

Context Aware Machine Learning

Yun Zeng

1600 Amphitheatre Parkway, Mountain View
CA 94043, United States

XZENG@GOOGLE.COM

Editor:

Abstract

We propose a principle for exploring *context* in machine learning models. Starting with a simple assumption that each observation (random variables) may or may not depend on its context (conditional variables), a conditional probability distribution is decomposed into two parts: *context-free* and *context-sensitive*. Then by employing the log-linear word production model for relating random variables to their embedding space representation and making use of the convexity of natural exponential function, we show that the embedding of an observation can also be decomposed into a weighted sum of two vectors, representing its context-free and context-sensitive parts, respectively. This simple treatment of context provides a unified view of many existing deep learning models, leading to revisions of these models able to achieve significant performance boost. Specifically, our upgraded version of a recent sentence embedding model (Arora et al., 2017) not only outperforms the original one by a large margin, but also leads to a new, principled approach for compositing the embeddings of bag-of-words features, as well as a new architecture for modeling attention in deep neural networks. More surprisingly, our new principle provides a novel understanding of the gates and equations defined by the long short term memory (LSTM) model, which also leads to a new model that is able to converge significantly faster and achieve much lower prediction errors. Furthermore, our principle also inspires a new type of generic neural network layer that better resembles real biological neurons than the traditional linear mapping plus nonlinear activation based architecture. Its multi-layer extension provides a new principle for deep neural networks which subsumes residual network (ResNet) as its special case, and its extension to convolutional neural network model accounts for irrelevant input (*e.g.*, background in an image) in addition to filtering. Our models are validated through a series of benchmark datasets and we show that in many cases, simply replacing existing layers with our context-aware counterparts is sufficient to significantly improve the results.

Keywords: Context Aware Machine Learning, Sentence Embedding, Bag of Sparse Features Embedding, Attention Models, Recurrent Neural Network, Long Short Term Memory, Convolutional Neural Network, Residual Network.

1. Introduction

“A brain state can be regarded as a transient equilibrium condition, which contains all aspects of past history that are useful for future use. This definition is related but not identical to the equally difficult concept of *context*, which refers to a set of facts or circumstances that surround an event, situation, or object.”

— Gyoögy Buzsáki in *Rhythms of the Brain*

Most machine learning tasks can be formulated into inferring the likelihood of certain outcome (*e.g.*, labels) from given input or context (Bishop, 2006). Also most models assume such a likelihood is completely determined by its context, and therefore their relations can be approximated by functions (deterministic or nondeterministic) of the context with various approximate, parameterized forms (*e.g.*, neural networks), which in turn can be learned from the data. Nevertheless, such an assumption does not always hold true as real-world data are often full of irrelevant information and/or noises. Explicitly addressing the relevance of the context to the output has been shown to be very effective and sometimes game-changing (*e.g.*, Sepp Hochreiter (1997); Vaswani et al. (2017)). Most notably, three circuit-like gates are defined in the LSTM model (Sepp Hochreiter, 1997) to address the relevance of previous state/previous output/current input to current state/current output in each time step. Despite its tremendous success in the past two decades in modeling sequential data (Greff et al., 2016), LSTM’s heuristic construction of the neural network keeps inspiring people to search for better architectures or more solid theoretical foundations. In this work, we continue with such an effort by showing that not only better architectures exist, the whole context-aware treatment also squarely fits a unified probabilistic framework that allows us to connect with and revise many existing models.

Our work is primarily inspired by a sentence embedding algorithm proposed by Arora et al. (2017), where the embedding of a sentence (bag of words) is decomposed into a weighted sum of a *global vector*, computed as the first principal component on the embeddings of all training sentences, and a *local vector* that changes with the sentence input. Its decomposition is achieved via adding a word frequency term into the word-production model (Mnih and Hinton, 2007), connected by a weight (a global constant) that is chosen heuristically. Word frequency can take into account those common words (*e.g.*, “the”, “and”) that may appear anywhere, and Arora et al. (2017) showed that in the embedding space, the impact of those common/uninformative words can be reduced thanks to the new decomposition. Essentially, that work connects the idea of term frequency-inverse document frequency (TF-IDF) from information retrieval (Manning et al., 2008) to sentence embedding. Despite its promising results, the model is constructed based on a number of heuristic assumptions, such as assuming the weights that connects the global and local vectors be constant, as well as the use of principal component analysis for estimating the global vector.

To find a more solid theoretical ground for the above sentence embedding model, we start by revamping its heuristic assumptions using rigorous probability rules. Rather than directly adding a new term, *i.e.*, word frequency, into the word-production model, we introduce an indicator variable to the conditional part of the probability distribution that takes into account the extent to which the observation (*e.g.*, a word) is related to its given context (*e.g.*, a sentence or a topic). This *context-aware* variable allows us to decompose the original conditional probability into two terms: a context-free part and a context-sensitive part, connected by a weighting function (a probability distribution) determined by the input. Furthermore, by restricting the probability distribution be within the exponential family and exploiting its convexity property, a new equation is obtained, which we call the *Embedding Decomposition Formula* or **EDF**. It decomposes the embedding of an observation into two vectors: context-free and context-sensitive, connected by a weighting function, called the **χ -function** that denotes the level of context-freeness for each observation. EDF is powerful in its *simplicity* in form, its *universality* in covering all models that can be formulated with a conditional probability distribution and therefore its potential in revolutionizing many existing machine learning problems from this context-aware point of view.

Note that the concept of context-aware treatment for machine learning problems is hardly new. Besides language related models mentioned above, it has also been explored in image models both explicitly (*e.g.*, Zhang et al. (2009)) and implicitly (*e.g.*, Jegou et al. (2012)). However, our *contribution* here is a new formulation of such a common sense in the embedding space, based on established statistics theory and probability rules, which leads to a new, unified principle that can be applied to improving many existing models. Specifically, we revisit and revise the following problems in this paper:

- *Sentence embedding* (CA-SEM): The basic idea of the original paper by Arora et al. (2017) is to explore global information in a corpus of training sentences, using any already learned word embeddings. Our EDF inspires a new formulation by casting it into an energy minimization problem, which solves for the optimal decomposition of the original word embeddings into a context free part and a context sensitive part, connected by the χ -function that represents the broadness of each word. We show that it can be solved iteratively by a block-coordinate descent algorithm with a closed-form solution in each step. In our experiment, our new approach significantly outperforms the original model in all the benchmarks tasks we tested.
- *Bag of sparse features embedding* (CA-BSFE): In a neural network with sparse feature input, it is essential to first embed each input feature column to a vector before feeding it to the next layer. Sometimes the values for a feature column may be represented by bag-of-words, *e.g.*, words in a sentence or topics in a webpage. Prior art simply computes the embedding for that feature column by averaging the embeddings of its values. Our framework suggests a new architecture for computing the embeddings of bag-of-sparse-features in neural networks. Its setup automatically learns the importance of each feature values and out-of-vocabulary (OOV) values are automatically taken care of by a context-free vector learned during training.
- *Attention model* (CA-ATT): If attention can be considered as sensitivity to its context/input, our framework is then directly applicable to computing the attention vector used in sequence or image models. Different from the CA-BSFE model above that focuses on training a word-to-embedding lookup dictionary, our CA-ATT model focuses on the training of the χ -function (attention) from input context, corresponding to the softmax weighting function in existing attention models. Our theory implies normalization of the weight is not necessary, since the number of salient objects from the input should vary (*e.g.*, single face versus multiple faces in an image).
- *Sequence model* (CA-RNN): Our framework is able to provide a new explanation of the LSTM model. First of all, a sequence modeling problem can be formulated into inferring the current output and hidden state from previous output and hidden state, as well as current input. Then by applying standard probability rules the problem is divided into the inferences of the current output and hidden state, respectively. For each of the sub-problems, by applying our new framework, we find both the forget and the output gates defined in LSTM can find their correspondences as a χ -function defined in our model. This alignment also allows us to unify various existing sequence models and find the “correct” version that outperforms existing sequence models significantly.
- *Generic neural network layer* (CA-NN): Traditional linear mapping of input + nonlinear activation based NN layer fails to capture the change-of-fire-patterns phenomena, known as bifurcation, which is common in real neurons. Our framework partially addresses this limitation by allowing an NN layer to switch between two modes: context-free and context-sensitive, based on the contextual input. We show that the mappings learned by CA-NN layer possess different properties from a normal NN layer.
- *Convolutional neural network* (CA-CNN): As an application of the CA-NN architecture, it is straightforward to design a new convolutional neural network whose χ -function explicitly classifies each input fragment (*e.g.*, an image patch) into irrelevant (context free) background/noise or meaningful (context sensitive) foreground. Qualitative results on popular datasets such as MNIST illustrate that our model can almost perfectly tell informative foreground from irrelevant background.
- *Residual network layer* (CA-RES): By carefully modeling CA-NN in the multi-layer case using probability theory, we have discovered a new principle for building deep neural networks, which also subsumes the currently popular residual network (ResNet) as our special case. Our

new treatment regards a deep neural network as inter-connected neurons without any hierarchy before training, much like human brain in the early ages, and hierarchical information processing is only formed when sufficient data are learned, controlled by the χ -function in each CA-NN sublayer that turns it on/off when information is passed from its input to output.

The rest of this paper is organized as follows. We first establish the theoretical foundation of our new approach in Sec. 2. Then we connect our new framework to the problems listed above and also revise them into better forms in Sec. 3. In Sec. 4 the performance of our models are evaluated on various datasets and models. Related works on diverse fields covered in this paper are reviewed in Sec. 5, after which we conclude our work in Sec. 6.

2. Mathematical Formulation

We start by discussing the close relations among language modeling, exponential family and word embedding models in Sec. 2.1. Then in Sec. 2.2 we propose a novel framework for context-aware treatment of probabilistic models and introduce our EDF. In Sec. 2.3 we discuss its extension to embedding bag-of-features.

2.1 Language Modeling, Exponential Family and Embedding

Generally speaking, the task of statistical language modeling is to find the distribution of a set of words $P(w_1, \dots, w_n)$, $w_i \in \mathcal{W}$, $i = 1, 2, \dots, n$ Manning et al. (2008). In practice, it is usually more useful to model the distribution of a word $w \in \mathcal{W}$ from its given context $c \in \mathcal{C}$: $P(w|c)$, where context can be a sequence of words preceding w or any other relevant information that contributes to the production of w Bengio et al. (2003). Note that the same formulation can be applied to any other problems involving probability distributions with discrete variables.

To obtain a parameterized representation of $P(\cdot)$, one usually considers the exponential family that covers a rich set of distributions, *i.e.*, $P(c) \propto \exp(\langle \theta_1, \mathbf{T}_1(c) \rangle)$ and $P(w, c) \propto \exp(\langle \theta_2, \mathbf{T}_2(w, c) \rangle)$, where $\theta_i \in \mathbb{R}^d$, $i \in \{1, 2\}$, $\mathbf{T}_1 : \mathcal{C} \mapsto \mathbb{R}^d$ and $\mathbf{T}_2 : \mathcal{W} \times \mathcal{C} \mapsto \mathbb{R}^d$ map the input variables to a d -dimensional Euclidean space. By further assuming $\mathbf{T}_2(w, c) = \mathbf{T}_3(w) \odot \mathbf{T}_1(c)$ where $\mathbf{T}_3 : \mathcal{W} \mapsto \mathbb{R}^d$ and $\odot : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}^d$ is the pointwise multiplication, then from $P(w|c) = P(w, c)/P(c)$, we have:

$$P(w|c) \propto \exp(\langle \theta_2 \odot \mathbf{T}_3(w) - \theta_1, \mathbf{T}_1(c) \rangle) \equiv \exp(\langle \mathbf{W}(w), \mathbf{C}(c) \rangle), \quad (1)$$

which is known as the *log-linear word production model*. Here $\mathbf{W} : \mathcal{W} \mapsto \mathbb{R}^d$ and $\mathbf{C} : \mathcal{C} \mapsto \mathbb{R}^d$ can be considered as the *embeddings* for w and c , respectively. Note that Eq. 1 is based on very restricted assumption on the forms of exponential family distribution to decouple w and c . One can certainly consider the more general case that $\mathbf{T}_2(w, c)$ is not factorable, then $\mathbf{W}(\cdot)$ becomes a function on $\mathcal{W} \times \mathcal{C}$ and $\mathbf{C}(\cdot)$ remains the same. In the rest of this paper, we will denote the embedding of a variable v simply by \vec{v} , then Eq. 1 becomes

$$P(w|c) \propto \exp(\langle \vec{w}, \vec{c} \rangle). \quad (2)$$

Also note that the same model of Eq. 2 has been proposed under various interpretations in the past (*e.g.*, Welling et al. (2005); Mnih and Hinton (2007); Arora et al. (2016)). Here we favor the exponential family point of view as it can be easily generalized to constructing probability distributions with more complex forms (*e.g.*, Rudolph et al. (2016)), or with more than one conditional variables. For example, from $P(w|c, s) = P(w, c|s)/P(c|s)$, we may consider a general family of distributions:

$$P(w|c, s) \propto \exp(\langle \vec{w}(s), \vec{c}(s) \rangle), \quad (3)$$

where we use $\vec{w}(s)$ and $\vec{c}(s)$ to denote that the embeddings of w and c also depends on the value of s . We will show that this is very instrumental for our new treatment of many existing models.

2.2 Context Aware Embedding Model

Given the definition of the log-linear model, we are ready to introduce the core concept in this paper: *context free* and its negation, *context sensitive*.

Consider the case of a language model $P(w|c)$ when w is a word and c is a sentence as its context. When w is an uninformative word like {"is", "the"}, it may appear in any sentence and its frequency is independent of its context (context free). Therefore we can loosely define $P(w|c, w \text{ is context free}) = \tilde{P}(w)$, meaning it is independent of any context c .

Formally, if we introduce an indicator $CF(w)$ to denote if a variable w is context free or context sensitive, *i.e.*,

$$CF(w) = \begin{cases} 1, & \text{if } w \text{ is context free} \\ 0, & \text{if } w \text{ is context sensitive} \end{cases} \quad (4)$$

then we can quantitatively define the concept of context-freeness as follows.

Definition 1 (Context free indicator). A random indicator variable $CF : \mathcal{W} \mapsto \{0, 1\}$ is called context free indicator if and only if the following equation holds for $\forall c$

$$P(w|CF(w) = 1, c) = \tilde{P}(w) \quad (5)$$

Here $\tilde{P}(w)$ is a probability distribution that is not necessarily the same as $P(w)$ (word frequency).

