Topic Models for Texts and Images in Representation Space

Kui Tang and Sameer Lal

Columbia University

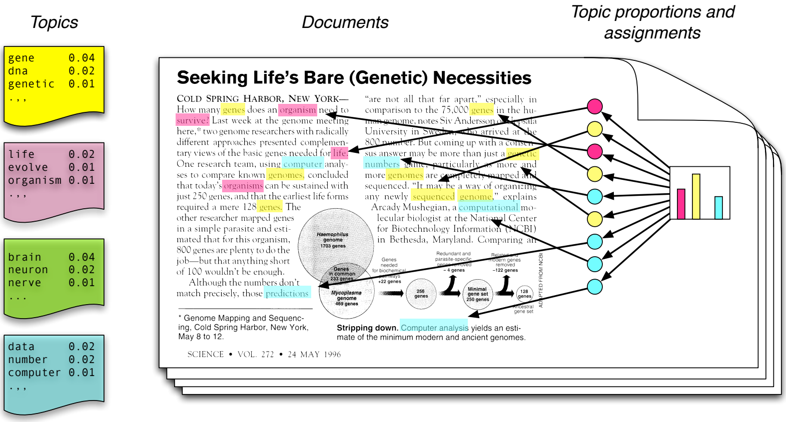
29 April 2015

Outline

1. Review topic models and multimodal embeddings.
    1.1 Proposed joint model
    1.2 Actual layer-wise model.
    1.3 Data
2. Image-word alignment model (DeViSE) (Frome et al., 2013).
    2.1 Results
    2.2 Demo
3. Mixture of Gaussian topic model (original work).
    3.1 Text training
    3.2 Text + image training.
4. Conclusions + future work.

Review Topic Models and Multimodal Embeddings.

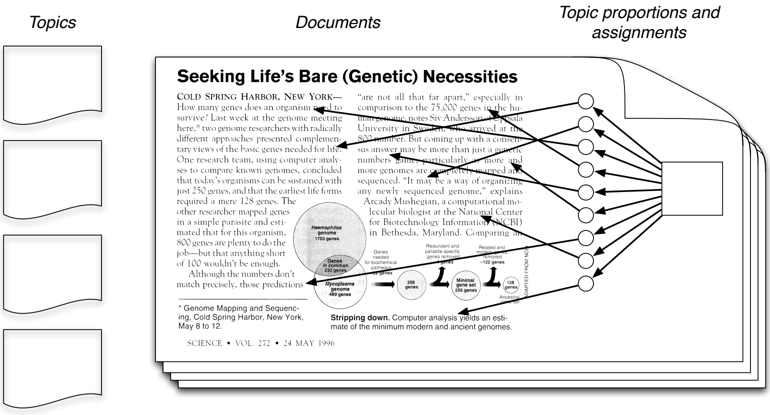
Topic Models



- LDA assumes that there are K topics shared by the collection.
- Each document exhibits the topics with different proportions.
- Each word is drawn from one topic.
- We discover the structure that best explain a corpus.

Slide stolen from D. Blei.

Latent Variable Models

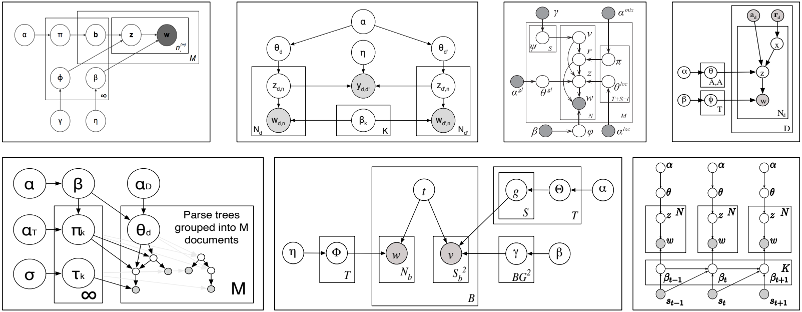


- Our goal is to **infer** the hidden variables
- I.e., compute their distribution conditioned on the documents

$$p(\text{topics}, \text{proportions}, \text{assignments} \mid \text{documents})$$

Slide stolen from D. Blei.

