# MULTIMODAL REPRESENTATION: KNESER-NEY SMOOTHING/SKIP-GRAM BASED NEURAL LANGUAGE MODEL

*Mingoo Song, Chang D. Yoo*

Korea Advanced Institute of Science and Technology
School of Electrical Engineering
291, Daehak-ro, Yuseong-gu, Daejeon-si 34141, Republic of Korea

## ABSTRACT

For image retrieval and caption generation, this paper considers a multimodal representation that associates image with its text description (caption) by defining a neural language model as the conditional probability of the next word given both $n$ past words in a caption and the image that the caption describes. To address the data sparsity problem, the use of the Kneser-Ney smoothing and skip-gram models is examined by integrating each into the multimodal neural language model. A language model (LM) known as Kneser-Ney smoothing is based on absolute-discounting interpolation while skip-gram LM is based on $n$-grams organized by allowing intermediate tokens to be *"skipped"*. The multimodal representation is evaluated on the IAPR TC-12 dataset. Using perplexity and BLEU-$n$ measures, both Kneser-Ney smoothing and skip-gram models are demonstrated to be more effective as approaches to addressing the data sparsity problem than the generic $n$-gram model used in previous multimodal representations. The *modality-biased log-bilinear* (MLBL-B) model is set as the base model in the experiment.

***Index Terms***— multimodal representation, neural language model, Kneser-Ney smoothing, skip-gram, data sparsity

## 1. INTRODUCTION

Humans form meaningful perceptual experience by integrating information from different sensory modalities such as sight, sound, touch, smell and taste into coherent representation. For instance, the McGurk effect [6] demonstrates the integration of hearing and vision in speech perception. In fact, a concept or an idea is often delivered in both image and text for effective communication. An algorithm that can model the association between image and its text description would be conducive to retrieving an image by its text description and automatically generating a text description of a given image.

Data sparsity has been frequently problematic for the training of all sorts of neural language models, even with the availability of a large text corpus.

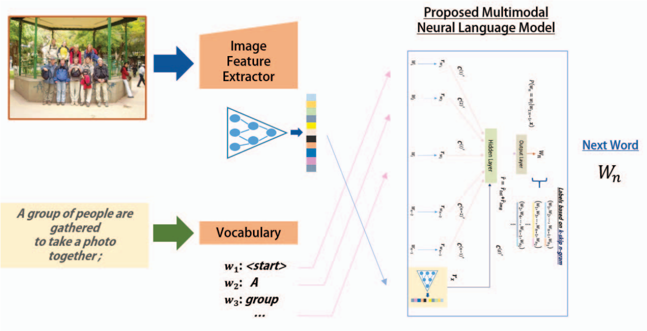


Image
Feature
Extractor
Proposed Multimodal
Neural Language Model
A group of people are gathered to take a photo together ;
Vocabulary
$w_1$: <start>
$w_2$: A
$w_3$: group
...
Hidden Layer
Output Layer
$P(w_n = w_i | w_{1:n-1}, x)$
$\hat{r} = \hat{r}_{txt} + \hat{r}_{img}$
Labels based on k-skip n-gram
Next Word
$W_n$

**Fig. 1**. An overview of our proposed model for the tasks of image retrieval and caption generation. The main framework is magnfied in Figure 3(b) for the skip-gram-based model but the Kneser-Ney-smoothing-based model in Figure 3(a) can be also equivalently referred.

To address the data sparsity problem, this paper considers a multimodal representation that is learned either by Kneser-Ney smoothing [2] or skip-gram [3] model, and it compares the performance with the representation learned with a generic $n$-gram [8].

Accordingly, the major contributions by this paper can be summarized as follows.

(1) *approaches to addressing data sparsity on the multimodal representation learned through neural language models.*

(2) *the refinement of the prediction in order that the contextual consistency of each sentence is strengthened and the correspondence of image and text modalities is enhanced.*

(3) *the use of state-of-the-art uni-modal representations for image and text for the multimodal representation: VGG-19 model [12] and Word2Vec [5] embedding.*

The rest of the paper consists of log-bilinear model (Section 2), modality-biased log-bilinear model (Section 3), proposed multimodal neural language models (Section 4), experiments (Section 5) and conclusion (Section 6).

