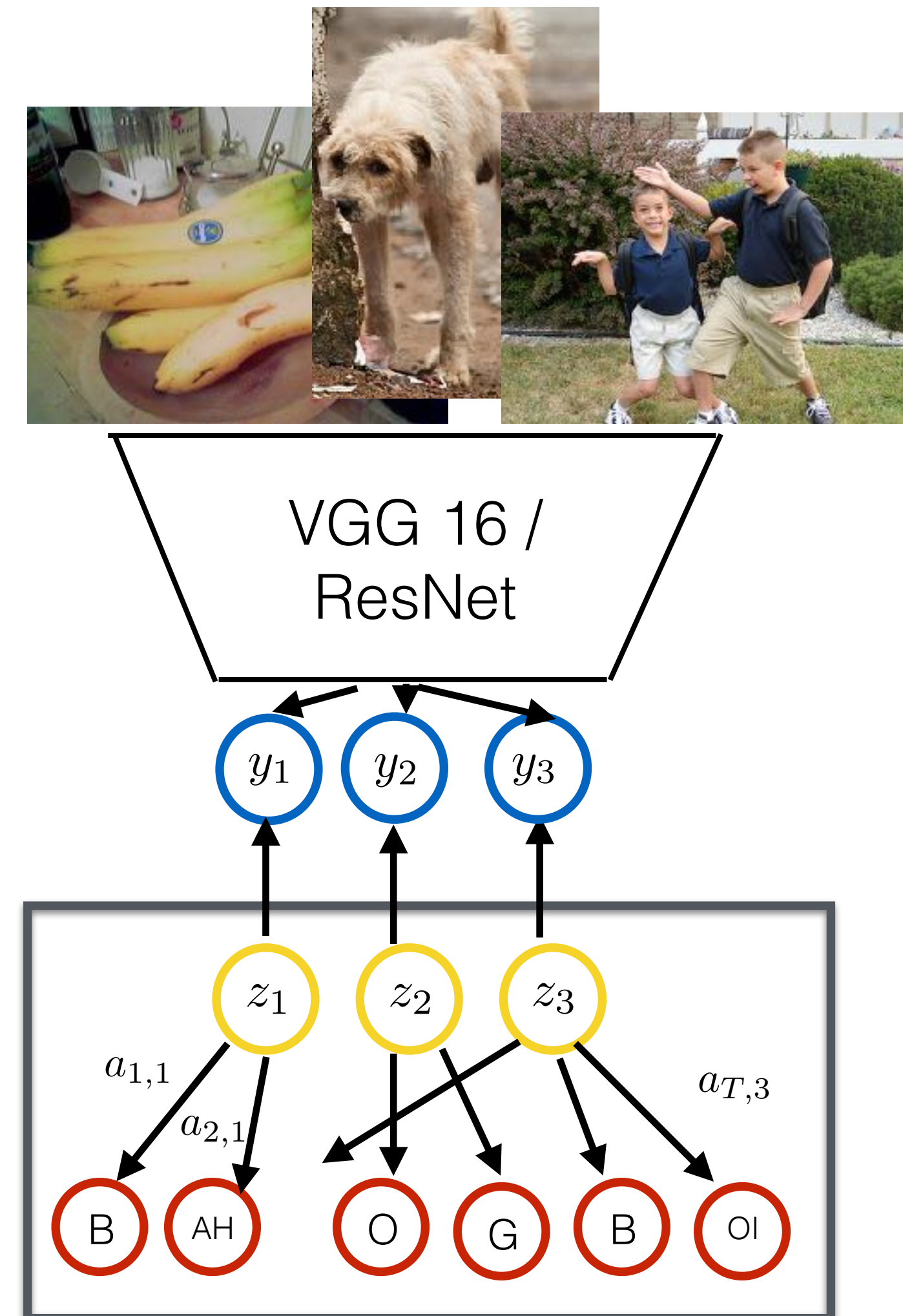
1. One region per phone, at least one phone per region (many-to-one correspondence)
2. One concept per region, but allow repeating concepts
3. Not realistic, but makes the math easier



Independence Assumptions

Given the distribution parameters:

1. Independent concepts $z_1, z_2 \dots z_N$
2. Markov alignment variables $i_1, i_2 \dots i_T$:
3. Conditionally independent image features y given z
4. Conditionally independent phone symbols x given z, A



Main Challenges

- **Potentially intractable likelihood and posteriors**

$$p(\mathbf{x}, \mathbf{y}, \Theta) = p(\Theta) \sum_z \prod_{i=1}^N p(z_i) \sum_{A \in \{0,1\}^{T \times N}} p(y_i | z_i) \prod_{t=1}^T p(i_t | i_{t-1}) p(x_t | z_{i_t})$$

K^N terms (intractable when the length of the image sequence is large)

- **Real value features:** infeasible to directly collect counts

Outline

- Motivation
- Prior Works
- Problem Formulation
- Model Description
- Experiments
- Image-to-Speech Discovery (in-progress work)
- Conclusion

Model 1: Approximate Posteriors with Variational Inference (VI)

- Based on the Correspondence LDA model proposed by Blei & Jordan (2003)