Bayesian Networks



Slide stolen from D. Blei.

- Shaded variables are *observed*, other variables are *hidden*.
- A model is our hypothesis for how data are generated.
- We *condition* on observations to update our hypothesis.

Multimodal Documents

A milking machine in action

farms also grow their own feed, typically including corn, alfalfa, and hay. This is fed directly to the cows, or stored as silage for use during the winter season. Additional dietary supplements are added to the feed to improve milk production. [10]

4.2 Poultry farms

Poultry farms are devoted to raising chickens (egg layers or broilers), turkeys, ducks, and other fowl, generally for

Typical cattle yard in Northern Iowa, USA

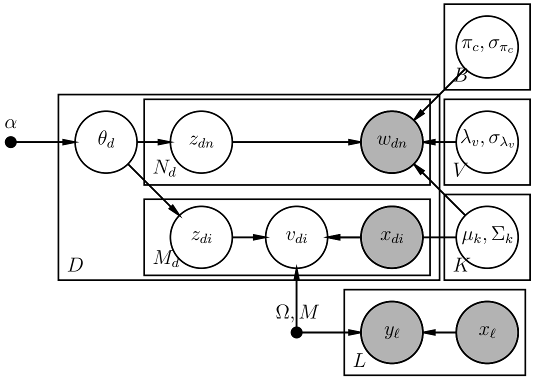
accompanied by the decoupling of political power from farm ownership.

5.1 Forms of ownership

In some societies (especially socialist and communist), collective farming is the norm, with either government ownership of the land or common ownership by a local group. Especially in societies without widespread industrialized farming, tenant farming and sharecropping are common; farmers either pay landowners for the right to use farmland or give up a portion of the crops.

- *We want to learn a topic model using text and images jointly.*
- Images and text complement each other.
- Captions aren't the whole story: cows in political contexts.

Proposed joint model (work in progress)



- Topics are (mixtures of) Gaussians.
- Words are latent vectors $\lambda_v \in \mathbb{R}^{D_W}$ using Bayesian word2vec.
- Images are latent vectors $v_{in} \in \mathbb{R}^{D_I}$ *conditioned* on raw images x_{di}. We have $v_{ni} \sim \mathcal{N}(M\text{CNN}_x(x_{ni}; \Omega), \Sigma)$ with Ω CNN parameters, M mapping to word vector space, and CNN_x feature representation output by CNN.

Variational Bayesian EM (eventually)

To learn latent variable models, maximize the *marginal likelihood*

$$\max_{\theta} p(\mathbf{x}|\theta) = \int p(\mathbf{x}, \mathbf{z}|\theta) p(\mathbf{z}|\theta) d\mathbf{z}$$

This integral is intractable. Approximate instead with the *evidence lower bound (ELBO)*

$$\log p(\mathbf{x}|\theta) \geq E_{q(z|\phi)} [\log p(\mathbf{x}, \mathbf{z}|\theta) - \log q(\mathbf{z}|\theta)] =: \mathcal{L}(\theta, \phi)$$

where $q(\mathbf{z}|\phi)$ is a simple *variational distribution* which approximates the posterior $p(\mathbf{z}|\mathbf{x}, \theta)$.

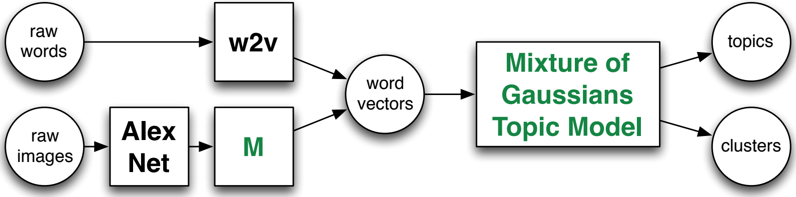
Variational Bayesian EM:

- E Step: Update $\phi^{(t+1)} \leftarrow \arg\max_{\phi} \mathcal{L}(\theta^{(t)}, \phi)$
- M Step: Update $\theta^{(t+1)} \leftarrow \arg\max_{\theta} \mathcal{L}(\theta, \phi^{(t)})$