978-1-4673-9961-6/16/$31.00 ©2016 IEEE

## 2. LOG-BILINEAR MODEL

The log-bilinear (LBL) language model proposed by Mnih and Hinton [9] is known as one of the simplest neural language models that predicts the $n$th word in a sentence given the context $n-1$ words. Its structure is equivalent to a feed-forward neural network with one linear hidden layer. Each of the words belonging to the vocabulary is converted into a distributed representation, i.e., $K$-dimensional real-valued vector $\mathbf{r}_w \in \mathbb{R}^{\mathbb{K}}$ referred to as word representation vector. For the context size of $n-1$, ($w_1$, ...,$w_{n-1}$) consists of a tuple of $n-1$ words and then the corresponding form in vector becomes ($\mathbf{r}_{w_1}$, ...,$\mathbf{r}_{w_{n-1}}$). From the intra-modal relationship depicted above, the LBL model predicts the next word representation $\hat{\mathbf{r}}$ by a linear combination of the context words:

$$\hat{\mathbf{r}} = \sum_{i=1}^{n-1} \mathbf{C}^{(i)} \mathbf{r}_{w_i}, \tag{1}$$

where $\mathbf{C}^{(i)}$ indicate $K \times K$ context parameter matrices for $i = 1, 2, ..., n-1$. Then, with the predicted representation, the conditional probability for the next $n$th word $P(w_n = w_l | w_{1:n-1})$ given preceding words $w_1, w_2, ..., w_{n-1}$ or *context* can be calculated as

$$P(w_n = w_l | w_{1:n-1}) = \frac{exp(\hat{\mathbf{r}}^T \mathbf{r}_{w_l} + b_{w_l})}{\sum_{j=1}^{V} exp(\hat{\mathbf{r}}^T \mathbf{r}_{w_j} + b_{w_j})}, \tag{2}$$

where $\mathbf{b} \in \mathbb{R}^V$ is a vector of word-specific bias $b_{w_l}$ for $l = 1, 2, ..., V$. We observe that the log-bilinear relation shown in Equation 2 is used to estimate the cosine similarity between the predicted next word representation $\hat{\mathbf{r}}$ and every word in the vocabulary $\mathbf{r}_{w_l}$ and return a score by softmax activation. Consequently, the parameters in the matrix are learned with the backpropagation algorithm.

## 3. MODALITY-BIASED LOG-BILINEAR MODEL (MLBL-B)

As an extended model of the LBL model, the modality-based log-bilinear model [1] associates multiple modalities while the former considers only one modality *e.g., text*.

The representation vector $\mathbf{r_x} \in \mathbb{R}^G$ *(G: dimension of features,* $\mathbf{x}$*: vector for features)* for another modality need be encoded on multimodal representation space, or equivalently shared knowledge space, where the corresponding representation vector for text modality co-exists. For bi-modalities, Equation 3 indicates a multimodal representation that intuitively incorporates multimodal data by involving an equal summation of uni-modal representations for image and text:

$$\hat{\mathbf{r}} = \hat{\mathbf{r}}_{\mathbf{txt}} + \hat{\mathbf{r}}_{\mathbf{img}} = \sum_{i=1}^{n-1} \mathbf{C}^{(i)} \mathbf{r}_{w_i} + \mathbf{C}^{(g)} \mathbf{r_x}, \tag{3}$$

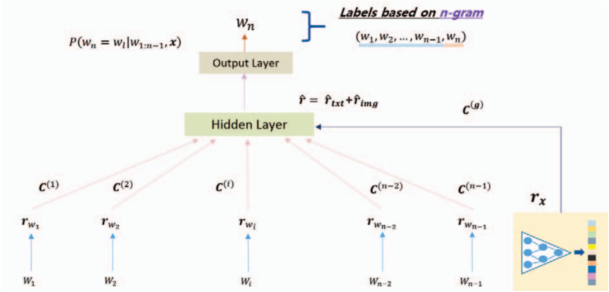


**Fig. 2**. An overview of the Modality-biased Log-bilinear (MLBL-B) model in Section 3 for image and text modalities proposed by Kiros, Salakhutdinov, and Zemel. [1]

where $\mathbf{C}^{(g)}$ is a $K \times G$ context parameter matrix for image in this case. Then, with the predicted representation, the conditional probability $P(w_n = l | w_{1:n-1}, \mathbf{x})$ of $w_n$ given the word context and image features can be calculated as

$$P(w_n = w_l | w_{1:n-1}, \mathbf{x}) = \frac{exp(\hat{\mathbf{r}}^T \mathbf{r}_{w_l} + b_{w_l})}{\sum_{j=1}^{V} exp(\hat{\mathbf{r}}^T \mathbf{r}_{w_j} + b_{w_j})}, \tag{4}$$