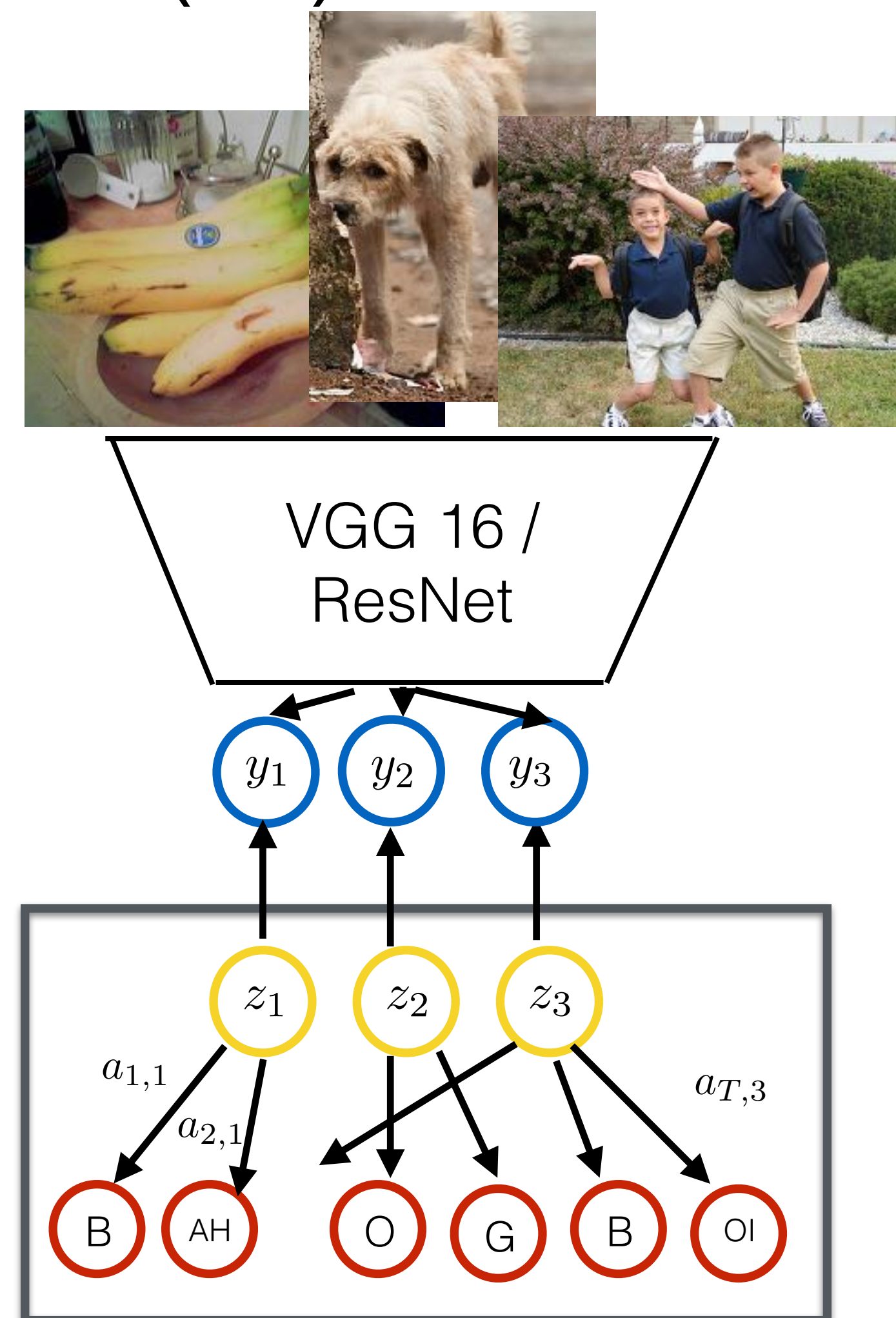
- Evidence Lower Bound (ELBO):

$$\log p(\mathbf{x}, \mathbf{y}, \Theta) \geq \mathbb{E}_{q(\mathbf{z}, \mathbf{A}, \Theta)} [\log p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{A}, \Theta)] - D(q(\mathbf{z}, \mathbf{A}, \Theta) || p(\mathbf{z}, \mathbf{A}, \Theta))$$

$$=: \text{ELBO}(q)$$

- Approximate posterior:

$$q^*(\mathbf{z}, \mathbf{A}, \Theta) = \arg \max_{q(\mathbf{z}, \mathbf{A}, \Theta)} \text{ELBO}(q(\mathbf{z}, \mathbf{A}, \Theta))$$



VI Assumptions

- **Parameter Prior:**

$$p(\sigma_{z_i, dd}^{-1} | \alpha_0, \beta_0) = \text{Gamma}(\sigma_{z_i, dd}^{-1} | \alpha_0, \beta_0)$$

$$p(\mu_{z_i} | \mu_0, \Sigma_{z_i}, \kappa_0) = \mathcal{N}(\mu_{z_i} | \mu_0, (\kappa_0)^{-1} \Sigma_{z_i})$$

$$p(p(i_t = \cdot | i_{t-1}) = q | \gamma_{i_{t-1}0}^a) = \text{Dir}(q | \gamma_{i_{t-1}0}^a)$$

$$p(v_k = v | \gamma_{k0}, \gamma_{k1}) = \text{Beta}(v | \gamma_{k0}, \gamma_{k1})$$

$$p(z_i = k | \mathbf{v}) = v_k \prod_{j < k} (1 - v_j)$$

- **Conditional Gaussian distribution:**

$$p(y_i | z_i) = \mathcal{N}(y_i | \mu_{z_i}, \Sigma_{z_i})$$

- **Mean-field Assumption:** Allow nice decomposition of ELBO

$$q(\mathbf{z}, \mathbf{A}, \Theta) = q(\Theta) \prod_{t=1}^T q(i_t | i_{t-1}) \prod_{i=1}^N q(z_i)$$

Coordinate Ascent Update for Approximate Posteriors

Visual observation term

$$q(z_i = k) = \frac{\mathcal{N}(y_i | \bar{\mu}_k, \bar{\Sigma}_k)}{Z_1(\mathbf{x}, \mathbf{y})} \exp \left(\psi(\gamma_{k0}) + \sum_{j < k} \psi(\gamma_{j1}) - \sum_{j \leq k} \psi \left(\sum_{m=0}^1 \gamma_{jm} \right) \right) \exp \left(\sum_{t=1}^T q(i(t) = i) \log p(x_t | z_i = k) \right)$$

Normalization factor

Prior terms

Audio observation terms dependent on $q(i(t)=i)$ from previous iteration

$$q(i(t) = i, i(t-1) = j) = \frac{q(i(t-1) = j) \left(\exp(\psi(\gamma_{ji}^a)) - \psi \left(\sum_{j=1}^L \gamma_{ji}^a \right) \right)}{Z_2(\mathbf{x}, \mathbf{y})} \exp \left(\sum_{k=1}^K q(z_i = k) \log p(x_t | z_i = k) \right),$$