Given the flexibility of the exponential family, we can always construct a probability distribution that satisfies the above equation. For example, any distribution with the form $P_\theta(w|CF(w), c) \propto \exp(CF(w)\langle \vec{w}_1, \theta \rangle + (1 - CF(w))\langle \vec{w}_2, \vec{c} \rangle)$, $\theta, \vec{w}_1, \vec{w}_2 \in \mathbb{R}^d$ would work. Note that the conditional distribution $P(w|CF(w) = 0, c)$ would become different from $P(w|c)$ in order for them to reside in a valid probability space.

To study the relation between $P(w|c)$ and $P(w|CF(w) = 0, c)$, let us combine Eq. 5 with standard probability rules to obtain

$$\begin{aligned} P(w|c) &= P(w, CF(w) = 1|c) + P(w, CF(w) = 0|c) \\ &= P(w|CF(w) = 1, c)P(CF(w) = 1|c) + P(w|CF(w) = 0, c)P(CF(w) = 0|c) \\ &= \tilde{P}(w)P(CF(w) = 1|c) + P(w|CF(w) = 0, c)(1 - P(CF(w) = 1|c)). \end{aligned} \quad (6)$$

If we denote by $\chi(w, c) \equiv P(CF(w) = 1|c)$, which we call the **χ -function** throughout this paper, the above equation then becomes

$$P(w|c) = \tilde{P}(w)\chi(w, c) + P(w|CF(w) = 0, c)(1 - \chi(w, c)), \quad (7)$$

where $\chi(w, c) \in [0, 1]$ indicates how likely w is independent of its context c , and the conditional probability $P(w|c)$ is decomposed into a weighted sum of two distributions, one denoting the context free part (*i.e.*, $\tilde{P}(w)$), another the context sensitive part (*i.e.*, $P(w|CF(w) = 0, c)$).

We can further connect Eq. 7 to its embedding space using the exponential family assumption discussed in Sec 2.1. First of all, let us assume $P(w|CF(w) = 0, c) = \exp(\langle \vec{w}', \vec{c} \rangle - \ln Z_c)$ where \vec{w}' is the embedding of w when it is context sensitive and Z_c is the partition function depending only on c . Then Eq. 7 becomes

$$\begin{aligned} P(w|c) &= \tilde{P}(w)\chi(w, c) + P(w|CF(w) = 0, c)(1 - \chi(w, c)) \\ &= \exp(\ln(\tilde{P}(w)))\chi(w, c) + \exp(\langle \vec{w}', \vec{c} \rangle - \ln Z_c)(1 - \chi(w, c)). \end{aligned} \quad (8)$$

Also from the convexity of the exponential function ($\mu e^a + (1 - \mu)e^b \geq e^{\mu a + (1 - \mu)b}$, $\mu \in [0, 1]$), we have (let $a = \ln(\tilde{P}(w))$, $b = \langle \vec{w}', \vec{c} \rangle - \ln Z_c$ and $\mu = \chi(w, c)$)

$$\begin{aligned} P(w|c) &\geq \exp(\chi(w, c) \ln \tilde{P}(w) + (1 - \chi(w, c))(\langle \vec{w}', \vec{c} \rangle - \ln Z_c)) \\ &= \frac{\exp(\chi(w, c)(1 + \frac{\ln \tilde{P}(w)}{\ln Z_c})\langle \vec{v}_c, \vec{c} \rangle + (1 - \chi(w, c))\langle \vec{w}', \vec{c} \rangle)}{Z_c} \\ &= \frac{\exp(\langle \chi(w, c)(1 + \frac{\ln \tilde{P}(w)}{\ln Z_c})\vec{v}_c + (1 - \chi(w, c))\vec{w}', \vec{c} \rangle)}{Z_c}. \end{aligned} \quad (9)$$

Here \vec{v}_c is any vector that satisfies

$$Z_c = \sum_w \exp \langle \vec{w}', \vec{c} \rangle = \exp(\langle \vec{v}_c, \vec{c} \rangle). \quad (10)$$

Intuitively, \vec{v}_c represents the partition function in the embedding space which summarizes the information on all \vec{w}' interacting with the same context c . The solution to Eq. 10 is not unique, which gives us the flexibility to impose additional constraints in algorithm design as we will elaborate in Sec. 3.

By comparing Eq. 9 with the original log-linear model of Eq. 2 in the general case that \vec{w} depends on both w and c , we can establish the following decomposition for the embedding of w :

$$\vec{w} \approx \chi(w, c)(1 + \frac{\ln \tilde{P}(w)}{\ln Z_c})\vec{v}_c + (1 - \chi(w, c))\vec{w}' \quad (11)$$

Furthermore, if we assume $\frac{\ln \tilde{P}(w)}{\ln Z_c} \approx 0$, a very simple embedding decomposition formula can be obtained as follows.

Theorem 1 (Embedding Decomposition Formula, a.k.a., EDF). *The log-linear model of a conditional probability $P(w|c)$ can be approximately decomposed into two parts: context sensitive and context free, connected by the $\chi(\cdot)$ function that indicates its level of context-freeness*

$$\vec{w} \approx \chi(w, c)\vec{v}_c + (1 - \chi(w, c))\vec{w}'. \quad (12)$$

Here \vec{v}_c represents the context-free part of an embedding, and \vec{w}' represents the context-sensitive part. This is because $P(w|c) \propto \exp(\langle \vec{w}, \vec{c} \rangle)$, and when $\chi(w, c) \rightarrow 1$, $P(w|c) \rightarrow \text{constant}$ (Eq. 10) which is insensitive to the observation w^1 . On the other hand, when $\chi(w, c) \rightarrow 0$, w 's embedding \vec{w}' represents something that could interact with its context, i.e., $\langle \vec{w}, \vec{c} \rangle$ changes with w , and therefore is sensitive to the context.

In practice, we may make reasonable assumptions such as $\chi(\cdot)$ is independent of c (e.g., “the” is likely to be context free no matter where it appears), or \vec{v}_c is orthogonal to \vec{w}' for all w (e.g., the oov tokens can appear in any sentence and their distribution should be independent of the context, i.e., $\langle \vec{w}, \vec{c} \rangle$ is a constant). We will discuss in details in Sec. 3 how these simplified assumptions can help reduce model complexity.

2.3 Bag-of-words Embedding

In the above we have only discussed the embedding model for a single variable w . It is straightforward to generalize it to the case when we want to find the embedding of a set of variables $s = \{w_1, \dots, w_n\}$.

1. This is related to how the out-of-vocabulary (oov) tokens are dealt with in neural network based language models, where an oov token may appear anywhere (context-free). The usual approach is to map them to a single vector, corresponding to v_c in our case.

We can combine the naive independence assumption and Eq. 9 to obtain an approximation of $P(s|c)$ as follows:

$$\begin{aligned}
P(s|c) &= \prod_i P(w_i|c) \\
&\geq \exp\left(\sum_i \langle \chi(w_i, c) \vec{v}_c (1 + \frac{\ln \tilde{P}(w_i)}{\ln Z_c}) + (1 - \chi(w_i, c)) \vec{w}'_i, \vec{c} \rangle\right) / Z_c^n \\
&\propto \exp\left(\sum_i \langle \chi(w_i, c) \vec{v}_c (1 + \frac{\ln \tilde{P}(w_i)}{\ln Z_c}) + (1 - \chi(w_i, c)) \vec{w}'_i, \vec{c} \rangle\right) \\
&\approx \exp(\langle \vec{v}_c \sum_i \chi(w_i, c) + \sum_i (1 - \chi(w_i, c)) \vec{w}'_i, \vec{c} \rangle)
\end{aligned} \tag{13}$$

By comparing it with the log-linear model $P(s|c) \propto \exp(\langle \vec{s}, \vec{c} \rangle)$, we obtain the *bag-of-features embedding formula*:

$$\vec{s} \approx \vec{v}_c \sum_i \chi(w_i, c) + \sum_i (1 - \chi(w_i, c)) \vec{w}'_i, \tag{14}$$

which is simply the sum of single-variable embeddings. This gives us a very straightforward way of compositing bag-of-feature embeddings beyond the averaging scheme $\vec{s} \approx \sum_i \vec{w}_i / n$ (e.g., Wieting et al. (2016)).

3. Existing Models Revisited and Revised

In this section, we explore the connections between our context-awareness based framework (Sec. 2) and various existing neural network models. Specifically, we demonstrate how our EDF (Eq. 12) can fit into various machine learning models and therefore provide a unified principle.

3.1 Sentence Re-embedding

The task of re-embedding is to retrain any existing (pre-trained) embedding mapping $(\mathbf{T} : \mathcal{W} \mapsto \mathbb{R}^d)$ for a new task to obtain a new mapping $\tilde{\mathbf{T}} : \mathcal{W} \mapsto \mathbb{R}^d$ that better fits the distribution of new data. This is useful when training new embeddings is infeasible due to lack of (labeled) data. Wieting et al. (2016) demonstrated that by simply averaging the (pretrained) embeddings of all the words in a sentence can already achieve better performance than sequence models. Later, Arora et al. (2017) further improved this bag-of-words based approach by introducing an additional term in the log-linear model to reduce the impact of those frequent and uninformative words, which is the primary inspiration for our work.

Specifically, in the work of Arora et al. (2017), the probability that a word w is produced in a sentence c (represented by a discourse vector, or embedding \vec{c}) is modeled by,

$$\begin{aligned}
P(w|c) &= \alpha P(w) + (1 - \alpha) \frac{\exp(\langle \vec{w}, \vec{c} \rangle)}{Z_c} \\
\text{where } \vec{c} &= \beta \vec{c}_0 + (1 - \beta) \vec{c}' \text{ and } \vec{c}_0 \perp \vec{c}'
\end{aligned} \tag{15}$$

Here α, β are scalar hyper-parameters that are globally constant, \vec{c}_0 is a global vector representing the common part of the embeddings (computed via PCA on all training sentences), and \vec{c}' changes with each sentence c .

Surprisingly or not, Eq. 7 resembles Eq. 15 in almost every way². It is easy to identify the rough correspondences $\chi(w, c) \leftrightarrow \alpha$, $\tilde{P}(w) \leftrightarrow P(w)$ and $P(w|CF(w) = 0, c) \leftrightarrow \frac{\exp(\langle \vec{w}, \vec{c} \rangle)}{Z_c}$ between them.

2. Note that Eq. 7 is derived using standard probability rules whereas Eq. 15 is heuristically defined.

However, despite their resemblance in appearance, the underlying assumptions are largely different as highlighted below.

1. $\chi(w, c)$ has a clear mathematical meaning ($P(CF(w)|c)$) under our context-awareness setting which changes with both w and c , whereas α is a global constant that is heuristically defined.
2. $\tilde{P}(w)$ in Eq. 7 is a different distribution (frequency of w when it is context free) from $P(w)$ (overall frequency of w). Otherwise it can be proven that there is no valid measurable space that satisfies Eq. 7. Hence our model does not involve overall word frequency.
3. Although \vec{c}_0 of Eq. 15 resembles \vec{v}_c in our case, the former is defined for each sentence and the latter is on the side of w . Besides, \vec{v}_c is defined for each value of c (Eq. 10), whereas \vec{c}_0 is a global constant.
4. There is no correspondence of β in our new model.
5. The log-linear model $\frac{\exp(\langle \vec{w}, \vec{c} \rangle)}{Z_c}$ in Eq. 15 is defined for the original distribution $P(w|c)$, which can lead to a contradictory equation $P(w|c) = P(w)$. In our case, we assume $P(w|CF(w) = 0, c) \propto \exp(\langle \vec{w}', \vec{c} \rangle)$ such that everything remains in a valid probability space.
6. The context c in Arora et al. (2017) represents the sentence that a word appears. This is confusing because if a word is already shown in its context, the probability $P(w|c)$ should be 1. In contrast, our treatment of context is flexible to represent any high-level topic underlying the sentence, or even a whole corpus of sentences.

Eq. 15 is solved via computing the first principle components of all sentence embeddings as \vec{c}_0 then remove them from each sentence to obtain \vec{w}' . In contrast, in the rest of this section, we show that the sentence re-embedding problem can be formulated as an energy minimization problem which no longer requires PCA to estimate the global information \vec{v}_c .

3.1.1 CONTEXT-AWARE SENTENCE RE-EMBEDDING (CA-SEM)

To find the “correct” solution for sentence re-embedding, we resort to the EDF (Eq. 12) and solve for a new embedding \vec{w}' for each w from a given corpus of sentences $S = \{s_1, \dots, s_n\}$ and pre-trained embedding \vec{w} . In other words, we would like to minimize the distance (*e.g.*, L2-norm) between \vec{w} and $\chi(w, c)\vec{v}_c + (1 - \chi(w, c))\vec{w}'$ over all words $w \in s_i, i = 1, \dots, n$. Once we have found the new embeddings $\vec{w}', w \in \mathcal{W}$, the embedding of any sentence can be computed based on the bag-of-features embedding composition of Eq. 14.

However, there are a number of ambiguities that needs to be clarified to make the problem well-defined. First, both $\chi(w, c)$ and \vec{v}_c change with the context c , so what exactly is the context for the re-embedding problem? The answer largely depends on the problem we want to solve. Regardless, we can always assume that all the sentences in S share some common topic (*e.g.*, about conversation, or movie review). This means we can be spared from having to explicitly model \vec{c} and just assume that $\chi(w, c)$ and \vec{v}_c do not change with c . Hence we only need to solve for a global context vector, denoted as \vec{v}_0 , and a weight $\tilde{\chi}(w) \in [0, 1]$ for each w^3 .

Second, the minimization problem is ill-posed if we do not impose constraints on \vec{w}' . For instance, a trivial solution $\vec{w}' = \vec{w}$ and $\tilde{\chi}(w) = 0, \forall w$ can always achieve a perfect but useless decomposition, in which case the sentence embedding (Eq. 14) becomes equivalent to summing the embeddings of words. To obtain more meaningful embeddings, we find the assumption $\vec{v}_0 \perp \vec{w}', \forall w$ used in 15 makes good theoretically sense under our new context-awareness framework. In fact, this assumption

3. Here we justified the use of a single global vector \vec{c}_0 in Eq. 15 with a new explanation. Moreover, our framework invites designs of new algorithms such as automatic discovery of hierarchical topics from given corpus: first compute the reembedding based on global context; then group the new embeddings, and discover the sub-context in each group, etc.

essentially normalizes the partition function of Eq. 10 to be a constant. With it, we can always guarantee that the solution is non-trivial.