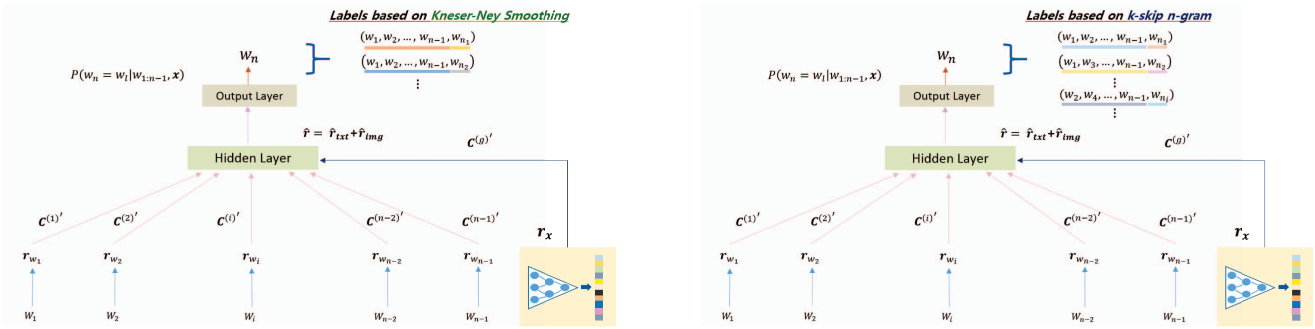
where the form of the conditional probability is not affected by $\mathbf{x}$ due to the predicted representation being taken into account for computation. In comparison to the LBL model, the MLBL-B model can be referred to as a feed-forward network with one additional channel from image modality and thus the backpropagation algorithm can also be used to calculate its weights, i.e., contextual matrices for image and text. However, we observe that data spasity and contextual inconsistency are anticipated as a result of learning and then incorrectly trained weights, on the grounds that the given labels based on statistically calculated $n$-grams are possibly insufficient in number or qualitatively coarse as in other multimodal neural language models.

## 4. PROPOSED MULTIMODAL NEURAL LANGUAGE MODELS

To address the data sparsity problem in constructing a neural language model with multimodal representation, the Kneser-Ney smoothing [2] and skip-gram [3] models can be separately integrated into the language model, and their performances are comparatively evaluated.

### 4.1. Kneser-Ney Smoothing based Multimodal Neural Language Model

The Kneser-Ney smoothing language model has demonstrated its effectiveness as a state-of-the-art method for the prediction of the next word based on the conditional probability. It is based on the observation that the conditional probability from lower-order $n$-gram labels should be considered provided that the conditional probability from high-order


Labels based on Kneser-Ney Smoothing
$(w_1, w_2, \ldots, w_{n-1}, w_{n_1})$
$(w_1, w_2, \ldots, w_{n-1}, w_{n_2})$
$\vdots$
$w_n$
$P(w_n = w_l | w_{1:n-1}, \boldsymbol{x})$
Output Layer
$\hat{\boldsymbol{r}} = \hat{\boldsymbol{r}}_{txt} + \hat{\boldsymbol{r}}_{img}$
$\boldsymbol{C}^{(g)'}$
Hidden Layer
$\boldsymbol{C}^{(1)'}$
$\boldsymbol{C}^{(2)'}$
$\boldsymbol{C}^{(i)'}$
$\boldsymbol{C}^{(n-2)'}$
$\boldsymbol{C}^{(n-1)'}$
$\boldsymbol{r}_x$
$\boldsymbol{r}_{w_1}$
$\boldsymbol{r}_{w_2}$
$\boldsymbol{r}_{w_i}$
$\boldsymbol{r}_{w_{n-2}}$
$\boldsymbol{r}_{w_{n-1}}$
$W_1$
$W_2$
$W_i$
$W_{n-2}$
$W_{n-1}$
Labels based on k-skip n-gram
$(w_1, w_2, \ldots, w_{n-1}, w_{n_1})$
$(w_1, w_3, \ldots, w_{n-1}, w_{n_2})$
$\vdots$
$(w_2, w_4, \ldots, w_{n-1}, w_{n_i})$
$\vdots$
$w_n$
$P(w_n = w_l | w_{1:n-1}, \boldsymbol{x})$
Output Layer
$\hat{\boldsymbol{r}} = \hat{\boldsymbol{r}}_{txt} + \hat{\boldsymbol{r}}_{img}$
$\boldsymbol{C}^{(g)'}$
Hidden Layer
$\boldsymbol{C}^{(1)'}$
$\boldsymbol{C}^{(2)'}$
$\boldsymbol{C}^{(i)'}$
$\boldsymbol{C}^{(n-2)'}$
$\boldsymbol{C}^{(n-1)'}$
$\boldsymbol{r}_x$
$\boldsymbol{r}_{w_1}$
$\boldsymbol{r}_{w_2}$
$\boldsymbol{r}_{w_i}$
$\boldsymbol{r}_{w_{n-2}}$
$\boldsymbol{r}_{w_{n-1}}$
$W_1$
$W_2$
$W_i$
$W_{n-2}$
$W_{n-1}$