Recursive term

Normalization factor

Audio observation terms dependent on $q(z_i = k)$

Parameter Update of VI

$$\bar{\mu}_z = \frac{\mu_0 + \kappa_0 \sum_{i=1}^L q(z_i) y_i}{1 + \kappa_0 \sum_{i=1}^L q(z_i)} \quad (19)$$

$$\bar{\sigma}_{z,dd}^* = \frac{\beta_0 + \kappa_0 \sum_{i=1}^L q(z_i) (y_{id} - \mu_{z,d}^*)^2}{\alpha_0 + \kappa_0 \sum_{i=1}^L q(z_i)} \quad (20)$$

$$\bar{p}(i_t = j | i_{t-1} = i) \quad (21)$$

$$= \frac{\gamma_{i0}^a + \sum_{t=2}^T q(i_{t-1} = i, i_t = j)}{\sum_{i=1}^L \gamma_{i0}^a + \sum_{t=2}^T q(i_{t-1} = i, i_t = j)} \quad (22)$$

$$\bar{p}(x_t = x | z_i = k) \quad (23)$$

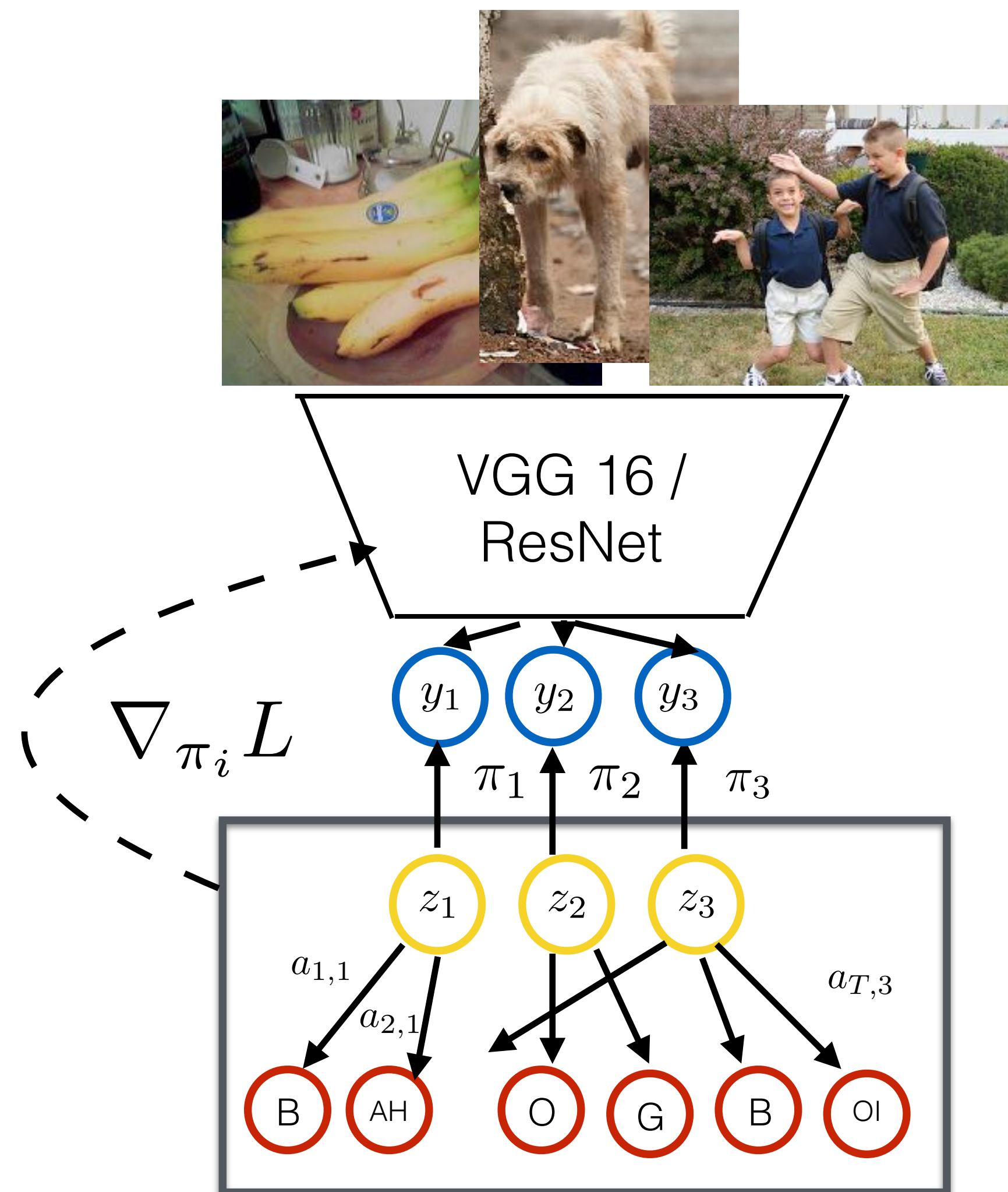
$$= \frac{\sum_{i=1}^L \sum_{t: x_t = x}^T q(i_t = i) q(z_i = k)}{\sum_{i=1}^L \sum_{t=1}^T q(i_t = i) q(z_i = k)} \quad (24)$$

$$\bar{p}(z_i = k) = \frac{\sum_{i=1}^L q(z_i = k)}{\sum_{i=1}^L \sum_{k=1}^K q(z_i = k)}. \quad (25)$$

Issues with Mean-field VI

- Potentially large gap between the approximate and true posteriors
- Prior somewhat arbitrary; no need to compute the full posterior with priors
- In practice, converge to bad local optimum, even with pretraining