E step is variational Bayesian inference (Ranganath, Gerrish, and D. M. Blei, 2014; Wang and D. M. Blei, 2013). M step is learning (updating) a CNN with objective

$$\min_{\Omega} \sum_{\ell} L(y_\ell; \mathrm{CNN}_y(x_\ell; \Omega)) + \frac{1}{2\sigma^2} \sum_{di} E_{q(v_{di}|\phi^{(t)})} \left[(v_{di} - \mathrm{CNN}_x(x_{di}; \Omega))^2 \right]$$

Actual layer-wise model



- Train image-word alignment M and mixture of Gaussian topic model separately.
- Pretrained word2vec model on 3 million word/phrase vocabulary, 100 billion word corpus.
- Pretrained Caffe reference network (Jia et al., 2014), derived from AlexNet (Krizhevsky, Sutskever, and Hinton, 2012).
- No fine-tuning (for now)

Data

Data

- Imagenet's 1.3 million training images over 1000 classes
- Only used 10% for training currently (100,000 image vectors)
- Wikipedia pages for each of the 1000 classes
- In reality, far less, due to synsets not being in pretrained word2vec
- Used Google's word2vec pretrained word vectors from the Google News Corpus. This corpus had 10 billion words, and generated 3 million word vectors.

Extraction of AlexNet Features

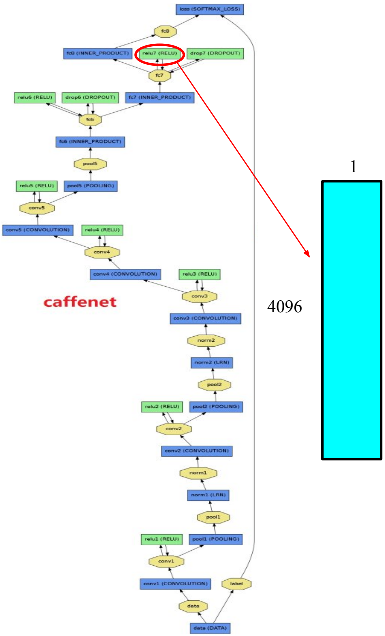
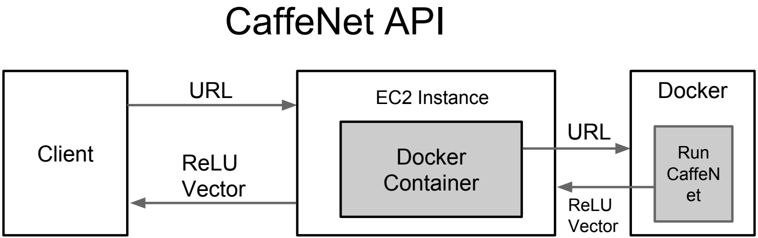


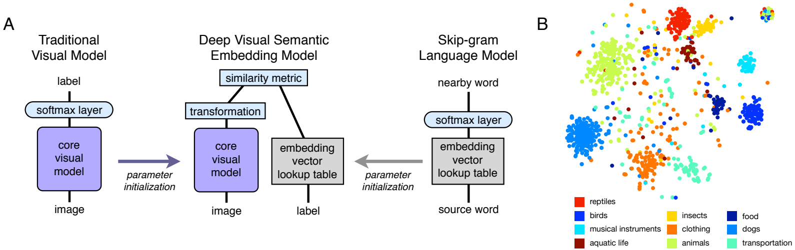
Image Vectorization API



- API utilizes pretrained CaffeNet Model
- Python interface to get classification and image features
- Example call:
  `image_vec, softmax_vec = transform(image_url)`
- Can expand to images on local (client) machine

Image-word alignment model (M).

Transform Raw Images to Word Vectors



- Learn M by minimizing a ranking loss:

$$\ell(v, y) = \sum_{y' \neq y} \max \left[0, \lambda - w_y^\top M v + w_{y'}^\top M v \right]$$

where v is image vector, y is image label, w is word vector. Sum this term over all (v, y) pairs in labeled data.