In sum, the problem of sentence embedding can be cast as solving the following constrained energy minimization problem:

$$\begin{aligned} \min_{\vec{w}', \vec{v}_0, \tilde{\chi}(w) \in [0, 1]} & \sum_{s \in \mathcal{S}} \sum_{w \in s} (\vec{w} - \tilde{\chi}(w)\vec{v}_0 - (1 - \tilde{\chi}(w))\vec{w}')^2 \\ \text{s.t. } & \vec{v}_0 \perp \vec{w}', \tilde{\chi}(w) \in [0, 1], \forall w \end{aligned} \quad (16)$$

However, question remains how to efficiently optimize the above energy. It turns out that there exists an efficient block-coordinate descent algorithm such that each step can be solved in a closed form.

3.1.2 OPTIMIZATION ALGORITHM FOR CA-SEM

Let us first rearrange Eq. 16 by grouping similar terms. If we denote by n_w the number of times that w appears among sentences in \mathcal{S} , then Eq. 16 becomes

$$\begin{aligned} \min_{\vec{w}', \vec{v}_0, \tilde{\chi}(w) \in [0, 1]} & \sum_{w \in \mathcal{S}} n_w (\vec{w} - \tilde{\chi}(w)\vec{v}_0 - (1 - \tilde{\chi}(w))\vec{w}')^2 \\ \text{s.t. } & \vec{v}_0 \perp \vec{w}', \tilde{\chi}(w) \in [0, 1], \forall w \end{aligned} \quad (17)$$

This is a nonlinear optimization problem and if we solve it using gradient descent it can easily stuck at local minima. Fortunately, if we solve each of \vec{v}_0 , $\{\vec{w}'|w \in \mathcal{W}\}$ and $\{\tilde{\chi}(w)|w \in \mathcal{W}\}$ by fixing the other values in a block-coordinate descent algorithm, it turns out that each sub-problem can be solved efficiently in a closed form:

1. Solve \vec{v}_0 : If we fix the values of $\tilde{\chi}(w)$ and \vec{w}' , $\forall w$, the optimal value of \vec{v}_0 can be solved via a standard linear least squares minimization problem.
2. Solve \vec{w}' : If we fix the value of \vec{v}_0 , we can also obtain \vec{w}' by removing the component along \vec{v}_0 from \vec{w} to respect the orthogonal constraint in Eq. 16, namely $\vec{w}' = (I - \vec{v}_0\vec{v}_0^T/|\vec{v}_0|^2)\vec{w}$. This gives us a closed form solution for each \vec{w}' once we know the value of \vec{v}_0 .
3. Solve $\tilde{\chi}(w)$: The optimization problem of Eq. 17 can be divided into independent sub-problems if we fix the values of \vec{v}_0 and \vec{w}' , meaning we only need to solve for the optimal value of $\tilde{\chi}(w) \in [0, 1]$ for each w by minimizing $(\vec{w} - \tilde{\chi}(w)\vec{v}_0 - (1 - \tilde{\chi}(w))\vec{w}')^2$. Geometrically, this is equivalent to finding the closest point on the line segment connecting \vec{v}_0 and \vec{w}' to the point \vec{w} (Fig. 1), which can be easily solved in a closed form.

Note that both $\tilde{\chi}(w)$ and \vec{w}' can be computed in a closed form if we know the value of \vec{v}_0 , which is the only thing we need to store for new word/sentence embeddings. Alg. 1 summarizes our CA-SEM algorithm for re-embedding. After the re-embedding of each word w (*i.e.*, \vec{w}') is computed, a sentence embedding can be computed using the following equation:

$$\vec{s} = \vec{v}_0 \sum_{w \in s} \tilde{\chi}(w) + \sum_{w \in s} (1 - \tilde{\chi}(w))\vec{w}', \quad (18)$$

which represents the context-sensitive part of Eq. 14.

One practical concern is the out-of-vocabulary (oov) words that are in the training/inference sentences but not in the pre-trained word embedding dictionary. These oov tokens can appear anywhere and hence context-free, so we only need to map them to \vec{v}_0 and let $\tilde{\chi}(\text{oov}) = 1$. In an extreme case, when all the words in a sentence are oov, the embedding becomes $|s|\vec{v}_0$, whose direction implies its level of context-freeness and its norm implies its extent (number of oov's in the sentence).

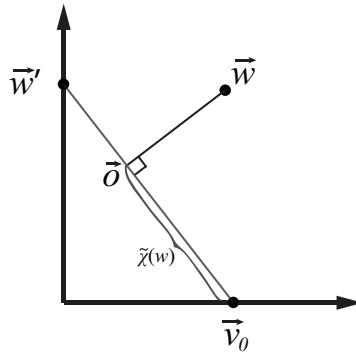


Figure 1: Illustration of solving the optimal $\tilde{\chi}(w)$ from \vec{w} , \vec{w}' and \vec{v}_0 in the 2D case. It first projects \vec{w} to the segment connecting \vec{w}' and \vec{v}_0 at \vec{o} , then the value of $\tilde{\chi}(w)$ can be computed as the ratio $|\vec{v}_0 - \vec{o}|/|\vec{w}' - \vec{v}_0|$.

Algorithm 1: Block-coordinate Algorithm for Sentence Reembedding.

Input : A set of sentences \mathcal{S} and a pre-trained embedding $\{\vec{w}|w \in \mathcal{W}\}$.

Output : The context vector $\vec{v}_0 \in \mathbb{R}^n$.

Initialization: Initialize \vec{v}_0 with the first principle component of all the points $\{\vec{w}|w \in s, s \in \mathcal{S}\}$.

Step 1: Fix \vec{v}_0 , solve for \vec{w}' and $\chi(w)$ for each w .

Step 2: Fix \vec{w}' and $\chi(w)$ for each w , solve for \vec{v}_0 .

Step 3: Evaluate the energy of Eq. 17 with new values, if it drops from previous iteration and maximal iteration is not reached, repeat Step 1; otherwise, stop.

3.2 Bag of Sparse Features Embedding (CA-BSFE)

Representing a set of sparse features (*e.g.*, words or categorical features) in a vector space is a prerequisite in neural network models involving sparse input. When the number of sparse feature varies with input (*i.e.*, bag of features), it is necessary to composite them into a single, fixed length vector to make sure that the architecture of an NN would remain fixed, a problem known as bag of sparse features embedding (BSFE) (or sentence/phrase embedding if the feature is a word). Compared to all the progresses made on the theoretical basis for embedding individual features (*e.g.*, (Bengio et al., 2003; Mnih and Hinton, 2007; Pennington et al., 2014; Arora et al., 2016)), fewer work has been done on the theoretical aspect of embedding bag of features (*e.g.*, (Mitchell and Lapata, 2010; Arora et al., 2018)). Previously, a de facto approach for compositing BSFE in NN is to map each sparse feature w to its embedding \vec{w} and the embedding is computed by their average $\sum_{w \in s} \vec{w}/|s|$, which is usually fed into the next layer in a deep network.

Our EDF (Eq. 12) provides a sound basis for compositing bag of sparse features embeddings that can be implemented in a neural network. We can design an NN layer that explicitly solves for $\chi(w, c)$, \vec{w}^e , and \vec{v}_c for each sparse feature w in the MED, then a BSFE can be computed by Eq. 14. Because neural networks are highly modularized such that model definition and training are well separated, here we only need to focus on the design of the BSFE layer and let back propagation algorithm take care of the training.

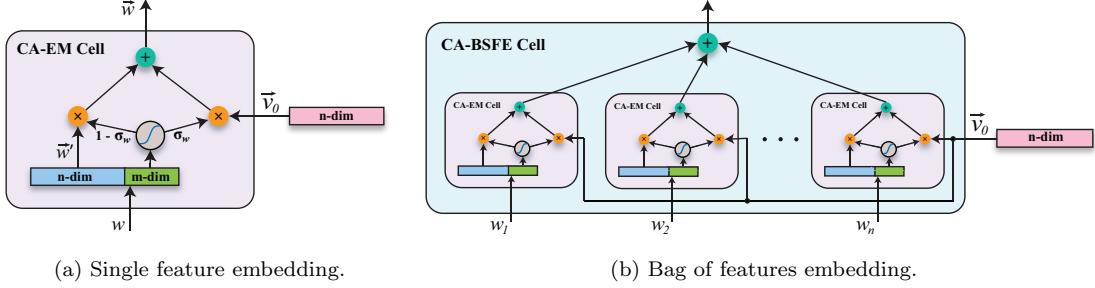


Figure 2: The architecture for single feature embedding (CA-EM) and multiple feature embedding (CA-BSFE). Here σ_w in (a) denotes the output of the χ -function defined in Eq. 19 and a global vector \vec{v}_0 is shared by all the input.

To start with, let us look at the meaning of context in BSFE. Similar to CA-SEM (Sec. 3.1), we can assume a global context \vec{v}_0 is shared among all the input. Therefore in the architecture of BSFE we need to solve for a global vector $\vec{v}_0 \in \mathbb{R}^n$ during training, where n is the embedding dimension.

The next thing to look at is the χ -function in our EDF which depends on both input feature w and context \vec{v}_0 . Because now \vec{v}_0 is a global variable that does not change with each input, the χ -function essentially becomes a function of w , namely $\tilde{\chi}(w)$. Also the role of χ -function is essentially a soft gating function that switches between \vec{v}_0 and \vec{w}' , similar to those gates defined in LSTM model Sepp Hochreiter (1997). Hence we simply need to employ the standard sigmoid function for its definition, which automatically guarantees that the value of χ -function are always within $[0, 1]$.

However, there is still one more obstacle to overcome before we can write down the equation for $\tilde{\chi}(w)$: the sigmoid function takes a real vector as its input whereas w is a sparse feature and it needs to be mapped to an embedding first. Here we have two choices: i) reuse the embedding vector \vec{w}' or ii) define a new embedding $\vec{w}_\sigma \in \mathbb{R}^m$ specifically for computing $\tilde{\chi}(w)$, where $m \in \mathbb{N}$ is not necessarily the same as the embedding dimension n . Theoretically speaking, ii) makes better sense since it decouples the embedding \vec{w}' and the χ -function. Otherwise, EDF is just a nonlinear function of \vec{w}' and the number of parameters of the sigmoid function would be constrained by n .

In sum, the χ -function can be implemented by the following sigmoid function:

$$\tilde{\chi}(w; \theta) = \frac{1}{1 + \exp(-\theta^T \vec{w}_\sigma)}, \quad (19)$$

where \vec{w}_σ is an m -dimensional vector associated to each w and θ is a global m -dimensional vector shared by all χ -functions. From Eq. 14, the embedding of a bag of features $s = \{w_1, w_2, \dots\}$ then becomes

$$\vec{s} = \vec{v}_0 \sum_{w \in s} \tilde{\chi}(w; \theta) + \sum_{w \in s} (1 - \tilde{\chi}(w; \theta)) \vec{w}. \quad (20)$$

Fig. 2 illustrates the architecture for single feature embedding (CA-EM) and BSFE embedding (CA-BSFE). Note that now the input embedding's dimension is $n + m$ and the model splits it into two parts for \vec{w}' and \vec{w}_σ , respectively.

Finally, let us connect our approach with the de facto, embedding lookup based network layer. If we define

$$\sigma(w) = \begin{cases} 0, & \text{if } w \in \mathcal{W} \\ 1, & \text{otherwise, or OOV} \end{cases} \quad (21)$$

and denote the embedding of OOV tokens as \vec{v}_0 , then the embedding of a sparse feature can be represented as

$$\vec{w} = \vec{v}_0\sigma(w) + (1 - \sigma(w))\vec{w}', \quad (22)$$

which is similar to our EDF in form, except that in our case, $\sigma(w) \in [0, 1]$. Hence our new model only provides a soft version of existing approaches, though it fundamentally changed the meaning of its variables.

3.3 Attention Model (CA-ATT)

The concept of attention used in neural networks (*e.g.*, Bahdanau et al. (2014)) is closely related to that of context, where attention decides which contextual input should be ignored or not. In a sequence model, the conditional distribution $P(w|c, s)$ is often considered, where w is the output (*e.g.*, y_t in a sequence-to-sequence model Sutskever et al. (2014) at stage t), c and s are different contexts for w (*e.g.*, in a sequence-to-sequence model $c = \{y_1, \dots, y_{t-1}\}$ is the output sequence before t and $s = \{x_1, \dots, x_T\}$ is the whole input sequence). In addition, attention is often applied when both the embeddings of c (*e.g.*, the hidden state of an RNN from previous steps, also known as the *query*) and each elements of s (*e.g.*, the hidden states of the whole input sequence, also known as the *memory*) are given. Intuitively, attention models the importance of input/memory $s = \{s_i | i = 1, \dots, T\}$ from given context/query c .

Therefore, we only need to draw our attention (pun intended) to the distribution $P(s|c)^4$, which is often represented by a vector \vec{s} (*a.k.a.*, attention vector) in a sequence model. From the EDF for bag of features (Eq. 14), the embedding of s can be represented as

$$\vec{s} = \vec{v}(\vec{c}) \sum_{i=1}^n \chi(\vec{s}_i, \vec{c}) + \sum_{i=1}^n (1 - \chi(\vec{s}_i, \vec{c})) \vec{s}_i, \quad (23)$$

where \vec{c} and $\vec{s}_i, i \in \{1, \dots, n\}$ are given as input, and we only need to define the forms for $\vec{v}(\cdot)$ and $\chi(\cdot, \cdot)$ to obtain \vec{s} . Here the value $1 - \chi(\vec{s}_i, \vec{c}), i = 1, \dots, n$ represents the importance of input s_i given the context c , which translates to context-sensitive-ness in our context-aware framework. Compared to the embedding models in Sec. 3.1 and 3.2, our attention model no longer assumes a global context and therefore the χ -function also depends on the context input \vec{c} . Our new architecture for attention model can be illustrated in Fig. 3.

Before discussing the definitions of $\vec{v}(\cdot)$ and $\chi(\cdot, \cdot)$, let us compare our context aware attention model (CA-ATT) with the original one (Bahdanau et al. (2014)).

1. The original model composites \vec{s} with an equation $\vec{s} = \sum_i \alpha(\vec{s}_i, \vec{c}) \vec{s}_i$ where

$$\alpha(\vec{s}_i, \vec{c}) = \frac{\exp(f(\vec{s}_i, \vec{c}))}{\sum_i \exp(f(\vec{s}_i, \vec{c}))}. \quad (24)$$

In contrast, we replace $\alpha(\vec{s}_i, \vec{c})$ with $1 - \chi(\vec{s}_i, \vec{c})$, which has a clear meaning: the context-sensitive-ness of s_i given the context c . However, one major difference is $\alpha(\cdot, \cdot)$ is normalized as a softmax distribution, which implicitly assumes only one of the input $\{s_i | i = 1, \dots, n\}$ should be attended to. Whereas in our case, we employ the more general assumption that the number of important input can be arbitrary.