Model 2: HMM-DNN with Exact EM Inference



Exact EM Inference Assumptions

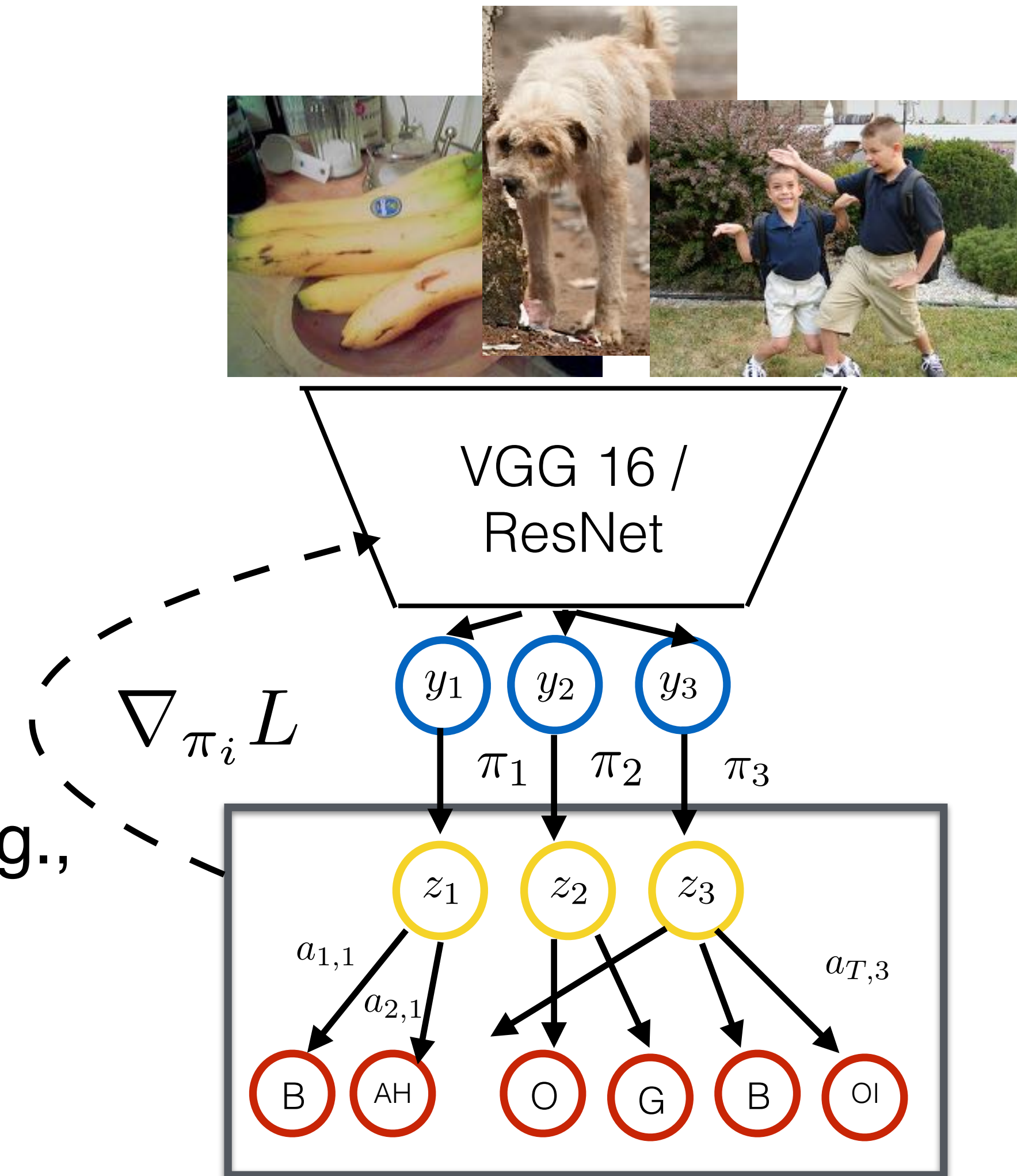
- Uniform prior on y :

$$p(\mathbf{x}, \mathbf{y}) \propto p(\mathbf{x}|\mathbf{y})$$

- Softmax **unimodal posterior**:

$$\pi_{ik} := p(z_i = k|\mathbf{y}) = \frac{\exp(g_k(y_i))}{\sum_{k'=1}^K \exp(g_{k'}(y_i))}$$

where $g_k(\cdot)$ is the **activation function** of y , e.g.,
linear, log-Gaussian, neural net



HMM-DNN: EM Algorithm with Augmented State Space

- **Augmented state space:** $s_t := (i_t, z_i)$
- **Forward probabilities:** $\alpha_t(i, z) := p(\mathbf{x}_{1:t}, i_t = i, z_i = z | \mathbf{y})$
- **Backward probabilities:** $\beta_t(i, z) := p(\mathbf{x}_{t+1:T} | i_t = i, z_i = z, \mathbf{y})$
- **State posterior probabilities:**

$$\gamma_t(i, z) := p(i_t = i, z_i = z | \mathbf{x}, \mathbf{y})$$
$$\zeta_t(i, z, i, z') := p(i_t = i, z_i = z, i_{t+1} = j, z_j = z' | \mathbf{x}, \mathbf{y})$$

HMM-DNN: Posterior Update

1. Update forward probabilities:

$$\alpha_1(i, z) = p(i_t = i | N) \pi_{ik} p(x_1 | z)$$

$$\alpha_{t+1}(i, z) = \boxed{\alpha_t(i, z) p(i | i, N) p(x_{t+1} | z)} + \boxed{\left(\sum_{j \neq i}^N \sum_{z'=1}^K \alpha_t(j, z') p(i_{t+1} = i | i_t = j) \right) \pi_{iz} p(x_{t+1} | z)}$$

2. Update backward probabilities:

$$\beta_T(i, z) := 1.$$

$$\beta_t(i, z) = \boxed{\beta_{t+1}(i, z) p(i | i, N) p(x_{t+1} | z)} + \begin{cases} \boxed{\left(\sum_{j \neq i}^N \sum_{z'=1}^K \beta_{t+1}(j, z') p(i_{t+1} = j | i_t = i) \right) \pi_{iz} p(x_{t+1} | z)} & i_t \neq i_{t-1} \\ \boxed{\beta_{t+1}(i, z) p(i | i, N) p(x_{t+1} | z)} & i_t = i_{t-1} \end{cases}$$

HMM-DNN: Posterior Update

3. Update state posterior at time t:

$$\gamma_t(i, z) = \frac{\alpha_t(i, z)\beta_t(i, z)}{\sum_{i=1}^N \sum_{z=1}^K \alpha_t(i, z)\beta_t(i, z)}$$

4. Update state transition posterior at time t:

$$\xi_t(i, z) = \begin{cases} \frac{\alpha_t(i, z)p(i|i, N)p(x_{t+1}|z)\beta_{t+1}(i, z')}{Z}, & \text{if } i = j, \\ \frac{\alpha_t(i, z)p(j|i, N)\pi_{jz'}p(x_{t+1}|z')\beta_{t+1}(j, z')}{Z}, & \text{otherwise.} \end{cases}$$

where $Z := \sum_{(i,z)=(1,1)}^{(N,K)} \alpha_t(i, z)p(i|i, N)p(x_{t+1}|z)\beta_{t+1}(i, z') + \sum_{(i,z)=(1,1)}^{(N,K)} \sum_{j \neq i, z'=1}^{N,K} \alpha_t(i, z)p(j|i, N)\pi_{jz'}p(x_{t+1}|z')\beta_t(j, z')$

HMM-DNN:

HMM Parameter Update

$$\hat{p}(i|N) = \frac{\sum_{t=1}^T \sum_{z=1}^K \gamma_t(i, z)}{\sum_{i=1}^N \sum_{t=1}^T \sum_{z=1}^K \gamma_t(i, z)}$$

$$\hat{p}(j|i, N) = \frac{\sum_{t=1}^T \sum_{z, z'=1}^{K, K} \xi_t(i, z, j, z')}{\sum_{j=1}^N \sum_{t=1}^T \sum_{z, z'=1}^{K, K} \xi_t(i, z, j, z')}$$

$$\hat{p}(x_t = x|z) = \frac{\sum_{t: x_t = x}^T \sum_{i=1}^N \gamma_t(i, z)}{\sum_{t=1}^T \sum_{i=1}^N \gamma_t(i, z)}$$