(a)

(b)

**Fig. 3**. Proposed neural language models with multimodal representation based on Kneser-Ney smoothing (Figure 3(a)) and skip-gram (Figure 3(b)). Although the MLBL-B model is used as an example for both cases, our models can be applied to other general types of multimodal neural language models. The denotation for context matrices $\mathbf{C}^{(i)}$ and $\mathbf{C}^{(g)}$ is replaced with $\mathbf{C}^{(i)'}$ and $\mathbf{C}^{(g)'}$, respectively, to indicate that the updated weights are changed.

$n$-gram labels is inclined to be almost zero. Therefore, this can lead to selecting plausible labels and then reflecting extant information. Kneser-Ney smoothing involves absolute discounting by which a fixed value of $\delta \in [0, 1]$ is substracted from the count of a set of lower-order labels to limit the influence of $n$-grams with lower frequencies.

Equation 5 is the modified conditional probability given context using $n$-grams for multimodal data, *i.e., image and text*, where Kneser-Ney smoothing is applied.

$$P_{KN}(w_n|w_{1:n-1}, \mathbf{x}) = \frac{max(C(w_1, w_2, ..., w_{n-1}) - \delta, 0)}{\sum_{w_n \in V} C(w_1, w_2, ..., w_{n-1})} + \lambda(w_{1:n-1}) P_{KN}(w_n|w_{1:n-2}, \mathbf{x}), \quad (5)$$

$$\lambda(w_{1:n-1}) = \frac{\delta}{\sum_{w_n \in V} C(w_1, w_2, ..., w_{n-1})} \cdot N_{1+}(w_{1:n-1}\bullet), (6)$$

where $C(w_1, w_2, ..., w_{n-1})$ denotes the count of a set of $n-1$ words and $N_{1+}(w_{1:n-1}\bullet)$ is the non-overlapping number of a possible set of $n-1$ words before word $\bullet$. $\lambda(\cdot)$ is a smoothing parameter to be determined as in Equation 6.

Using Kneser-Ney smoothing, it is expected that adjusted labels from their coarse ones would be obtained by which the data sparsity problem is fairly resolved. Also, we may obtain different multiple labels from $n-1$ words, i.e., $w_{n_1}, w_{n_2}, ...$, as shown in Figure 3(a). Moreover, there will be more labels generated in the presence of multimodal data given a number of different image features.

### 4.2. Skip-gram based Multimodal Neural Language Model

As another approach, the skip-gram language model can be utilized to substitute $n$-gram in the calculation of a set of labels whereby $n$-grams organized with intermediate tokens being allowed to be *"skipped"* is considered. The term skip-gram is an abbreviation for $k$-skip-$n$-gram from which $n$-gram can be a particular case for $k = 0$. In other words, skip-grams are a generalization of $n$-grams.

The following mathematical expression in 7 describes the definition of skip-grams.

$$\{w_{s_1}, w_{s_2}, ..., w_{s_n} | \sum_{i=1}^{n} |s_i - s_{i-1}| < k\}, \quad (7)$$

where $w_{s_1}, w_{s_2}, ..., w_{s_n}$ is a set of $n$ words, $s_i$ indicates the $i$th index, and $k \geq 0$, a positive integer, such that $n > k$.

From above, we notice that a variety of sets of words can be generated using skip-grams such that the sum of indices of neighboring words is less than $k$. We observe that $n$-grams are automatically included in the generated sets and also sets for $k$ belong to those for $k + 1$.