2. Another major difference between our model and the original one is the introduction of $\vec{v}(\cdot)$ to account for the situation when none of the input is relevant. In such a case, for the original
4. The full context-aware treatment of the the conditional probability $P(w|s, c)$ will be given in the next section.

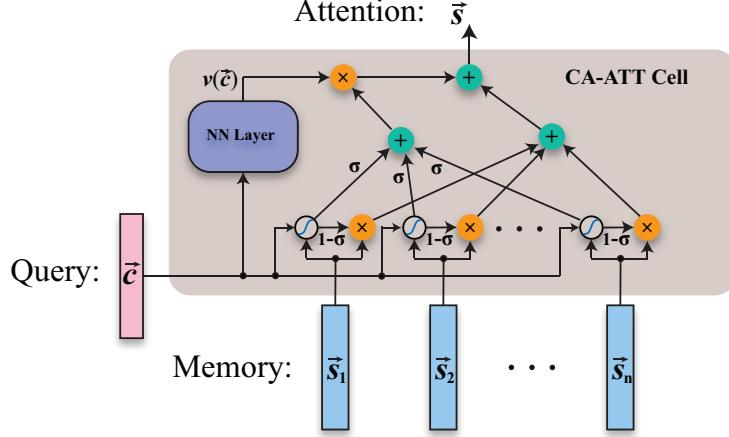


Figure 3: The architecture for context aware attention network (CA-ATT).

model, $f(\vec{s}_i, \vec{c}) \rightarrow 0$ and $\alpha(\vec{s}_i, \vec{c})$ would approach $1/|s|$, and \vec{s} is simply the average of the input embeddings, which still allows the information of s to be contributable to the final outcome of w . In contrast, we use a separate vector $v(\vec{c})$ to represent the state when the output is independent of the input (context-free). Hence when $1 - \chi(\vec{s}_i, \vec{c}) \rightarrow 0, i = 1, \dots, T$, the influence of s is eliminated (Eq. 23).

Now let us elaborate on the details of $\vec{v}(\vec{c})$, which is equivalent to defining a new layer of activation from previous layer's input \vec{c} . A general form of $\vec{v}(\cdot)$ can be represented as

$$\vec{v}(\vec{c}; \mathbf{W}_h, \mathbf{b}_h) = \sigma_h(\mathbf{W}_h \vec{c} + \mathbf{b}_h), \quad (25)$$

where σ_h can be any activation function such as hyperbolic tangent $\tanh(\cdot)$, $\mathbf{W}_h \in \mathbb{R}^{\dim(\vec{v}) \times \dim(\vec{c})}$ and $\mathbf{b}_h \in \mathbb{R}^{\dim(\vec{v})}$ are the kernel and bias, respectively, where $\dim(\cdot)$ represents the dimension of an input vector.

As for $\chi(\vec{s}_i, \vec{c})$, a straightforward choice is the sigmoid function:

$$\chi(\vec{s}_i, \vec{c}; \mathbf{w}_g, \mathbf{u}_g, b_g) = \sigma_g(\mathbf{w}_g \vec{s}_i + \mathbf{u}_g \vec{c} + b_g), \quad (26)$$

where σ_g is the sigmoid function, $\mathbf{w}_g \in \mathbb{R}^{\dim(\vec{s})}$, $\mathbf{u}_g \in \mathbb{R}^{\dim(\vec{c})}$ and $b_g \in \mathbb{R}$. Note that here we map each (\vec{s}_i, \vec{c}) to a scalar value within $[0, 1]$, which denotes the opposite of attention, *i.e.*, context-freeness of each input s_i .

Finally, we would like to add a note that our framework also implies many possibilities of implementing a neural network with attention mechanism. For example, in the CA-BSFE model (Sec. 3.2), we have proposed to use another m -dim embedding for the input to the χ -function, which allows the attention to be learned independent of the input embedding \vec{s}_i . We can certainly employ the same scheme as an alternative implementation.

3.4 Recurrent Neural Network (CA-RNN)

Now let us deal with some relatively complex sequence models. Generally speaking, in a RNN problem, the task is to infer the hidden state c_t and output y_t from previous hidden state c_{t-1} , previous output y_{t-1} and current input x_t , which can also be represented as a conditional probability $P(c_t, y_t | c_{t-1}, y_{t-1}, x_t), t = 1, 2, \dots$

The first step to solve this problem under our context-aware setting is to factorize the conditional probability into two parts: *cell state* and *output*, namely

$$P(c_t, y_t | c_{t-1}, y_{t-1}, x_t) = \underbrace{P(c_t | c_{t-1}, y_{t-1}, x_t)}_{\text{cell state}} \underbrace{P(y_t | c_t, c_{t-1}, y_{t-1}, x_t)}_{\text{output}}. \quad (27)$$

As shown in Table 1, it turns out that both the LSTM model (Sepp Hochreiter (1997)) and the GRU model (Cho et al. (2014)) fall into solving the above sub-problems in the embedding space using different assumptions (circuits).

	Cell State: $P(c_t c_{t-1}, y_{t-1}, x_t)$	Output: $P(y_t c_t, c_{t-1}, y_{t-1}, x_t)$
LSTM	$\bar{z}_t = \mathbf{W}_z x_t + \mathbf{R}_z y_{t-1} + \mathbf{b}_z$ $z_t = g(\bar{z}_t)$ - block input $\bar{i}_t = \mathbf{W}_i x_t + \mathbf{R}_i y_{t-1} + \mathbf{p}_i \odot c_{t-1} + \mathbf{b}_i$ $i_t = \sigma(\bar{i}_t)$ - input gate $\bar{f}_t = \mathbf{W}_f x_t + \mathbf{R}_f y_{t-1} + \mathbf{p}_f \odot c_{t-1} + \mathbf{b}_f$ $f_t = \sigma(\bar{f}_t)$ - forget gate $c_t = z_t \odot i_t + c_{t-1} \odot f_t$	$\bar{o}_t = \mathbf{W}_o x_t + \mathbf{R}_o y_{t-1} + \mathbf{p}_o \odot c_t + \mathbf{b}_o$ $o_t = \sigma(\bar{o}_t)$ - output gate $y_t = h(c_t) \odot o_t$
GRU	$\bar{r}_t = \mathbf{W}_r x_t + \mathbf{U}_r c_{t-1}$ - reset gate $r_t = \sigma(\bar{r}_t)$ $\bar{z}_t = \mathbf{W}_z x_t + \mathbf{U}_z c_{t-1}$ $z_t = \sigma(\bar{z}_t)$ - update gate $\bar{c}_t = \tanh(\mathbf{W}_x x_t + \mathbf{U}(r \odot c_{t-1}))$ $c_t = z_t \odot c_{t-1} + (1 - z_t) \bar{c}_t$	$y_t = h(\mathbf{W}_y c_t + \mathbf{b}_y)$

Table 1: A comparison between LSTM and GRU. Here $\sigma(\cdot)$ is the sigmoid function and $g(\cdot)$ and $h(\cdot)$ are activation functions such as $\tanh(\cdot)$.

To reformulate the RNN problem under our context-aware framework, let us first recall the EDF under an additional conditional variable s in a distribution $P(w|c, s)$ (Sec. 2.1). In such a case, we can still consider c as the context of interest, and EDF simply decomposes an embedding into one part that depends both on c and s (*i.e.*, $\vec{v}(c, s)$) and one that does not depend on c (*i.e.*, $\vec{w}'(s)$):

$$\vec{w} = \chi(w, c, s) \vec{v}(c, s) + (1 - \chi(w, c, s)) \vec{w}'(s) \quad (28)$$

Because w is conditioned on s^5 , $\chi(w, c, s)$ actually depends only on c and s . Hence we can simplify the above formula into

$$\vec{w} = \chi(c, s) \vec{v}(c, s) + (1 - \chi(c, s)) \vec{w}'(s) \quad (29)$$

To shed a light on how Eq. 29 is connected to the existing models listed in Table 1, one can consider the following correspondences between $P(w|c, s)$ and the cell state distribution $P(c_t | c_{t-1}, y_{t-1}, x_t)$.

$$\begin{aligned} w &\leftrightarrow c_t \\ c &\leftrightarrow c_{t-1} \\ s &\leftrightarrow \{y_{t-1}, x_t\} \end{aligned} \quad (30)$$

Then it is not difficult to verify that all the gating equations defined in Table 1 (*i.e.*, input and forget gates for LSTM and reset and upgrade gates for GRU) fall into the same form $\chi(c_{t-1}, y_{t-1}, x_t)$,

5. Note that here we only consider c as the context, and s is a variable that both w and c are conditioned on.

corresponding to $\chi(c, s)$ in Eq. 29. In the following, we explore the deep connections between Eq. 29 and those RNN models while searching for the “correct” version of RNN model based on the powerful EDF.

3.4.1 MODEL CELL STATE

Now let us consider the representation of $P(c_t|c_{t-1}, y_{t-1}, x_t)$ using the EDF of Eq. 29 from the correspondences given by Eq. 30. By substituting Eq. 30 into Eq. 29, we obtain the following embedding decomposition:

$$\vec{c}_t = \chi(c_{t-1}, y_{t-1}, x_t) \vec{v}(c_{t-1}, y_{t-1}, x_t) + (1 - \chi(c_{t-1}, y_{t-1}, x_t)) \vec{c}'_t(y_{t-1}, x_t), \quad (31)$$

where $\vec{v}(c_{t-1}, y_{t-1}, x_t)$ denotes the context-free part of \vec{c}_t and $\vec{c}'_t(y_{t-1}, x_t)$ denotes the context-aware part of \vec{c}_t . Comparing it with the cell state equation of LSTM (*i.e.*, $c_t = z_t \odot i_t + c_{t-1} \odot f_t$), one may find a rough correspondence below

$$\begin{aligned} \chi(c_{t-1}, y_{t-1}, x_t) &\leftrightarrow f_t \\ \vec{v}(c_{t-1}, y_{t-1}, x_t) &\leftrightarrow c_{t-1} \\ 1 - \chi(c_{t-1}, y_{t-1}, x_t) &\leftrightarrow i_t \\ \vec{c}'_t(y_{t-1}, x_t) &\leftrightarrow z_t \end{aligned} \quad (32)$$

Furthermore, Eq. 31 provides us with some insights into the original LSTM equation.

1. Under our framework, the gate functions f_t and i_t no longer map to a vector (same dimension as c_t), but rather to a scalar, which has a clear meaning: the probability that c_t is independent of its context c_{t-1} conditioned on $\{y_{t-1}, x_t\}$.
2. The correspondences established in Eq. 32 suggest that the forget gate and the input gate are coupled and should sum to one ($f_t + i_t = 1$). This may not make good sense in the LSTM model since it amounts to choosing either c_{t-1} (previous state) or z_t (current input). But if one looks at the counterpart of c_{t-1} , namely $\vec{v}(c_{t-1}, y_{t-1}, x_t)$, it already includes all the current (x_t) and past (c_{t-1}, y_{t-1}) information. Therefore, we can safely drop one gate in LSTM as well as reduce their dimension to one.
3. Both $\vec{c}'_t(y_{t-1}, x_t)$ in our framework and z_t in LSTM share the same form: the embedding of c_t without the information from c_{t-1} . However, we call $\vec{c}'_t(y_{t-1}, x_t)$ the context-sensitive vector as when interacted with the context vector in the log-linear model, it changes with c_{t-1} . Hence our framework endows new meanings to the terms in LSTM that makes mathematical sense.

Similarly for the GRU model listed in Table 1, we can also connect it to our context-aware framework by considering the following correspondences:

$$\begin{aligned} \chi(c_{t-1}, y_{t-1}, x_t) &\leftrightarrow z_t \\ \vec{v}(c_{t-1}, y_{t-1}, x_t) &\leftrightarrow c_{t-1} \\ \vec{c}'_t(y_{t-1}, x_t) &\leftrightarrow \bar{c}_t \end{aligned} \quad (33)$$

At a high level, the GRU equation $c_t = z_t \odot c_{t-1} + (1 - z_t) \bar{c}_t$ resembles our EDF in appearance. However, \bar{c}_t is not independent of c_{t-1} , contrary to our assumption. To remedy this, the GRU model employs the reset gate r_t to potentially filter out information from c_{t-1} , which is not necessary in our case. Also GRU model assume the gate functions are vectors rather than scalars.

Given the general form of context-aware cell state equation 31, one can choose any activation functions for its implementation. Following similar notations used in LSTM equations, we can write down a new implementation of the RNN model, namely CA-RNN, as follows.

$$\begin{aligned}\chi(c_{t-1}, y_{t-1}, x_t) : \quad & f_t = \sigma(\mathbf{v}_f^T \vec{c}_{t-1} + \mathbf{w}_f^T \vec{x}_t + \mathbf{u}_f^T \vec{y}_{t-1} + b_f) \\ \vec{v}(c_{t-1}, y_{t-1}, x_t) : \quad & \vec{v}_t = h_v(\mathbf{W}_v \vec{x}_t + \mathbf{U}_v \vec{y}_{t-1} + \mathbf{p}_v \odot \vec{c}_{t-1} + \mathbf{b}_v) \\ \vec{c}'(y_{t-1}, x_t) : \quad & \vec{c}'_t = h_c(\mathbf{W}_c \vec{x}_t + \mathbf{U}_c \vec{y}_{t-1} + \mathbf{b}_c) \\ \vec{c}_t : \quad & \vec{c}_t = f_t \vec{v}_t + (1 - f_t) \vec{c}'_t\end{aligned}\tag{34}$$

Here σ is the sigmoid function with parameters $\{\mathbf{v}_f \in \mathbb{R}^{\text{dim}(\vec{c}_t)}, \mathbf{w}_f \in \mathbb{R}^{\text{dim}(\vec{x}_t)}, \mathbf{u}_f \in \mathbb{R}^{\text{dim}(\vec{y}_t)}, b_f \in \mathbb{R}\}$, h_v and h_c are activation functions (*e.g.*, tanh, relu), with parameters $\{\mathbf{W}_v \in \mathbb{R}^{\text{dim}(\vec{c}_t) \times \text{dim}(\vec{x}_t)}, \mathbf{U}_v \in \mathbb{R}^{\text{dim}(\vec{c}_t) \times \text{dim}(\vec{y}_t)}, \mathbf{b}_v \in \mathbb{R}^{\text{dim}(\vec{c}_t)}\}$ and $\{\mathbf{W}_c \in \mathbb{R}^{\text{dim}(\vec{c}_t) \times \text{dim}(\vec{x}_t)}, \mathbf{U}_c \in \mathbb{R}^{\text{dim}(\vec{c}_t) \times \text{dim}(\vec{y}_t)}, \mathbf{b}_c \in \mathbb{R}^{\text{dim}(\vec{c}_t)}\}$.