HMM-DNN: Unimodal Posterior Parameter Update

- Analogous to the HMM-DNN proposed in Bengio et. al. (1992), we have:

$$\begin{aligned} L_{MLE} &:= \log p(\mathbf{x}|\mathbf{y}) \\ \frac{\partial L_{MLE}}{\partial g_{ik}} &= \frac{1}{p(\mathbf{x}|\mathbf{y})} \frac{\partial p(\mathbf{x}|\mathbf{y})}{\partial g_{ik}} \\ &= \frac{1}{p(\mathbf{x}|\mathbf{y})} \sum_{j=1}^K \frac{p(\mathbf{x}, z_i = j|\mathbf{y})}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial g_{ik}} = \sum_{k=1}^K p(z_i = k|\mathbf{x}, \mathbf{y}) \left(\frac{\partial g_{ij}}{\partial g_{ik}} - \pi_{ik} \right) \\ &= \sum_{k=1}^K \boxed{p(z_i = k|\mathbf{x}, \mathbf{y}) - \pi_{ik}} := \Delta_{ik} \end{aligned}$$

Gap between the unimodal and multimodal posterior at $z = k$

Unimodal Posterior Parameter Update for Different Activation Functions

- Log-Gaussian activation:

$$\mu_k^* = \frac{\sum_{i=1}^N \Delta_{ik} y_i}{\sum_{i=1}^N \Delta_{ik}}$$

- Linear activation:

$$w_{kd}^{next} = w_{kd} - \eta \frac{L_{MLE}}{w_{kd}} = w_{kd} - \eta \sum_{I=1}^N \Delta_{ik} y_{kd}$$

- Neural network activation:

$$w_{kd}^{next} = w_{kd} - \eta \sum_{i=1}^L \Delta_{ik} h_{id}$$
$$V_{ld}^{next} = V_{jd} - \eta \sum_{i=1}^N \sum_{j=1}^K \Delta_{ik} \frac{\partial g_{ij}}{\partial h_{id}} y_{il}$$

HMM-DNN: Decoding

- Two-Step Search:
- Viterbi search for alignment (exact):

$$A^* = \arg \max_A p(x, \mathbf{A} | \mathbf{y})$$

- Approximate search for cluster:

$$z^* = \arg \max_z p(x, \mathbf{z} | \mathbf{A}^*, \mathbf{y})$$

Outline

- Motivation
- Prior Works
- Problem Formulation
- Model Description
- Experiments
- Image-to-Speech Discovery (in-progress work)
- Conclusion

Datasets

- Synthetic captions from MSCOCO:
 - Randomly draw 5 image regions of distinct concepts from the validation set of 800000 images and concatenate the pronunciation of concept name as the caption
 - Keep only images with names appear in the actual captions, expand the person class
- Phone set from SpeechCOCO
- MSCOCO 2 k: 65 image classes, 2541 images, ~200 images per image class
- MSCOCO 20 k: 65 image classes, 19925 images, ~2000 images per image class

Hyper parameter setting - CorrLDA

- Stick-breaking concept prior: Uniform, $\gamma_{k0} = 1, \gamma_{k1} = \frac{1}{n - k + 1}$
- Alignment prior and prior of the visual centroids: $\gamma_{i0}^a = 0.1, \kappa_0 = 0.01$
- K-Means initialization of the visual centroid
- Fixed variance equal to 0.3 * sample variance

Hyper parameter setting - HMM-DNN

- Uniform initialization of alignment and phone observation probabilities
- K-Means initialization of the Gaussian softmax with width 1
- Initialize W with the classification layer of the VGG 16/ Res 34 network trained on ImageNet and fine-tuned on MSCOCO Training set (50 % accuracy)

Evaluation Metrics

- Accuracy: The ratio between the number of correctly aligned phones and the total number of phones
- F1: The harmonic mean of the alignment recall and precision
 - Alignment recall: The **average** ratio between the number of correctly aligned phones and the total number of phones **given each concept**
 - Alignment precision: The **average** ratio between the number of correctly aligned phones and the total number of phones predicted to be a **given concept**
- Purity: The average ratio between the number dominant image label and the cluster size in each cluster

Overall Results

		Gaussian (VGG 16)	Gaussian (Res 34)	Neural Net (VGG 16)	Neural Net (Res 34)	CorrLDA
MSCOCO 2k	Accuracy	73.5 %	75.7 %.	14.2 %	66.0 %	52.6 %
	F1	71.2 %	74.3 %	17.6 %	61.6 %	51.2 %
	Purity	43.7 %	47.0 %.	33.9 %	68.7 %	40.4 %
MSCOCO 20k	Accuracy	67.5 %	70.2 %	26.8 %	68.8 %	51.5 %
	F1	63.6 %	68.1 %	14.0 %	65.8 %	50.2 %
	Purity	34.6 %	39.6 %	33.1 %	58.6 %	33.8 %