Therefore, the conditional probability of the next word given skip-gram-based context for multimodal data, *i.e., image and text*, can be formulated as

$$P_S(w_n | \{w_{s_1}, w_{s_2}, ..., w_{s_{n-1}} | \sum_{i=1}^{n-1} |s_i - s_{i-1}| < k\}, \mathbf{x}), \quad (8)$$

With skip-gram applied, we are able to cope with the problem of data sparsity as the number of labels itself is expected to grow larger for a bigger $k$. The comparison of how different labels can be comprised between the two models is depicted in Figure 3(a) and Figure 3(b). In this model, labels, i.e., $w_{n_1}, w_{n_2}, ...$, are generated from different $k$-skip-$n$-grams. In this paper, we explore the case based on traditional skip-gram while a fixed value of $k$ is set for $n = 5$ and $k = 1$.

## 5. EXPERIMENTS

For our proposed models used for experiments, we named MLBL-KN and MLBL-Skip for the MLBL-B model with Kneser-Ney smoothing in 4.1 and MLBL-B with skip-gram 4.2, respectively. Note that all referred models for comparison in [1] are based on $n$-gram. We engage the popular IAPR TC-12 dataset for the evaluation of our proposed models.

IAPR TC-12 Benchmark [10] is a collection of 20,000 images with a text description up to three sentences. Visualized samples are available in the supplementary material.

### 5.1. Implementaion Details

For the training of the LBL based model, we have attempted to maintain the same hyper-parameters, e.g., the weight decays and learning rates for each modality as in [1] so as to compare the differences of our model and the previous multimodal language models. However, we have adopted a state-of-the-art CNN, VGG-19 [12] for image feature extraction, intended to constitute a better uni-modal representation for image modality, while we utilize Word2Vec [5] as a state-of-the-art word embedding for the same reason, as cited.

### 5.2. Evaluation Criteria

**Perplexity**

In natural language processing, perplexity is one of the most popular evaluation criteria to evaluate a language model. Low perplexity generally indicates satisfactory sentences. The uni-modal perplexity $\chi(w_{1:n})$ is modified to consider image modality which becomes:

$$\chi(w_{1:n}|\mathbf{x}) = 2^{-\frac{1}{N}\sum_{w_{1:n}} log_2 P(w_n=w_i|w_{1:n},\mathbf{x})}, \qquad (9)$$

where $N$ is the length of a sentence and $w_{1:n}$ directs subsequences of context $n-1$. For image retrieval, the image with the lowest perplexity $\chi(w_{1:n}|\mathbf{x})$ is selected but the perplexity need be normalized to ensure the query image being independent. Thus, the image is retrieved whose ratio of $\chi(w_{1:n}|\mathbf{x})/\chi(w_{1:n}|\bar{\mathbf{x}})$ ($\bar{\mathbf{x}}$: mean image in the training dateset) is lowest.

**BLEU-$n$ Scores**

BLEU *(Bilingual Evaluation Understudy)* is another evaluation criterion that is widely used to estimate the quality of text for a variety of tasks in natural langue processing. It demonstrates how statistically the results from human and machine translation are correspondent and then indicates the quality of a language model. A BLEU-$n$ score indicates that it uses the $n$-gram to calculate the precision.

### 5.3. Quantitative Results

Table 1 demonstrates the quantitative results of our MLBL-KN/MLBL-Skip models and previous $n$-gram based models of MLBL-B, MLBL-F [1] and LBL [9] on the IAPR TC-12 dataset. Some of the models involved different combinations of image features from popular models, most of which are CNN-based, such as *AlexNet* [11], *DeCAF* [13], *VGG-19* [12], *skmeans* [14], and word embeddings such as Turian [4] and Word2Vec [5] *(available in the supplementary material).*