3.4.2 MODEL OUTPUT

In Table 1, the equations for the output y_t look much simpler than those for the cell state for both LSTM and GRU models. However, in our case, it is actually more complex as it models a distribution involving more random variables: $P(y_t|c_t, c_{t-1}, y_{t-1}, x_t)$. Let us first write down the corresponding EDF, assuming the context of interest is c_t :

$$\vec{y}_t = \chi(c_t, c_{t-1}, y_{t-1}, x_t) v(c_t, c_{t-1}, y_{t-1}, x_t) + (1 - \chi(c_t, c_{t-1}, y_{t-1}, x_t)) \vec{y}'_t(c_{t-1}, y_{t-1}, x_t)\tag{35}$$

From the output equations for LSTM ($y_t = h(c_t) \odot o_t$) and GRU ($y_t = h(\mathbf{W}_y c_t + \mathbf{b}_y)$), one can see that both only consider the context-free part of Eq. 35. The LSTM model uses a single output gate o_t , meaning the output would be turned off (*i.e.*, $y_t \rightarrow 0$), if $o_t \rightarrow 0$. However, when LSTM is trained (usually y_t is used as the activation of a softmax layer), the output can hardly be matched to a near zero vector target. Hence o_t as an output gate may not really have any real affect. In contrast, our formulation of Eq. 35 allows the (soft) switch between two different modes: context free and context sensitive, which models a more realistic situation. Note that Eq. 35 also subsumes the GRU output equation as its special case if we set $\chi(c_t, c_{t-1}, y_{t-1}, x_t) = 1$ and $v(c_t, c_{t-1}, y_{t-1}, x_t) = h(\mathbf{W}_y c_t + \mathbf{b}_y)$.

Again, following the notation of LSTM equations, we can write down one implementation of our CA-RNN model for the output as follows.

$$\begin{aligned}\chi(c_t, c_{t-1}, y_{t-1}, x_t) : \quad & o_t = \sigma(\mathbf{z}_o^T \vec{c}_t + \mathbf{v}_o^T \vec{c}_{t-1} + \mathbf{w}_o^T \vec{x}_t + \mathbf{u}_o^T \vec{y}_{t-1} + b_o) \\ v(c_t, c_{t-1}, y_{t-1}, x_t) : \quad & \vec{v}_t^o = h_v^o(\mathbf{Z}_v \vec{c}_t + \mathbf{V}_v^o \vec{c}_{t-1} + \mathbf{W}_v^o \vec{x}_t + \mathbf{U}_v^o \vec{y}_{t-1} + \mathbf{b}_v^o) \\ \vec{y}'_t(c_{t-1}, y_{t-1}, x_t) : \quad & \vec{y}'_t^o = h_c^o(\mathbf{V}_c^o \vec{c}_{t-1} + \mathbf{W}_c^o \vec{x}_t + \mathbf{U}_c^o \vec{y}_{t-1} + \mathbf{b}_c^o) \\ \vec{y}_t : \quad & \vec{y}_t = o_t \vec{v}_t^o + (1 - o_t) \vec{y}'_t^o\end{aligned}\tag{36}$$

Here the superscript o indicates that the variables are for the output, and $\{\mathbf{z}_o \in \mathbb{R}^{\text{dim}(\vec{c}_t)}, \mathbf{v}_o \in \mathbb{R}^{\text{dim}(\vec{c}_t)}, \mathbf{w}_o \in \mathbb{R}^{\text{dim}(\vec{x}_t)}, \mathbf{u}_o \in \mathbb{R}^{\text{dim}(\vec{y}_t)}, b_o \in \mathbb{R}\}$, $\{\mathbf{Z}_v \in \mathbb{R}^{\text{dim}(\vec{c}_t) \times \text{dim}(\vec{c}_t)}, \mathbf{V}_v^o \in \mathbb{R}^{\text{dim}(\vec{c}_t) \times \text{dim}(\vec{c}_t)}, \mathbf{W}_v^o \in \mathbb{R}^{|\mathcal{C}| \times \text{dim}(\vec{x}_t)}, \mathbf{U}_v^o \in \mathbb{R}^{\text{dim}(\vec{c}_t) \times \text{dim}(\vec{y}_t)}, \mathbf{b}_v^o \in \mathbb{R}^{\text{dim}(\vec{c}_t)}\}$, $\{\mathbf{V}_c^o \in \mathbb{R}^{\text{dim}(\vec{c}_t) \times \text{dim}(\vec{c}_t)}, \mathbf{W}_c^o \in \mathbb{R}^{\text{dim}(\vec{c}_t) \times \text{dim}(\vec{x}_t)}, \mathbf{U}_c^o \in \mathbb{R}^{\text{dim}(\vec{c}_t) \times \text{dim}(\vec{y}_t)}, \mathbf{b}_c^o \in \mathbb{R}^{\text{dim}(\vec{c}_t)}\}$.

In sum, the architecture of our CA-RNN model can be illustrated in Fig. 4. Comparing it with the circuits of the LSTM model Greff et al. (2016), the number of gates is reduced from three to two. Its symmetric structure also makes the model more comprehensible.

Finally, note that our model invites many possible implementations beyond Eq. 34 and Eq. 36. For example, one can make a simpler assumption that $P(y_t|c_t, c_{t-1}, y_{t-1}, x_t) = P(y_t|c_t, y_{t-1}, x_t)$ as \vec{y}_{t-1} is a function of \vec{c}_{t-1} and other variables.

The model complexity (*i.e.*, number of parameters) of CA-RNN depends on the implementation so there is no fair comparison with other models. Nevertheless, the gate functions (*i.e.*, χ -function) of CA-RNN reduces the output dimension from n to 1, where n is the dimension of cell state.

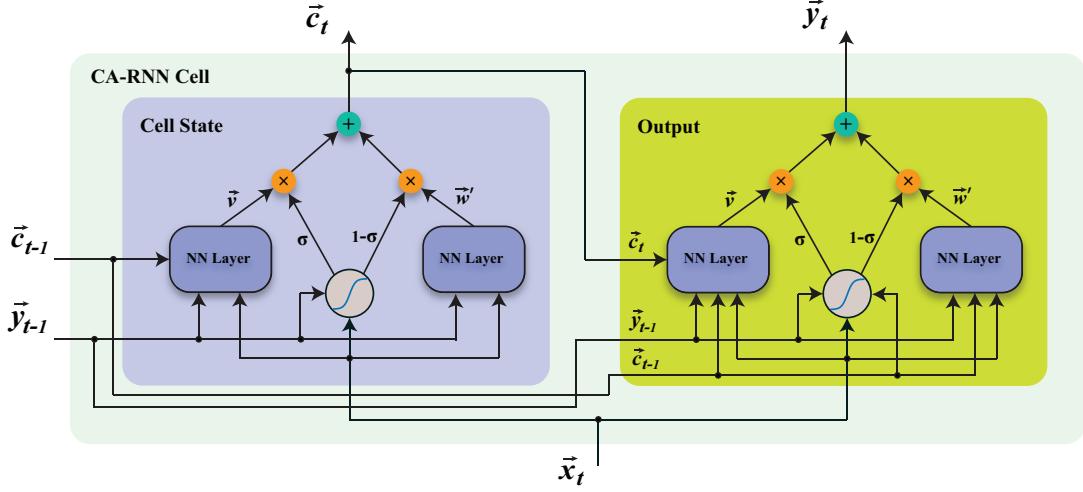


Figure 4: The architecture of CA-RNN model.

3.5 Generic Neural Network Layer (CA-NN)

Next, let us design a generic neural network layer based on the EDF. An NN layer simply takes the output from a lower level layer as input, and output an n-dimensional embedding. If we let c denote the input and w denote the output, an NN layer can be considered as modeling the probability $P(w|c)$. Hence, from our EDF of Eq. 12, the embedding (output) \vec{w} can be represented as

$$\vec{w} = \chi(\vec{w}_0, \vec{c})v(\vec{c}) + (1 - \chi(\vec{w}_0, \vec{c}))\vec{w}_0. \quad (37)$$

Here $v(\vec{c})$ is a function of the input signal \vec{c} that maps to the context-free part of the output; \vec{w}_0 is a vector representing the “default value” of the output when the input is irrelevant; The χ -function is determined both by the context and the default value, and because the default value \vec{w}_0 is a constant after training, it is just a function of \vec{c} . Hence we can simplify Eq. 37 as

$$\vec{w} = \chi(\vec{c})v(\vec{c}) + (1 - \chi(\vec{c}))\vec{w}_0. \quad (38)$$

Compared to a traditional NN layer, which simply assumes $\chi(\vec{c}) = 1$, our model allows the input to be conditionally blocked based on its value⁶.

We argue that our new structure of an NN layer make better neurological sense than the traditional NN layers. A biological neuron can be considered as a nonlinear dynamical system on its membrane potential Izhikevich (2007). The communication among neurons are either based on their fire patterns (resonate-and-fire) or on a single spike (integrate-and-fire). Moreover, a neuron’s fire pattern may change with the input current or context, *e.g.*, from excitable to oscillatory, a phenomena known as *bifurcation*. Such a change of fire pattern is partially enabled by the *gates* on the neuron’s membrane which control the conductances of the surrounding ions (*e.g.*, Na^+). Traditional NN layers under the perceptron framework (*i.e.*, linear mapping + nonlinear activation) fail to address such bistable states of neurons, and assumes a neuron is a simple excitable system with a fixed threshold, which is disputed by evidences in neuroscience (Chapters 1.1.2 and 7.2.4 in Izhikevich (2007)). Our CA-NN model (Fig. 5), in contrast, shares more resemblance with a real neuron and is able to model its bistable property. For instance, the probability of the ion channel

6. Note that this model is different from our single feature embedding model (Fig. 2 (a)) which seeks to computes an embedding with a global constant context.

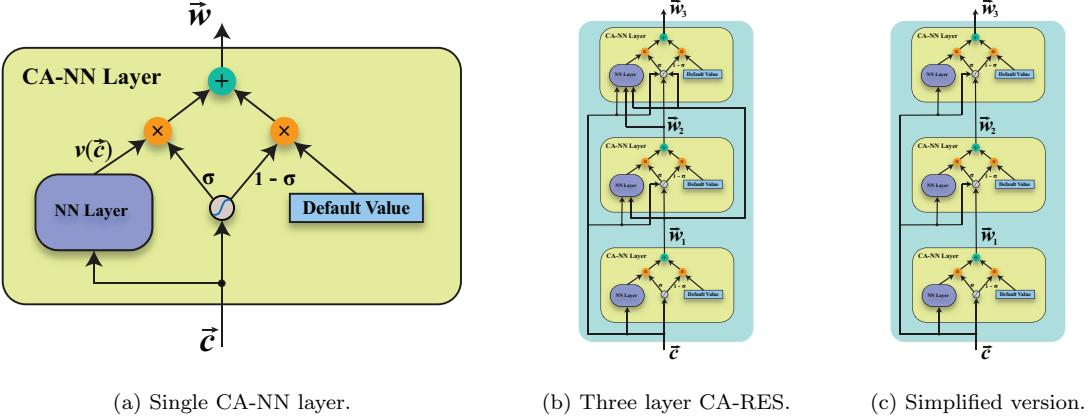


Figure 5: The architecture for a single layer CA-NN (a), its multi-layer extension CA-RES (b) and a simplified version of 3-layer CA-RES (c) that resembles ResNet.

gates to be open or close are determined by the current potential of the membrane represented by a neuron's extracellular state, which resembles our χ -function. Here the extracellular (external) state is modeled by \vec{c} , *i.e.* the contextual input.

To implement a CA-NN model, one can choose any existing NN layer for $v(\vec{c})$, and the χ -function can just be modeled using the sigmoid function, *i.e.*,

$$\chi(\vec{c}) = \sigma(\mathbf{v}^T \vec{c} + b), \quad (39)$$

where $\{\mathbf{v} \in \mathbb{R}^{dim(\vec{c})}, b \in \mathbb{R}\}$ and $\sigma(\cdot)$ is the sigmoid function. Fig. 5(a) shows the architecture of our CA-NN model.

Our CA-NN framework invites many possible implementations, depending on how the context is chosen. In the rest of this paper, we discuss its two straightforward extensions, one resembles the residual network (ResNet) He et al. (2015) and another the convolutional neural network (CNN) LeCun et al. (1990).

3.5.1 CONTEXT AWARE CONVOLUTIONAL NEURAL NETWORK (CA-CNN)

It is straightforward to modify our CA-NN model to serve as a CNN model LeCun et al. (1990), by just applying it on a patch of the input as filters. For example, in the linear case, we can define $v(\vec{c})$ in Eq. 38 as a linear transformation

$$v(\vec{c}) = \mathbf{W}\vec{c} + \mathbf{b}, \quad (40)$$

where $\mathbf{W} \in \mathbb{R}^{n \times dim(\vec{c})}$, $\mathbf{b} \in \mathbb{R}^n$ and n is the depth of the output. The presence of χ -function and $\vec{w}_0 \in \mathbb{R}^n$ in Eq. 38 allows irrelevant input (*e.g.*, background in an image) to be detected and replaced with a constant value. Therefore, CA-CNN model can be considered as convolution (CNN) with attention. Fig. 6 illustrates the architecture of a basic CA-CNN layer from a single layer CA-NN model. In a multilayer convolutional network, one can certainly employ the CA-RES architecture (Sec. 3.5.2) as the convolution kernel.