Clustering and Alignment Performance across Concepts

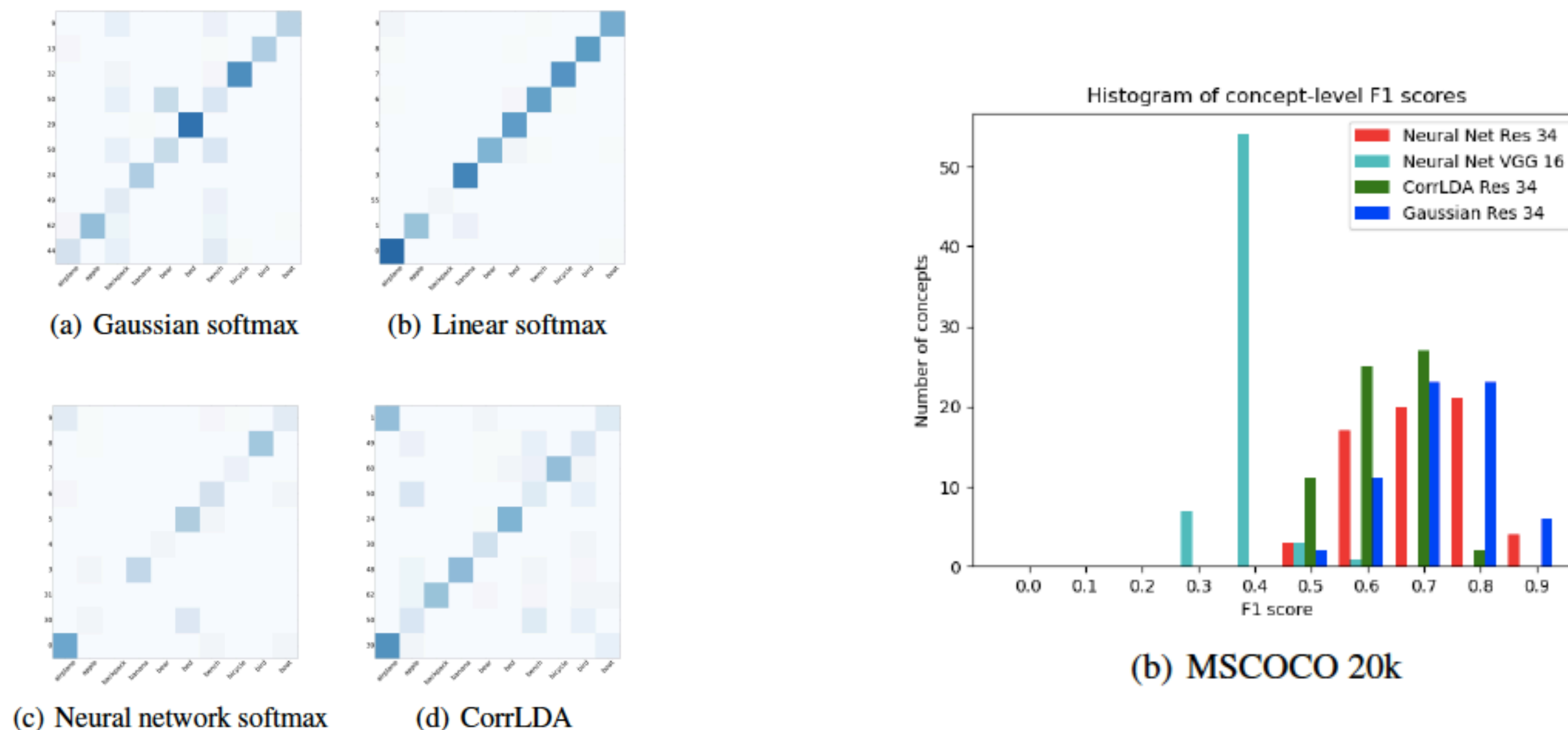