- Instead of summing all $y' \neq y$, randomly iterate y' and return first example violating the margin.

Results

Strawberries in a kitchen

"strawberry" word vector neighbors: strawberries blueberry berry tomato peaches peach blueberries rhubarb berries cherries watermelon apricot melon asparagus citrus grape mango pear ripe_strawberries raspberry

$Mv_{\textbf{above}}$ **neighbors**: strawberry pecan lime_mousse Chocolate_Marshmallow Burmese_microplate_along Alberto_Callapso freshly_baked_pie grilled_ciabatta pinch_hitter_Felipe_Lopez earthenware_dish tater boysenberry_pie chocolate_sauce clair minor_leaguer_Joba_Chamberlain fanned_Aaron_Harang reliever_Dennis_Sarfate almond currant_jelly pear

Strawberries in cereal

"strawberry" word vector neighbors: strawberries blueberry berry tomato peaches peach blueberries rhubarb berries cherries watermelon apricot melon asparagus citrus grape mango pear ripe_strawberries raspberry

$Mv_{\textbf{above}}$ **neighbors**: cinnamon_glaze gravy_peas gelatin_salad salty_pancetta pistachios_almonds sweet_potato_casseroles strawberry gravy_mashed_potatoes gravy_broccoli stewed_rhubarb lentil coarsely_crushed tangy_salsa candied_fruit salty_spicy creamed_cabbage veggie_salad Coarsely_grate Bodega_Chocolates apricot_preserves

Strawberries in a bowl

"strawberry" word vector neighbors: strawberries blueberry berry tomato peaches peach blueberries rhubarb berries cherries watermelon apricot melon asparagus citrus grape mango pear ripe_strawberries raspberry

$Mv_{\textbf{above}}$ **neighbors**: rimmed_martini_glass Bodega_Chocolates clove_studded butter_lettuce fresh_raspberries apricot_jelly raisin_growers Golden_Delicious_apples macerated Wendy_chili manioc_bananas vanilla_mousse jar fried_chicken_sunflower_seeds tomatoes_cilantro tortilla_chips_salsa citrus quince_jam diced_pineapple dragonfruit

Volcano erupting

"volcano" word vector neighbors: volcanoes eruption volcanic_eruption dormant_volcano rumbling_volcano lava Merapi volcanic volcano_erupted volcanic_activity Merapi_volcano Mt_Merapi active_volcanoes eruptions volcano_eruption Mount_Bulusan spews_ash lava_flow Kilauea_volcano Grmsvtn

Mv_{above} **neighbors**: volcano stray_firework erupt_explosively lava_spewing Sivand_dam incandescent_lava toxic_gasses eruption noxious_smelling Kapakis Flares_lit pyroclastic_flow volcanic_eruption Indonesia_Mount_Merapi Scott_Kardel mudflow Grimsvtn_volcano chief_Turhan_Yussef Mount_Unzen spewing_ash

Volcano behind promontory

"volcano" word vector neighbors: volcanoes eruption volcanic_eruption dormant_volcano rumbling_volcano lava Merapi volcanic volcano_erupted volcanic_activity Merapi_volcano Mt_Merapi active_volcanoes eruptions volcano_eruption Mount_Bulusan spews_ash lava_flow Kilauea_volcano Grmsvtn

$Mv_{\textbf{above}}$ **neighbors**: volcano promontory_overlooking lava Mount_Semeru Pennypack_Creek Testalinden_Creek Vesuvius densely_treed Pigeon_Creek Llaima_volcano volcanic erupting_volcano undammed Karthala outcropping rocky_outcrop_overlooking rocky_escarpment Dan_Oltrogge lava_spewing volcanic_mudflow

Volcano behind wooded hills

"volcano" word vector neighbors: volcanoes eruption volcanic_eruption dormant_volcano rumbling_volcano lava Merapi volcanic volcano_erupted volcanic_activity Merapi_volcano Mt_Merapi active_volcanoes eruptions volcano_eruption Mount_Bulusan spews_ash lava_flow Kilauea_volcano Grmsvtn