3.5.2 CONTEXT AWARE RESIDUAL NETWORK (CA-RES)

Lastly, let us look at how we can build a multilayer network using CA-NN layers as basic building blocks. Simply stacking multiple CA-NN layers on top of one another and using the output from

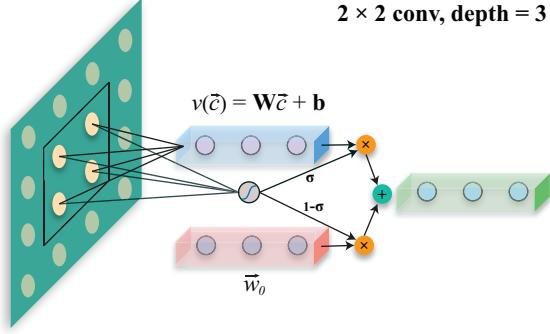


Figure 6: The architecture for context aware convolutional neural network (CA-CNN).

lower layer as context may not be a good idea because once the input to one layer is blocked, the information can no longer be passed to deeper layers⁷. However, if we look at the multiple layer network problem in a probabilistic setting, it can be actually formulated as solving the distribution $P(w_1, w_2, \dots, w_n | c)$ where $w_i, i = 1, 2, \dots, n$ are the output from different layers and $n > 1$ is the total layer number, which can be factorized as

$$P(w_1, w_2, \dots, w_n | c) = P(w_1 | c)P(w_2 | c, w_1) \dots P(w_n | c, w_{n-1}, \dots, w_1), n > 1 \quad (41)$$

A traditional multi-layer network simply assumes $P(w_l | c, w_{l-1}, \dots, w_1) = P(w_l | w_{l-1}), l > 1$, i.e., the output at layer l only depends on the output of lower layer $l-1$ ⁸. To construct a multi-layer CA-NN network, we have to use the full model $P(w_l | c, w_{l-1}, \dots, w_1)$ for each layer l and by assuming the context to be $\{c, w_{l-1}, \dots, w_1\}$ in Eq. 38, we have the following decomposition:

$$\vec{w}_l = \chi(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_1)\vec{v}_l(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_1) + (1 - \chi(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_1))\vec{w}_{l0}, l > 1. \quad (42)$$

Here \vec{w}_{l0} is the default value vector at level l . $\vec{w}_{l-1}, \dots, \vec{w}_1$ are the output from previous layers. $\chi(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_1)$ can be any activation function. Fig. 5(b) shows an example of 3-layer CA-NN network.

The number of parameters for the full version of our multi-layer NN model would grow squarely with the number of layers, which may not be desirable in a deep neural network. One can certainly make simplified assumptions, e.g., $\chi(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_1) = \chi(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_{l-k}), k > 1$. In an overly simplified case, if we assume $\vec{v}_l(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_1) = \vec{v}_l(\vec{c})$ (Fig. 5(c)) and $\chi(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_1) = 1$, it becomes the popular residual network (ResNet) He et al. (2015). In essence, both our multi-layer CA-NN and ResNet are able to skip underlying nonlinear NN layers, hence we also name our multi-layer CA-NN models CA-RES⁹.

In fact, our construction of multi-layer network provides some insights into existing models and we argue that it also better resembles a real neural network. First, contrary to a superficial view

-
7. Although this can be overcome by allowing the default value vector to change with input \vec{c} as proposed in Srivastava et al. (2015), it falls out of our framework and we are unable to find a sound theoretical explanation for such a heuristic.
 8. For comparison, a deep belief network Hinton et al. (2006) rather considers a deep network in the opposite direction: from deepest layer n to the observation c , which is represented by $P(c, w_1, w_2, \dots, w_n) = P(c|w^1) \prod_{k=1}^{n-2} P(w^k|w^{k+1}) P(w^{n-1}, w^n)$. It assumes the states of c and $w_i, i = 1, \dots, n$ are represented by the binary states of several neurons, whereas we assume the state of each layer (a single neuron) can be represented in a continuous embedding space.
 9. Despite their resemblance in structure, ResNet and CA-RES are derived from very different assumptions. We find the assumptions behind ResNet, decomposing a mapping $\mathcal{H}(x)$ into an identity mapping x and a residual mapping $\mathcal{F} = \mathcal{H}(x) - x$, may not fit well to our context-aware framework.

that an artificial neural network (ANN) is simply a nonlinear mapping from input to output, it actually models the internal states $\{w_i | i = 1, \dots, n\}$ from given context c . Also each internal state w_i can be represented in the embedding space if we assume the distribution $P(w_l | c, w_{l-1}, \dots, w_1)$ belongs to the exponential family.

Second, it is known that a real neural network evolves from almost fully connected to sparsely connected as a human grows from childhood to adulthood. Our CA-RES architecture is actually able to model this dynamic process in principle. The role of default value in each layer has the effect of blocking certain input from lower layers to all its upper layers, therefore it can reduce the effective connections among different neurons as more training data are seen. Note that the ResNet also achieves similar effect, but without a gate to explicitly switch on/off a layer therefore not adaptive.

3.6 Connections to other models

Before moving on to the experimental results, we would like to conclude this section by discussing how our context-aware framework can be further connected to other machine learning models.

It is possible to extend our framework to multiple contexts, in which case the χ -function can be modeled as a softmax function. For example, the mixture of experts model Shazeer et al. (2017) can be improved by accounting for the situation that no expert’s output is relevant. However, this is out of the scope of this paper.

Our CA-RES architecture can also be a good candidate for multitask learning (*e.g.*, Collobert and Weston (2008); Ruder (2017)). By wiring any of the output w_l from different layers (Sec. 3.5.2) to fit different targets (*e.g.*, Szegedy et al. (2015)), in principle, the χ -functions in each layer could play the role of switching to different networks for different tasks.

4. Experimental Results

Before delving into the details of the experiments for each of our context-aware models proposed in Sec. 3, it is helpful to compare them side-by-side to highlight their differences and relationships (Table. 2).

- Except for the CA-SEM model, all the other ones can be implemented as a neural network layer such that the optimization can be done via back propagation algorithm.
- Both CA-SEM and CA-BSFE assume a global context. Therefore the trained result is only an embedding dictionary.
- CA-RES is the multi-layer extension of CA-NN and CA-CNN is the single layer extension of CA-NN to convolutional network. CA-RES can also be extended to convolutional network.

A neural network model usually involves many layers of various types, and its overall performance is determined by not only the choice of individual layers, but also other hyper-parameters like initialization, batch size, learning rate, etc. The focus of this paper is not on the design of complex neural network architectures to achieve state-of-the-art performance on benchmark datasets (*e.g.* Krizhevsky et al. (2012); Vaswani et al. (2017)), but rather on the superiority of our proposed layers over their “traditional” counterparts, which can be validated by replacing the layers in various existing models with ours while keeping all other parameters unchanged to see the performance gain.

Dictionary-free embedding: another practical concern that can affect the performance of a model involving sparse features is the selection of a dictionary/vocabulary for embedding and all other sparse values are mapped to a special “unknown” or “oov” embedding. Such a dictionary is often selected based on the frequencies of the sparse features shown in the training data. We can eliminate this model tuning option simply by allowing every input features be mapped to its own embedding. Therefore in our experiments there is no such concept of “dictionary” for us to process.

Model	Form of EDF applied	Variables/functions to be learned
CA-SEM (Sec. 3.1)	$\vec{w} = \tilde{\chi}(w)\vec{v}_0 + (1 - \tilde{\chi}(w))\vec{w}'$	$\vec{v}_0, \tilde{\chi}(w), \vec{w}', w \in \mathcal{W}$
CA-BSFE (Sec. 3.2)	$\vec{s} = \vec{v}_0 \sum_{w \in s} \tilde{\chi}(w; \theta) + \sum_{w \in s} (1 - \tilde{\chi}(w; \theta))\vec{w}$	$\vec{v}_0, \tilde{\chi}(\cdot; \theta), \vec{w}, w \in \mathcal{W}$
CA-ATT (Sec. 3.3)	$\vec{s} = \vec{v}(\vec{c}) \sum_{i=1}^n \chi(\vec{s}_i, \vec{c}) + \sum_{i=1}^n (1 - \chi(\vec{s}_i, \vec{c}))\vec{s}_i$	$v(\cdot), \chi(\cdot, \cdot)$
CA-RNN (Sec. 3.4)	$\vec{w} = \chi(c, s)\vec{v}(c, s) + (1 - \chi(c, s))\vec{w}'(s)$	$\chi(\cdot, \cdot), \vec{v}(\cdot, \cdot), \vec{w}'(\cdot)$
CA-NN (Sec. 3.5)	$\vec{w} = \chi(\vec{c})v(\vec{c}) + (1 - \chi(\vec{c}))\vec{w}_0$	$\chi(\cdot), v(\cdot), \vec{w}_0$
CA-CNN (Sec. 3.5.1)	$\vec{w} = \chi(\vec{c})v(\vec{c}) + (1 - \chi(\vec{c}))\vec{w}_0$	$\chi(\cdot), v(\cdot), \vec{w}_0$
CA-RES (Sec. 3.5.2)	$\vec{w}_l = \chi(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_1)\vec{v}_l(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_1) + (1 - \chi(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_1))\vec{w}_{l0}$	$\chi(\cdot, \dots), v(\cdot, \dots), \vec{w}_{l0}$

Table 2: Comparison among all the models proposed in Sec. 3.

All the neural network models are implemented using the Tensorflow library Abadi et al. (2016) that supports distributed computation. For the public dataset tested in this paper, we usually deploy 2 – 100 workers without GPU acceleration. Below are general descriptions of the datasets that we use to evaluate our models.

The **SemEval** semantic textual similarity (STS) dataset Cer et al. (2017) consists of 5749 training and 1379 testing examples for comparing the semantic similarity (a number in [0, 5]) between two sentences, and the data is grouped by year from 2012 to 2017. The data contains both a testing file and a training file labeled by year.

The **SICK-2014** dataset Marelli et al. (2014) is designed to capture semantic similarities among sentences based on common knowledge, as opposed to relying on specific domain knowledge. Its annotations for each sentence pair include both a similarity score within [0, 5] and an entailment label: neutral, entailment, contradiction. The data consists of 4500 sentence pairs for training and 4927 pairs for testing.

The **SST** (Stanford Sentiment Treebank) Socher et al. (2013) is a dataset of movie reviews which was annotated for 5 levels of sentiment: strong negative, negative, neutral, positive, and strong positive. Training 67349 and testing 1821 examples.

The **IMDB** movie review dataset Maas et al. (2011) contains 50,000 movie reviews with an even number of positive and negative labels. The data is split into 25,000 for training and 25,000 for testing.

The **MNIST** dataset Lecun et al. (1998) contains 60,000 training and 10,000 testing grey-level images each normalized to size 28×28 .

The **CIFAR-10** dataset Krizhevsky and Hinton (2009) contains 60,000 color images in 10 classes, with 6000 images per class and each image is normalized to size 32×32 . The data is divided into 50,000 for training and 10,000 for testing.

The **WMT** 2016 dataset contains machine translation tasks among a wide range of languages. Its training data size is sufficient large to be considered useful for real-world applications. For example, its English-German dataset consists of 4,519,003 sentence pairs. Its testing data is relatively small, around 2,000 for each year (2009 – 2016).

4.1 Results for CA-SEM

Recall that our CA-SEM model solves the sentence re-embedding problem by a block-coordinate algorithm that minimizes Eq. 17. Once we have solved the global context vector \vec{v}_0 , all other variables, namely $\sigma(w)$ and w' , $\forall w \in \mathcal{W}$, can be obtained via a closed form solution (Sec. 3.1). Therefore \vec{v}_0 contains all the information we need for a re-embedding which makes model serving fast and simple.

One very useful application of CA-SEM is *text search*: find closest text snippets in a pre-defined corpus (*e.g.*, articles in news website or text in books) from given text query in a natural language form¹⁰. In such a case, one can pre-compute the embeddings of all the sentences in the given corpus from any pre-trained word embedding based on our re-embedding algorithm. In our experiments, our algorithms easily outperform that of Arora et al. (2017) on all datasets we tried (SemEval, SST, and several Google’s internal datasets) by a significant large margin.

4.1.1 IMPLEMENTATION DETAILS

We implemented both the algorithm described in Arora et al. (2017) and our method using C++. To handle large dataset (*e.g.*, $> 100M$ sentences), we also implemented a distributed version of our algorithm using standard distributed programming model that is similar to MapReduce Dean and Ghemawat (2004). In addition, the original embedding dictionary may not be restricted to single words (*e.g.*, “united”, “states”) but also contains phrases (*e.g.*, “united states”). In such a case, there exists multiple ways of segmenting a sentence into phrases and we select the segmentation with least number of phrases and OOV tokens (*e.g.*, to segment “united state” with a dictionary of {“united”, “states”, “united state”}, a segmentation of {“united state”} is preferable than {“united”, “states”}), which can be implemented by a fast dynamic programming algorithm.

In our implementation of Alg. 1, \vec{v}_0 is initialized with the first principal component computed from all the word/phrase embeddings from all the sentences in the training corpus. Then the energy minimization can be optimized iteratively until the energy of Eq. 17 no longer decreases between two iterations. We observed that for small datasets like SemEval, the energy can keep decreasing after hundreds of iterations (Fig. 7(a), but for large datasets with more than 10M sentences, the algorithm usually stops within 10 – 20 iterations. As for runtime, it takes less than 2 seconds to finish 100 iterations for small datasets like SemEval on a machine equipped with Intel(R) Xeon(R) 3.50GHz CPU with 12 cores, and for large datasets it usually takes less than 1 hour with 10 distributed machines.

Both the SemEval-STS and SICK datasets contain training instances that come with pairs of sentences. Whereas sentence reembedding only needs instances of sentences that reflect the distribution of interested text corpus. Hence we simply assume each training pair gives us two independent sentences, which then doubles the size of training instances from the original dataset.

For the pre-trained embedding dictionary as input to our algorithm, we employ the publicly available GloVe embedding Pennington et al. (2014) with dimension size 200. To evaluate the similarity between two sentences, we compute the cosine distance between their sentence embeddings. Similar to Arora et al. (2017), we compare with the ground truth score based on the Pearson’s coefficient $\times 100$ (Pearson’s $r \times 100$).

4.1.2 SEMEVAL-STS DATASET

Here we run Alg. 1 on the training data for each year and apply it to the corresponding testing data. Fig. 7 (a) shows the change in average loss (total energy of Eq. 17 / number of training examples)

10. One advantage of this bag-of-words embedding over sequence model based embedding is speed in serving, since computing an embedding vector for a sentence only requires a few lookups and fast re-embedding computation.