| Model | Image Feature | PPL | B-1 | B-2 | B-3 |
|---|---|---|---|---|---|
| LBL† | - | 20.1 | 0.327 | 0.144 | 0.068 |
| MLBL-B† | skmeans [14] | 18.0 | 0.349 | 0.161 | 0.079 |
| MLBL-B† | AlexNet [11] | 20.6 | 0.349 | 0.165 | 0.085 |
| MLBL-F† | AlexNet [11] | 21.7 | 0.341 | 0.156 | 0.073 |
| **MLBL-KN** | **AlexNet** [11] | **21.4** | **0.361** | **0.175** | **0.095** |
| **MLBL-Skip** | **AlexNet** [11] | **19.9** | **0.354** | **0.173** | **0.093** |
| MLBL-B† | DeCAF [13] | 24.7 | 0.373 | 0.187 | 0.098 |
| MLBL-F† | DeCAF [13] | 21.8 | 0.361 | 0.176 | 0.092 |
| **MLBL-KN** | **DeCAF** [13] | **23.0** | **0.378** | **0.193** | **0.098** |
| **MLBL-Skip** | **DeCAF** [13] | **21.1** | **0.381** | **0.192** | **0.100** |
| MLBL-B† | VGG-19 [12] | 25.8 | 0.333 | 0.148 | 0.069 |
| MLBL-F† | VGG-19 [12] | 24.2 | 0.331 | 0.147 | 0.070 |
| **MLBL-KN** | **VGG-19** [12] | **25.6** | **0.358** | **0.176** | **0.089** |
| **MLBL-Skip** | **VGG-19** [12] | **24.2** | **0.350** | **0.166** | **0.085** |

**Table 1**. Results on IAPR TC-12 dataset [10]. According to the model type, image feature and word embedding (Turian [4] for above), the quantitative results are evaluated on perplexity (PPL) and BLEU-$n$ criteria where $n$-grams ($n$ = 1, 2, 3) are used as the references. The symbol † indicates that the values are taken from published results [1].

Analyzing the numerical results, we can first observe that the MLBL-Skip models generally achieve higher performance on perplexity *(lowered)* and BLEU-$n$ criteria *(highered)* than the other models while many of the MLBL-KN models improve on BLEU-$n$ criteria. Second, we find that different combinations of image feature and word embedding may affect the performance but the influence seems limited as that of the techniques for the learning of multimodal neural language models is more dominant.

## 6. CONCLUSION

We have proposed a multimodal representation that is learned to overcome the chronic problem of data sparsity when using neural language models. For the demonstration, we have utilized the modality-biased log-bilinear (MLBL-B) model [1] as an example.

Using our models on which Kneser-Ney smoothing and skip-gram are based, we have observed improved results on the widely used IAPR TC-12 [10] dataset, especially compared with the results from the previous models [1, 9]. We were also successful in the investigation of the possibility of extending our models to general neural language models as far as weights in these models are learned using $n$-gram based labels which indicates that the same problem of data sparsity is existent for the models as well.

# Acknowledgement

This work was partly supported by the Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.B0101-16-0307, Basic Software Research in Human-level Lifelong Machine Learning (Machine Learning Center) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2010-0028680).

## 7. REFERENCES

[1] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal Neural Language Models. In *International Conference on Machine Learning*, 2014.

[2] R. Kneser and H. Ney. Improved Backing-off for $m$-gram Language Modeling. In *International Conference on Acoustics, Speech and Signal Processing*, 1995.

[3] D. Guthrie et al. A Closer Look at Skip-gram Modelling. In *International Conference on Language Resources and Evaluation*, 1998.

[4] J. Turian, L. Ratinov, and Y. Bengio. Word Representations: A Simple and General Method for Semi-supervised Learning. In *Annual Meeting of the Association for Computational Linguistics*, 2010.

[5] T. Mikolov et al. Distributed Representations of Words and Phrases and Their Compositionality. In *Annual Conference on Neural Information Processing Systems*, 2013.

[6] H. McGurk and J. MacDonald. Hearing Lips and Seeing Voices. *Nature*, 264:746-748, 1976.

[7] Y. Bengio et al. A Neural Language Model. *Journal of Machine Learning Research*, 3:1137-1155, March 2003.

[8] P. Brown et al. Class-based $n$-gram Models of Natural Language. *Computational Linguistics*, 18: 467-479, April 1992.

[9] A. Mnih and G. Hinton. Three New Graphical Models for Statistical Language Modelling. In *International Conference on Machine Learning*, 2007.

[10] M. Grubinger et al. The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems. In *International Workshop OntoImage*, 2006.

[11] A. Krizhevsky et al. ImageNet Classification with Deep Convolutional Neural Networks. In *Annual Conference on Neural Information Processing Systems*, 2012.

[12] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015.

[13] J. Donahue et al. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *arXiv preprint arXiv:1310.1531*, 2013.

[14] R. Kiros and C, Szepesv$\acute{a}$ri. Deep Representations and Codes for Image Auto-annotation. In *Annual Conference on Neural Information Processing Systems*, 2012.