$Mv_{\textbf{above}}$ **neighbors**: wooded_slopes fever_swamps precipitous_cliffs barren_hills Maoist_insurrection sq._kilometer ancient_lava_flows thickly_forested sandy_ridges forested_slopes granite_peaks alpine_vistas towering_cliffs Judean_hills ruggedly_beautiful unspoiled_wilderness Tohme_financier verdant spectacular_gorges Dangrek

Corn and many other foods

"corn" word vector neighbors: soybean soybeans wheat corn_crop Corn soy_bean corn_soybean sweet_corn soyabeans grain wheatfields_underwater grain_sorghum soy_beans corn_acreage crops Soybean corn_kernels sorghum Soybeans soy

$Mv_{\textbf{above}}$ **neighbors**: chargrilled_chicken crispy_shallots Mozzarella_sticks Miso_soup bun_french_fries peas_cranberry_sauce turkey_gravy Dipping_Sauce flatbread_sandwich succulent_scallops roasted_pecans Panang_Curry freshly_steamed char_siu sesame_chicken taco_salad Coconut_Soup Shrimp_Tacos homemade_sausage_gravy nacho_cheese

Corn, brown kernels

"corn" word vector neighbors: soybean soybeans wheat corn_crop Corn soy_bean corn_soybean sweet_corn soyabeans grain wheatfields_underwater grain_sorghum soy_beans corn_acreage crops Soybean corn_kernels sorghum Soybeans soy

$Mv_{\textbf{above}}$ **neighbors**: tuber_crops Ag_Processing corn Archer_Daniels_Midland_ADM insect_larvae crispy_shallots low_linolenic_soybeans Nasdaq_CVGW nut_butter procures_transports_stores distributor_Chiquita_Brands SRW_wheat catfish_filet Olive_Garden_chains brioche_toast oilseed_processing microgreen potato_chips_pretzels Bunge_Ltd_BG.N agribusiness_conglomerate

Acorn and oak leaves

"acorns" word vector neighbors: acorns hickory_nut pine_cone squirrel_nibbling pinecone beechnuts hairy_woodpecker red_oak_acorns suet_cake yaupon hickory_nuts squirrel oak_leaf hickory_trees pine_cones maple_sapling acorns_beechnuts ligustrum yaupon_holly hornworm

$Mv_{\textbf{above}}$ **neighbors**: acorn sprout oak_leaf sap_sucking sugar_beet_stecklings Sen._Jarrett_Barrios germinating Moringa_Oleifera planting acorns leaf_undersides seedpod blooming oxalis almond_groves tongues_wagging planted nutlets yaupon Mr._Swindal_apologizes

Acorn on brown ground

"acorns" word vector neighbors: acorns hickory_nut pine_cone squirrel_nibbling pinecone beechnuts hairy_woodpecker red_oak_acorns suet_cake yaupon hickory_nuts squirrel oak_leaf hickory_trees pine_cones maple_sapling acorns_beechnuts ligustrum yaupon_holly hornworm

$Mv_{\textbf{above}}$ **neighbors**: pine_oak eucalyptus_pine pine_cones reseeds_itself replanted Huckleberries_nuts maple_basswood sawtooth_oak chewed_wad clivias pine_straw ant_hill planted lupine_seeds mountains_denuded pickerel_weed cedar_elms thorny_scrub mulched_beds uprooted

Demo

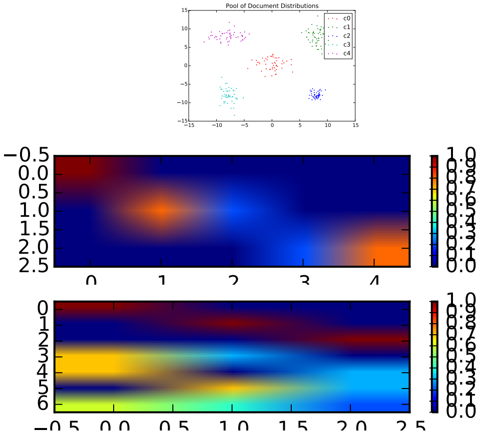
Mixture of Gaussians Topic Model

Generative Process

The model assumes a large dictionary of "concepts", which are Gaussian clusters in semantic space. A topic is a mixture of these concepts, and each vector x_{dn} (word or image) is described by a mixture of topics. The generative process is as follows:

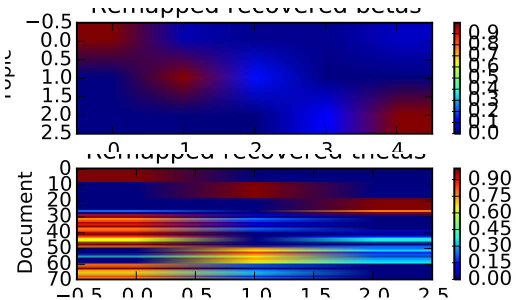
- For $k = 1, \ldots, K$ (for each topic):
  - Draw $\beta_k \sim \mathrm{Dir}(\alpha)$
  - Draw $\lambda_k \sim \mathrm{Gamma}(10^{-6}, 10^{-6})$
  - Draw $\mu_k \sim \mathcal{N}(0, \mathrm{diag}(\tau))$
- For $d = 1, \ldots, D$ (for each document):
- Draw $\theta_d \sim \mathrm{Dir}(\gamma)$
- For $n = 1, \ldots, N_d$ (for each word in document):
  - Draw $z_{dn} \sim \mathrm{Mult}(\theta_{dn})$
  - Draw $c_{dn} \sim \mathrm{Mult}(\beta_{z_{dn}})$
  - Draw $x_{dn} \sim \mathcal{N}(\mu_{c_{dn}}, \mathrm{diag}(\tau_{c_{dn}})$

Gaussian LDA Synthetic Problem



- Implemented variational message passing (Winn and Bishop, 2005) and stochastic variational inference using the BayesPy (Luttinen, 2014) package.
- Generate 5 Gaussian clusters (top), 3 topics consisting of mixtures of these clusters (mid) and documents as a mixture of topics.

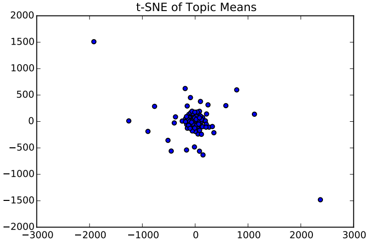
Gaussian LDA Synthetic Problem

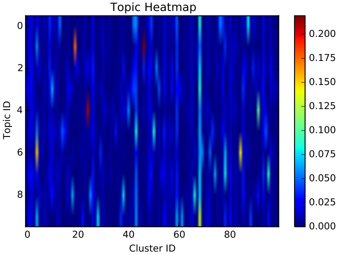


- Recovered essentially the same parameters we used to generate data.
- Model is well-specified and approximation algorithm works.

Mixture of Gaussians Topic Model Results — Text Only

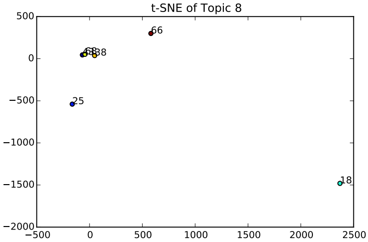
Recovered Topics and Clusters

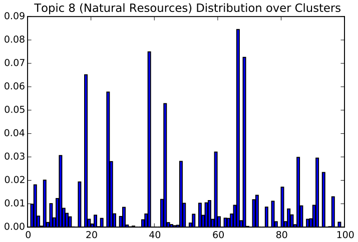




- Batch variational inference on 100 docs, 133,866 words.
- Selected topics (words from Google News corpus, 3 million word vocabulary.)
  - **Geography**: east west south north southeast southwest above signs united people regions areas levels states folk properties places sites locations cities
  - **Natural resources**: water temperature cold heat temperatures areas regions properties places locations cities towns parts sites natural_gas electricity gasoline gas fuel electric
  - **Music**: game mother folk culture traditions cultural strings vibrato pizzicato trombone instrument Mozart_concerto orchestra oboe flute cello harpsichord clarinet cellist soprano_saxophone