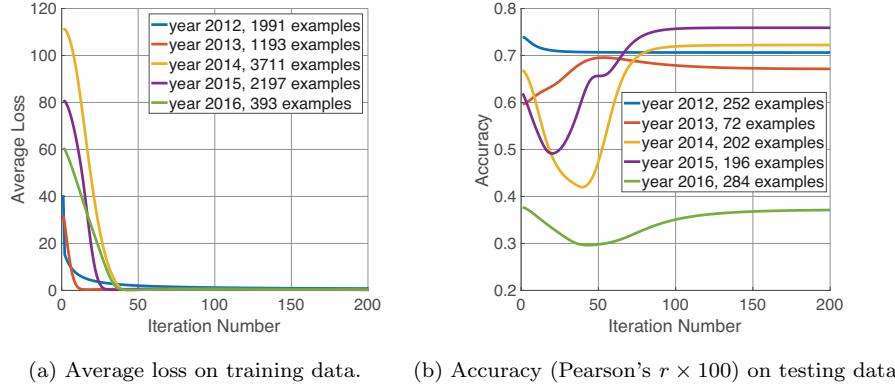
(a) Average loss on training data. (b) Accuracy (Pearson's $r \times 100$) on testing data.

Figure 7: (Best viewed in color) Loss and accuracy on STS dataset.

in each iteration. Fig. 7 (b) shows the corresponding accuracy computed over the corresponding testing data, which does not always increase but stabilizes after 100 iterations.

Table 3 summarizes the comparisons among different approaches.

	GloVe + Average	GloVe + PCA	GloVe + CA-SEM (our approach)
STS' 12	54.8(52.5)	72.3(56.2)	73.9
STS' 13	60.5(42.3)	66.1(56.6)	69.6
STS' 14	42.2(54.2)	68.4(68.5)	72.2
STS' 15	50.4(52.7)	67.6(71.7)	75.9
STS' 16	30.1	34.7	37.2

Table 3: Results on SemEval 2017 dataset. The numbers inside the parentheses are reported by Arora et al. (2017).

4.1.3 SICK SIMILARITY DATASET

Fig. 8 illustrates the loss and accuracy for each block coordinate iteration. In this case, the loss keeps decreasing after 20 iterations (Fig. 8 (a)) but the accuracy remains near 0.71 (Fig. 8 (b)). Nevertheless, this is a significant improvement over the original paper (Table 4).

	GloVe + Average	GloVe + PCA	GloVe + CA-SEM (our approach)
SICK' 14	62.4(65.9)	64.7(66.2)	71.1

Table 4: Results on SICK' 14 dataset. The numbers inside the parentheses are reported by Arora et al. (2017).

4.2 Results for CA-BSFE

While CA-SEM is useful for unsupervised learning, CA-BSFE is designed for supervised learning, which provides a solution for handling any input containing bag of sparse features that are not

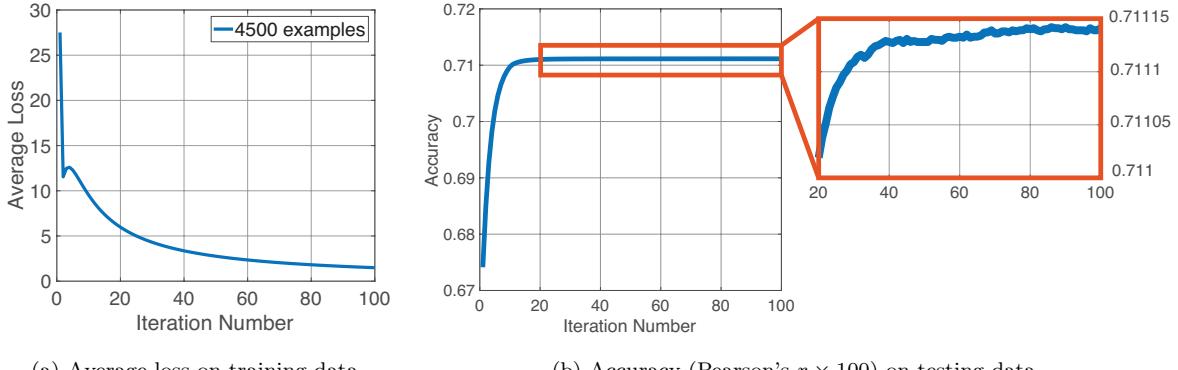


Figure 8: Training results on SICK dataset.

limited to text. In this section, we verify that its intrinsic support for computing the importance of each input (through χ -function) can help i) improve model training accuracy, ii) provide us with additional information from the input data and iii) provide us the flexibility in controlling generalization errors in training, especially when training data size is small.

4.2.1 IMPLEMENTATION DETAILS

The architecture of Fig. 2 can be implemented using Tensorflow APIs, where the embedding of each word w is sliced into two parts: word embedding \vec{w}' and χ -function input \vec{w}_σ (Eq. 20). A global vector \vec{v}_0 should also be defined as a trainable variable. In the following we use $\dim(\vec{w}')$ and $\dim(\vec{w}_\sigma)$ to denote the embedding dimensions of \vec{w}' and \vec{w}_σ , respectively.

We evaluate our CA-BSFE layer by plugging it into a sequence-to-binary, logistic regression problem (*e.g.*, positive/negative sentiment prediction for movie reviews). Here the input words $\{w|w \in s\}$ are mapped into an embedding \vec{s} via our CA-BSFE layer, then it is directly fed into a sigmoid output layer. A standard sigmoid cross entropy is employed as the loss function.

Block-coordinate gradient update scheme. However, in our first attempt to apply a naive implementation of CA-BSFE to the IMDB dataset, the gain in accuracy over average-based composition is quite marginal ($\sim 1\%$) as shown in Fig. 9 (a). One possible explanation is that the data size is too small that the model can quickly memorize the whole training dataset with zero error (see the training accuracies in Fig. 9 (a)), *i.e.*, the model can easily stuck at one of many possible minima that happens to not generalize well. Fortunately, our model construction allows additional freedom in minima search that prevents the energy from converging too quickly. The idea is similar to Alg. 1 for CA-SEM: only update a subset of all the variables in each training step. Specifically, we can alternate between fixing the values of $\{\chi(\vec{w}_\sigma)|w \in \mathcal{W}\}$ or the values of $\{\vec{w}'|w \in \mathcal{W}\}$ every em_steps (here we employ the name ‘‘em’’ as our algorithm resembles the expectation maximization, or EM algorithm) during the back-propagation update¹¹. In the following we will provide detailed analysis on how changing em_steps can help reduce generalization error.

For all our experiments in this section, we use a batch size of 16, a learning rate of 0.1, $\dim(\vec{w}') = 5$ and an optimizer of Adagrad Duchi et al. (2011) as the basic training config. For all our CA-BSFE models, we set $\dim(\vec{w}_\sigma) = 5$.

11. This can be achieved using Tensorflow 1.0 by two lines of code:

```
should_update = tf.equal(tf.mod(tf.train.get_global_step() / em_steps, 2), 0)
value = tf.cond(should_update, lambda : value, lambda : tf.stop_gradient(value)).
```

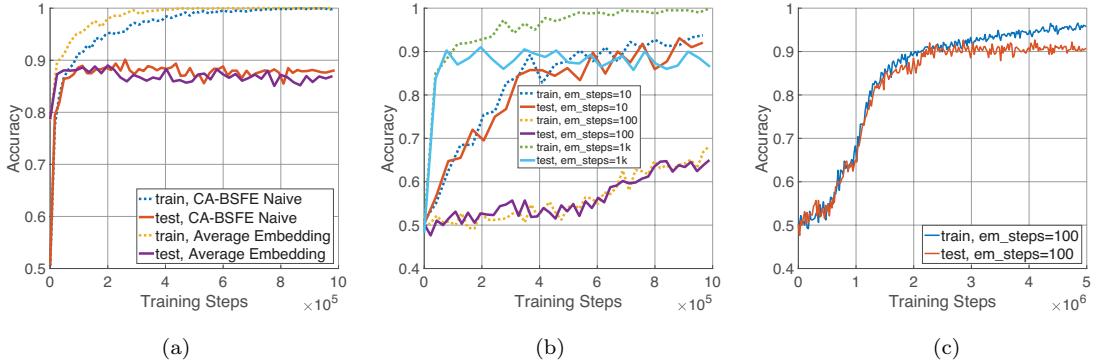


Figure 9: (Best viewed in color) Training results on IMDB dataset. Our block coordinate gradient update scheme can help improve the model’s ability in generalization by preventing the model from memorizing the whole training dataset too quickly.

4.2.2 IMDB DATASET

Fig. 9 shows the training and testing accuracy for different hyper-parameters. Directly updating the gradients of $\chi(\vec{w}_\sigma)$ and \vec{w}' simultaneously only gains a small margin over embedding composition schema based on averaging (Fig. 9(a)). One can see that both methods can memorize the whole training dataset after 600K steps. Applying our block-coordinate style gradient update makes obvious differences in narrowing the gap between training and test accuracy (Fig. 9(b)). However, too large em_steps (*e.g.*, 1000) still makes the model remember the training data quickly and leads to poor generalization. On the other hand, smaller em steps can lead to slower convergence but better accuracy (Fig. 9(c)). For em_steps = 100, the testing accuracy is stabilized above 0.91 after 6M steps, which is a significant improvement over the averaging scheme that achieves 0.88 at best during the first 200K training steps and degrades gradually afterwards.

4.2.3 SST DATASET

Results on the SST dataset further confirms our observation on the benefits of our block-coordinate gradient update scheme. In this case there are relatively a large number of training instances (67,349 *v.s.* 25,000 for IMDB) and a small number of testing cases (1,821 *v.s.* 25,000 for IMDB), and the same model cannot memorize all the training data. Nonetheless, using an em_steps > 0 still helps the generalization as shown in Fig. 10. It still hold true that CA-BSFE without em steps outperforms average-based scheme by a small margin (Fig. 10(a)), but with em steps the accuracy can easily stabilize above 0.82 (Fig. 10(b) and (c)). However, this time for em_steps = 1000 it performs poorly for the first 1M steps (Fig. 10(b)), but it starts to catch up after 2M training steps (Fig. 10(c)), which means the choice of em_steps strongly depends on the data.

4.3 Results for CA-ATT

Although our attention model share similar forms with existing ones (*e.g.*, Bahdanau et al. (2014); Luong et al. (2015)), the assumptions behind them are quite different. Recall that in Sec. 3.3, we have connected attention to the distribution $P(s|c_{t-1})$, where s is the input sequence (*a.k.a* memory in Fig. 3) and c_{t-1} is the cell state that summarizes the target output up to time $t-1$, or query. This is surprisingly consistent with existing implementation of NMT model Luong et al. (2017), which uses the cell state from previous time to query the attention vector (Fig. 11).

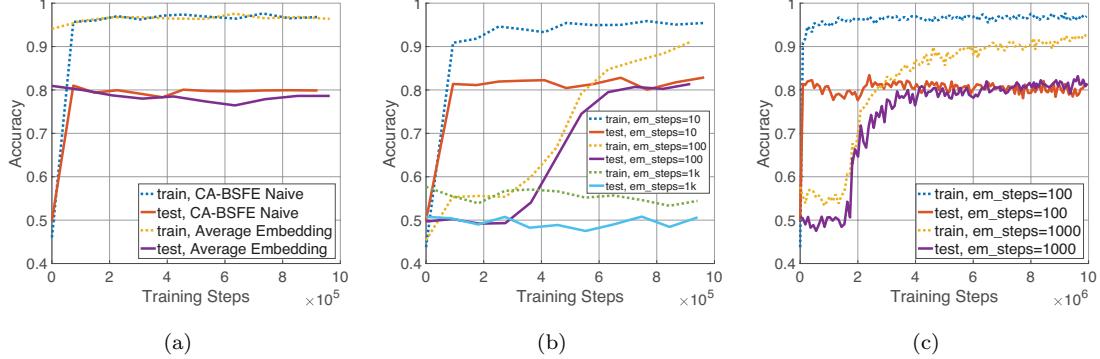


Figure 10: (Best viewed in color) Training results on SST dataset.

We compare our attention model with two popular ones: *Bahdanau attention* Bahdanau et al. (2014) and *Luong attention* Luong et al. (2015). The major difference between the above mentioned two attentions is how the scoring function $f(\cdot, \cdot)$ in Eq. 24 is defined:

$$f(\vec{s}_i, \vec{c}) = \begin{cases} \mathbf{v}^T \tanh(\mathbf{W}_1 \vec{c} + \mathbf{W}_2 \vec{s}_i) & \text{Bahdanau attention} \\ \vec{c}^T \mathbf{W} \vec{s}_i & \text{Luong attention} \end{cases} \quad (43)$$

In fact, our χ -function in Eq. 26 can also be defined using one of the above two forms. In this section, we show detailed analysis on the performance of these different implementations.

4.3.1 IMPLEMENTATION DETAILS

Our new attention model is implemented using Tensorflow by making some slight changes to existing `tf.contrib.seq2seq.AttentionWrapper` class¹² to allow for customized attention function and implementing a new subclass for `tf.contrib.seq2seq.AttentionMechanism`¹³. This way, we can easily compare different attention mechanisms side by side.

Our model is based on the publicly available implementation of the neural machine translation (NMT) architecture Bahdanau et al. (2014), which involves a sequence to sequence model with attention mechanism Luong et al. (2017). As mentioned above, unlike any existing implementations, we do not use any pre-defined vocabulary for token embedding. The hyper-parameters for our model are selected as follows: batch size is 16; learning rate is 0.01; embedding dimension for all tokens in source, target and output sequences is 100; optimizer for gradient descent is Adagrad. Our training is distributed into 100 worker machines without any hardware accelerations and it can achieve 20–30 global steps per second.

4.3.2 WMT’ 16 DATASET

When comparing our new attention model with existing ones, unfortunately, we find that there is no significant performance gain: all three attention mechanisms are quite comparable. Fig. 12 illustrates the training accuracy computed by looking at if the top k ($k = 5, 10, 100$) tokens returned by the TopK layer in Fig. 11 contain the target token. We have also tried to rewire the model in Fig. 11, e.g., by using the cell state from the top RNN cell as query and feed the returned attention vector to the output. Still none of these efforts result in any better performance gain; most of them perform significantly worse than the original design.