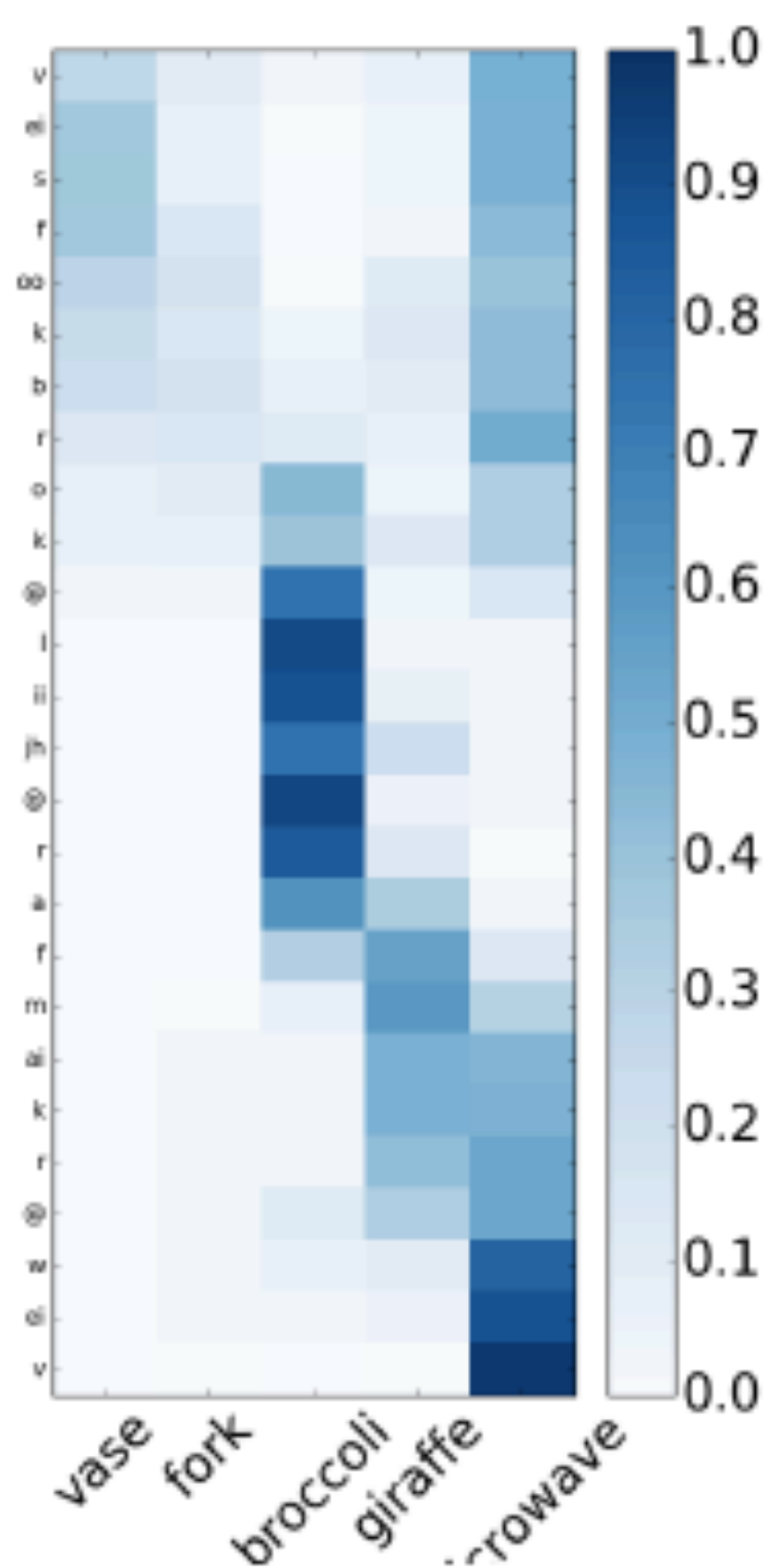
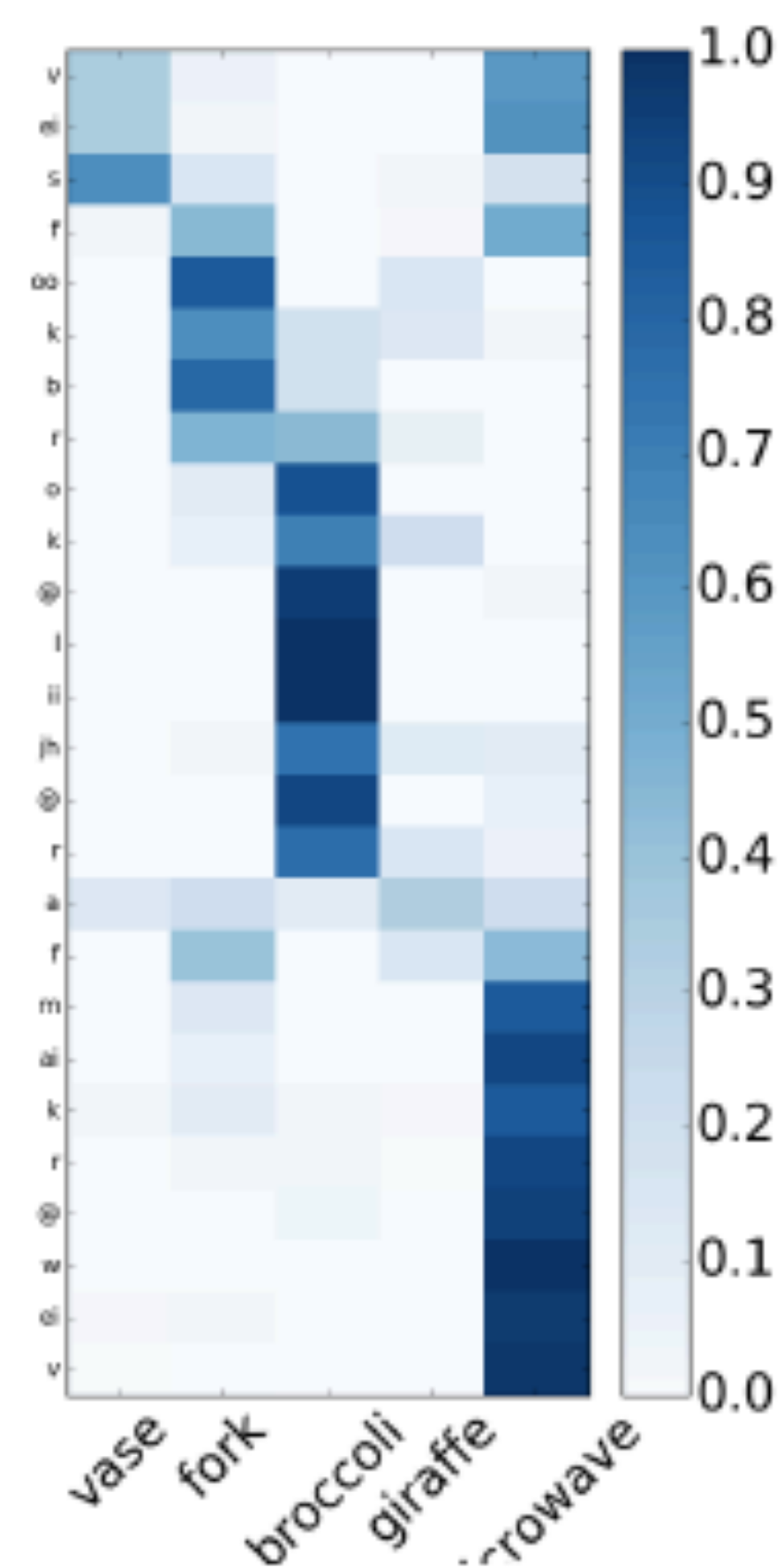