Example topic and cluster breakdown 1/2





Cluster	Words
18	**[Saltwater]** shoreline ocean coastline nearshore_reefs sandy_shorelines coastal_waters shallow_reefs tidal_creek shallow_waters mud_flats sea tidal_inlet pier_pilings underwater reef shoreward abyssal_plain inter_tidal shifting_sandbars sandy_bottomed
25	**[Freshwater]** water ice surface green porpoise_vaults surficial_aquifer rainwater Floridan_aquifer radar_deflectors wa_ter absorbs_carbon_dioxide bermed absorbs_sunlight bugs_wiggling Mosquitoes_breed overflowing_septic_tanks mild_dishwashing_liquid reverse_osmosis_filtration hyper_saline secondary_clarifier

Example topic and cluster breakdown 1/2

Cluster	Highest Probability Words
38	**[Chemicals]** hydrous calcium_oxide cyclohexane inorganic_salts calcium_sulphate fluorocarbons Sodium_cyanide silicate_rocks Nitric_acid chemically_reactive calcium_carbonates magnesium_silicate outgas raffinate potassium_salts bacterial_decomposition methane trihalomethanes_THMs element_boron Sulphur_dioxide
66	**[Volcanoes]** coral reefs reef corals coral_reefs ocean volcanoes sea coral_reef volcanic islands lava volcano oceans undersea_volcanoes oceanic ocean_basins lava_flows eruptions Kilauea_Volcano

Properties of Mixture of Gaussian LDA Model

- Captures both local (word2vec neighborhoods, context) semantic and syntactic similarity, as well as broader topical similarity.
- Mixture of Gaussian crucial: components of topic can be *far* in semantic space. Existing global semantic models, e.g. paragraph vectors (Le and Mikolov, 2014) still require locality in semantic space.
- Semantic space representation permits *explaining* topics using a much larger corpus than the training corpus.
  - Generalize across corpora.
  - Get good qualitative results even with small data.

Mixture of Gaussians Topic Model Results — Text + Images

Coming soon!

Conclusions + Future Work

- We have shown a proof-of-concept of a multimodal topic model in representation space:
  - Re-implemented DeViSE in CUDA; wrapped into a fast test-time API.
  - Derived and fit *mixture of Gaussian topic models* (MoGTA), a novel model that can be fit with standard techniques with intriguing properties on pre-trained word vectors.
- We have much work to do to make this a proper probabilistic model:
  - Demonstrate multimodal inference, modeling vectors for words and images simultaneously.
  - Improve Bayesian word2vec to be competitive with non-Bayesian versions.
  - Join Bayesian word2vec with MoGTA to form one joint model.
  - Fine-tune image vector updates (variational Bayesian EM).

Mixture of Gaussians Topic Model — Text + Images

Thank You

- Questions?

References I

Frome, A. et al. (2013). "Devise: a deep visual-semantic embedding model". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. Burges et al. Curran Associates, Inc., pp. 2121–2129.

Jia, Y. et al. (2014). "Caffe: convolutional architecture for fast feature embedding". In: *arXiv preprint arXiv:1408.5093*.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., pp. 1097–1105.

Le, Q. and T. Mikolov (2014). "Distributed representations of sentences and documents". In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. Ed. by T. Jebara and E. P. Xing. JMLR Workshop and Conference Proceedings, pp. 1188–1196.

References II

Luttinen, J. (Oct. 2014). "BayesPy: Variational Bayesian Inference in Python". In: *ArXiv e-prints*.

Ranganath, R., S. Gerrish, and D. M. Blei (Dec. 2014). "Black Box Variational Inference". In: *ArXiv e-prints*.

Wang, C. and D. M. Blei (2013). "Variational inference in nonconjugate models". In: *Journal of Machine Learning Research* 14.1, pp. 1005–1031.

Winn, J. M. and C. M. Bishop (2005). "Variational message passing". In: *Journal of Machine Learning Research* 6, pp. 661–694.