Figure 4. Confusion matrix of the first ten concepts in alphabetical order for various models on MSCOCO 20k

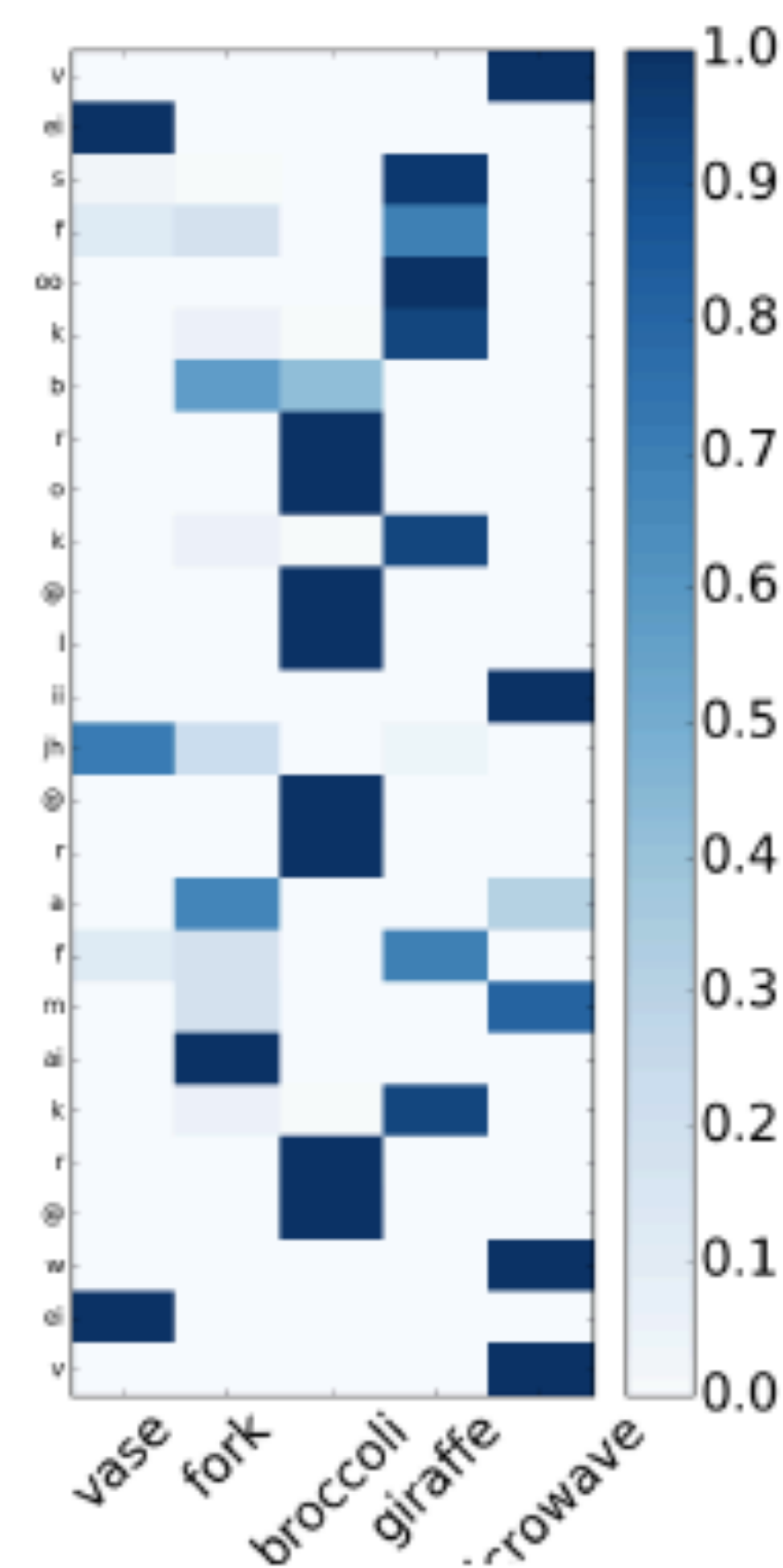
Alignment Matrices



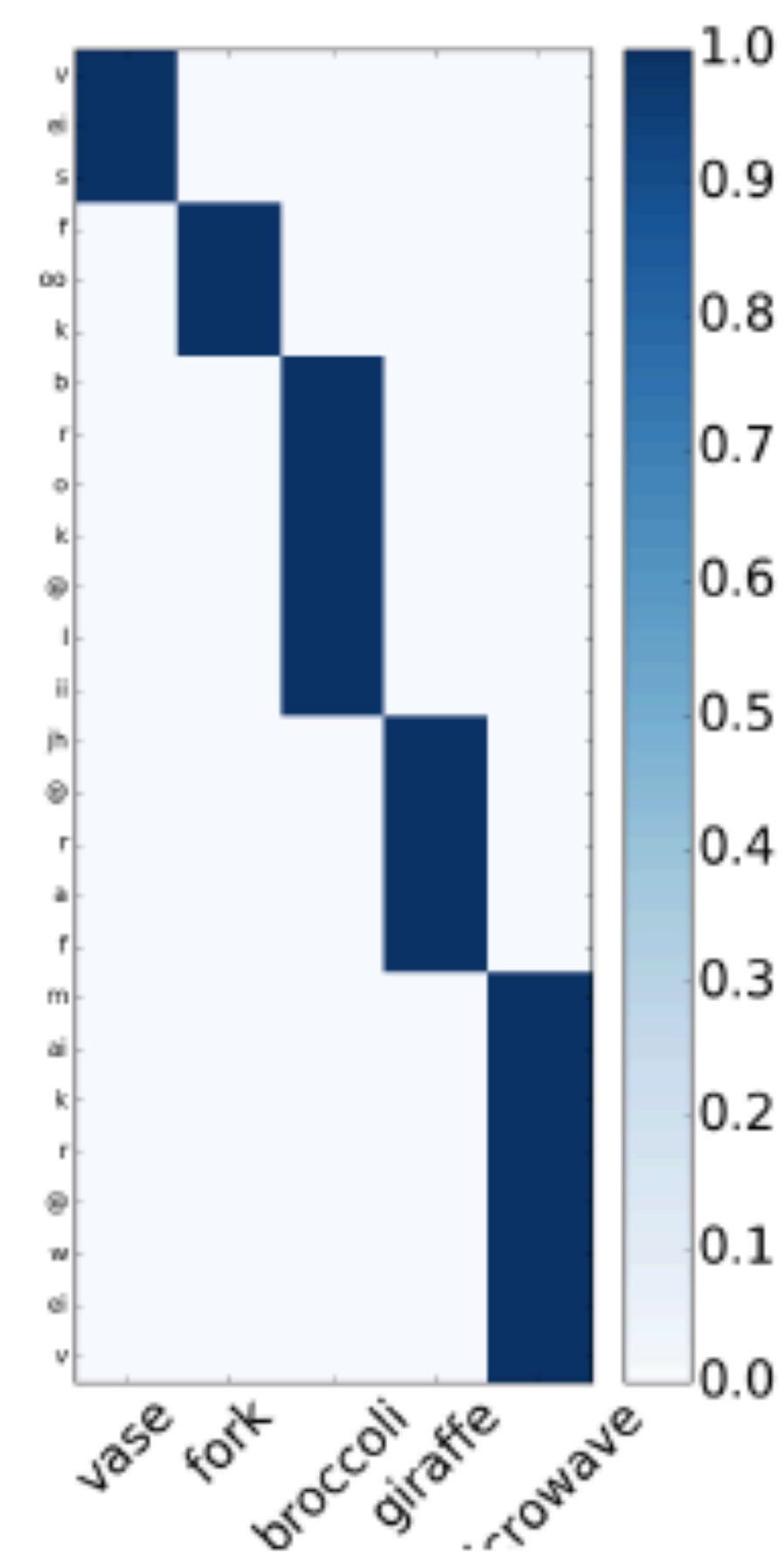
(a) Linear softmax



(b) Gaussian softmax



(c) CorrLDA



(d) Ground truth

Outline

- Motivation
- Prior Works
- Problem Formulation
- Model Description
- Experiments
- Image-to-Speech Discovery (in-progress work)
- Conclusion

Current Work

- Replace phone strings with spoken phone segments

- Key approximation:

$$p(x_t|\phi_t) = \frac{p(x_t)p(\phi_t|x_t)}{p(\phi_t)} \propto p(\phi_t|x_t)$$

- Use methods similar to the visual concept posterior to model $p(\text{phone}|x_t)$
- Audio features:
 - Resampled fixed-length embedding (inspired by Kamper 2016, 2017)
 - LSTM-CTC hidden layer (pre-trained on audio and phone strings from a rich resource language)
 - Unsupervised acoustic embeddings?

Preliminary Results

	Resample	CTC Last	CTC Mean
Accuracy	73.5	61.8	72.7
F1	72.0	60.0	72.8
Purity	99.5	98.4	99.8

Table 1: Word discovery results on MSCOCO 2k with synthetic (Gaussian) visual features and spoken phone sequence

Conclusion

HMM-DNN:

- Polynomial-time, data efficient, interpretable
- Outperforms CorrLDA on word discovery task
- Assumes availability of image and phone segmentation -> still a long way to go for the actual low-resource setting