12. https://www.tensorflow.org/api_docs/python/tf/contrib/seq2seq/AttentionWrapper

13. https://www.tensorflow.org/api_docs/python/tf/contrib/seq2seq/AttentionMechanism

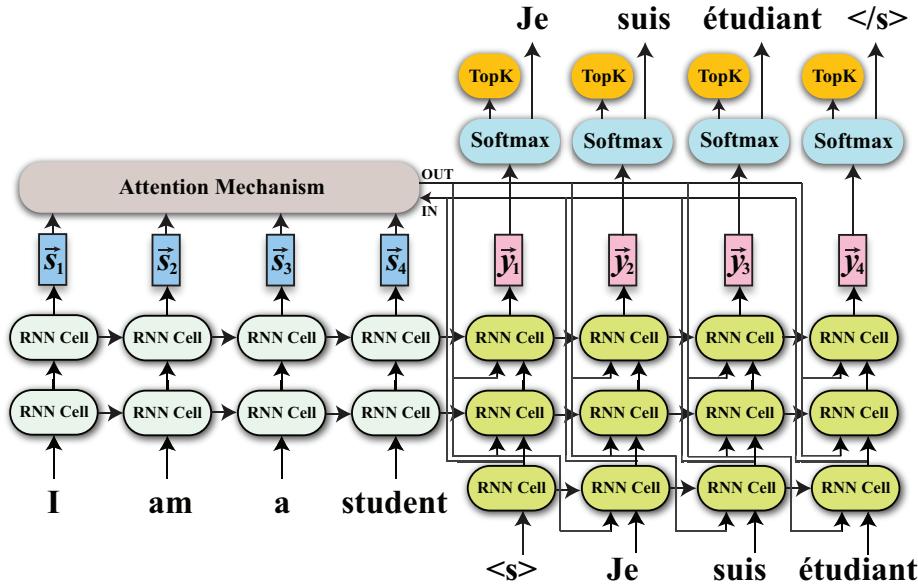


Figure 11: The architecture of NMT model Bahdanau et al. (2014).

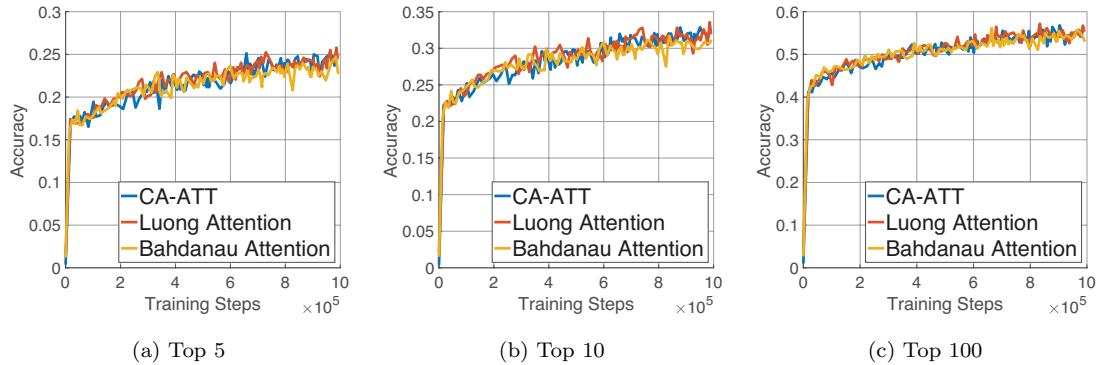


Figure 12: (Best viewed in color) Comparison among different attention mechanisms in the NMT architecture Fig. 11 (a) for WMT' 16 dataset.

Hence, the lesson learned from these experiments is that designing a model based on the factorization can lead to the ‘correct’ solution, as we will see in the next section.

4.4 Results for CA-RNN

4.4.1 IMPLEMENTATION DETAILS

In Tensorflow each sequence model (*e.g.*, RNN, GRU, LSTM) is implemented as a subclass of `tf.contrib.rnn.LayerRNNCell`¹⁴. Therefore we only need to implement our CA-RNN model as a new RNN cell under the same interface, whose output is a tuple of (output embedding, cell state embedding).

14. https://www.tensorflow.org/api_docs/python/tf/contrib/rnn/LayerRNNCell

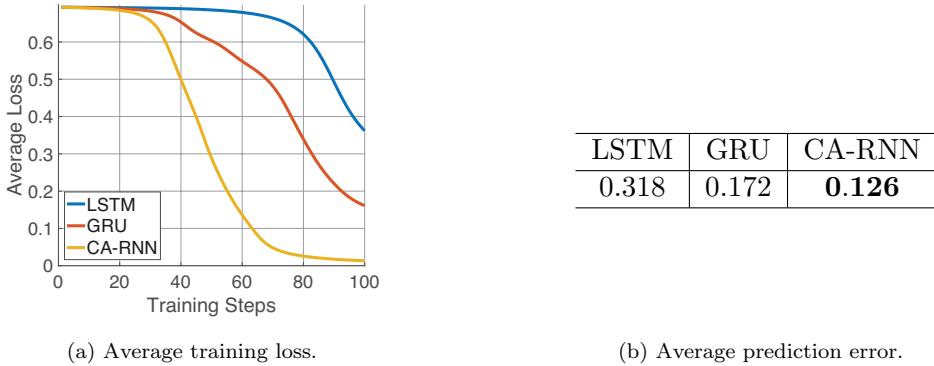


Figure 13: (Best viewed in color) Results on synthetic dataset for model validation described in Sec. 4.4.1.

Because our EDF only provides a form of the CA-RNN model (Eq. 31 and 35), choosing the right implementation is a trial-and-error process that often depends on the data of interest. Nevertheless, we require that our model should at least pass some very basic test to be considered valid. This said, we design a very simple model, which take a sequence as input and it is connect to any RNN cell (CA-RNN, GRU or LSTM), whose cell state embedding is further connect to a logistic regression output as a 0/1 classifier. Our synthetic training data is also simple. For training we let the model learn the mappings $\{\text{'I', 'am', 'happy'}\} \rightarrow 1$ and $\{\text{'You', 'are', 'very', 'angry'}\} \rightarrow 0$ for 100 iterations. For testing we expect the model to predict $\{\text{'I', 'am', 'very', 'happy'}\} \rightarrow 1$ and $\{\text{'You', 'look', 'angry'}\} \rightarrow 0$ (our model predicts 1 if its logistic output is greater than 0.5, otherwise it predicts 0). This model validation process can help us quickly rule out impossible implementations. For example, we found the peephole parameter \mathbf{p}_v in Eq. 34 cannot be replaced by matrix as linear mapping. Otherwise the model would not predict the correct result for the testing data.

Fig. 13 shows the average loss and prediction error from running our model 10 times on the above mentioned testing data. It can be seen that CA-RNN model is better at both memorizing training data and generalizing to novel input.

4.4.2 WMT' 16 DATASET

Here we employ the same model and hyper-parameters mentioned in Sec. 4.3.1 with a fixed attention mechanism (Luong et al. (2015)), only to change the RNN cell for both encoder and decoder for comparisons.

To evaluate the accuracy of the model on testing data, a BLEU score Papineni et al. (2002) metric is implemented for evaluation during training: for each target output token, simply feed it into the decoder cell and use the most likely output tokens (*i.e.*, the top 1 token) as the corresponding translation. We argue that this is a better metric to reflect the performance of a model than the BLEU score computed based on translating the whole sentence, since the later may vary with the choice of post-processing such as beam search.

From the results of first 1M training steps, one can see that CA-RNN consistently outperforms LSTM and GRU in top-k accuracy (Fig. 14(a), (b) and (c)) and BLEU score for prediction (Fig. 14(d)).

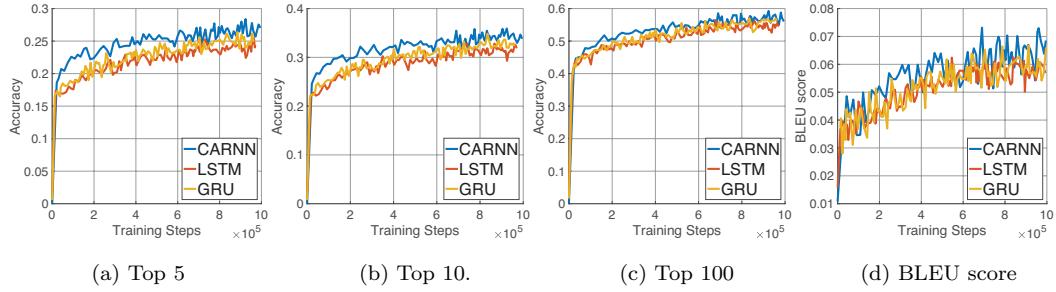


Figure 14: (Best viewed in color) Training accuracy on WMT’ 16 English to German dataset for first 1M training steps.

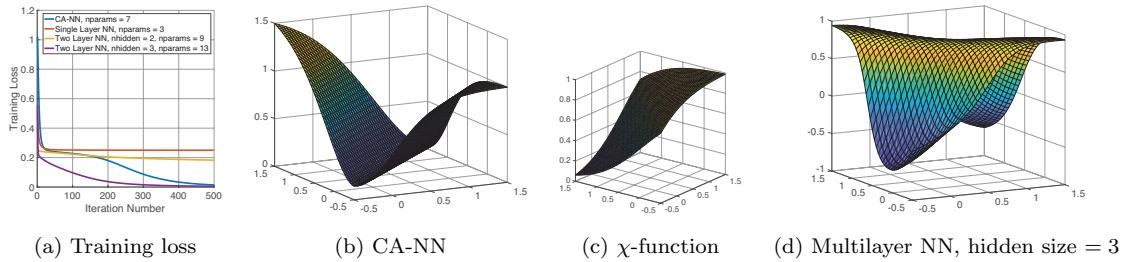


Figure 15: (Best viewed in color) Surfaces generated by using different models to fit four point mappings: $(1, 0) \mapsto 1$, $(0, 1) \mapsto 1$, $(0, 0) \mapsto 0$, $(1, 1) \mapsto 0$.

4.5 Results for CA-NN

A single layer CA-NN model can be considered as the context-aware counterpart of a NN hidden layer, whose applications is mainly focused on regression. It is already established that a NN with a single hidden layer can approximate arbitrary function Funahashi (1989) given sufficient amount of training data and training steps. However, a simple NN hidden layer’s ability to generalize may be poor given its simple structure. In contrast, our CA-NN model allows additional flexibility to improve the model’s generalization power as we will show in this section.

4.5.1 SURFACE FITTING

Synthetic data. In this case, we simply let the model learn a mapping from 2D to 1D height from four training data: $(1, 0) \mapsto 1$, $(0, 1) \mapsto 1$, $(0, 0) \mapsto 0$, $(1, 1) \mapsto 0$. The loss function is defined by the L2 distance. Note that our CA-NN model allows a subnetwork to define $v(\vec{c})$ (Fig. 5). Here we use a simple linear mapping with tanh activation. Fig. 15 shows the comparison between CA-NN and common single/multiple layer neural networks. CA-NN tends to fit the data with fewer parameters (Fig. 15 (a)). Also the fitted surfaces are very different from NN models when we plot the learned network in the extended domain $[-0.5, 1.5] \times [-0.5, 1.5]$ (Fig. 15 (b) and (d)). The surface learned by CA-NN tends to be less symmetric than that learned by NN-model (adding more hidden units for the NN model would result in very similar shapes as Fig. 15 (d)). The shape of the surface fitted by CA-NN model can be explained by looking at the χ -function (Fig. 15 (c)), which leads the shape into two different modes.

Smooth surface. In another example, we test CA-NN’s ability to fit a smooth surface with a small number of discrete samples. The surface we choose can be represented in a closed form as

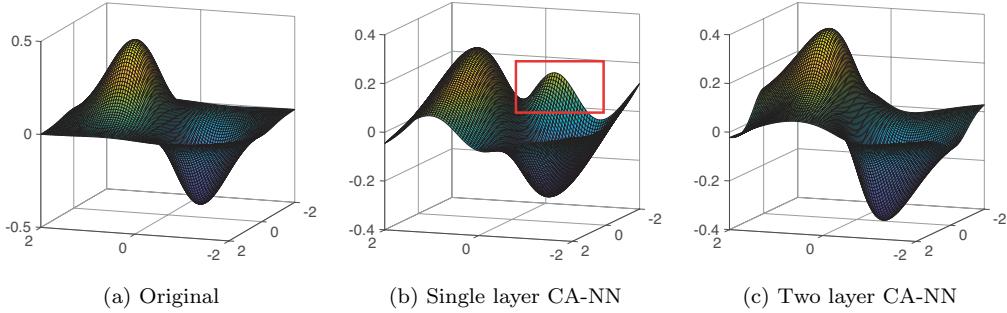


Figure 16: (Best viewed in color) Fitting a smooth surface (a) using a single layer CA-NN can introduce artifacts (b). Simply stacking multiple layers of CA-NN can sufficiently reduce them using the same training steps.

$(x, y, xe^{-x^2-y^2}), x, y \in [-2, 2]$ (Fig. 15 (a)). Here we discretize the surface on a 81×81 grid and use 1% of them for training and the rest for testing. For the CA-NN model, we choose the output/hidden layer size to be $\dim(\vec{w}_0) = 5$ in both single and multiple layer cases, and the model is trained 1000 steps with a learning rate of 0.1. Fig. 15 (b) illustrates one side effect of the χ -function, namely it may introduce undesired artifacts when the model is not fully learned. This can be remedied by stacking another layer of CA-NN to it, which is able to better approximate the original shape. The average predication error for single layer CA-NN is 0.0063, and for multi-layer case it is 0.0029.

4.6 Results for CA-CNN

Here we focus only on the application of convolutional layer in image related applications. We use the MNIST dataset to develop an intuition of the performance for a single CA-CNN layer, which can guide us build more complex model on more challenging dataset such as CIFAR Krizhevsky and Hinton (2009).

4.6.1 IMPLEMENTATION DETAILS

Our model is implemented by making some modifications to Tensorflow’s *conv2d* function¹⁵, which takes a high dimension tensor (*e.g.*, an image), an integer specifying output depth and a tuple representing kernel size, etc, as input and outputs a tensor representing the result of convolution. For example, in the 2D case, if the kernel size is $(3, 3)$ and the depth is 128, we can simply represent the CNN model as “ $(3, 3), 128$ ”. In our training, we select a batch size of 100, a stride size of 1, a learning rate of 0.01 and the same Adagrad optimizer for gradient descent update.

4.6.2 MNIST DATASET

For the relatively small MNIST dataset, we find a single layer CA-CNN model with one additional hidden layer (a fixed size of 128 is used in all our experiments) to be sufficient for achieving $> 99\%$ accuracy. Compared to its CNN counterpart, the same configuration (kernel size and depth) always achieves significantly lower accuracy as shown in Fig. 17. Fig. 17 shows the influence of different kernel sizes and output depths to the accuracy, and it turns out that for CA-CNN model, a larger kernel size usually leads to a better performance. Intuitively, a larger kernel size helps better

15. https://www.tensorflow.org/api_docs/python/tf/nn/conv2d

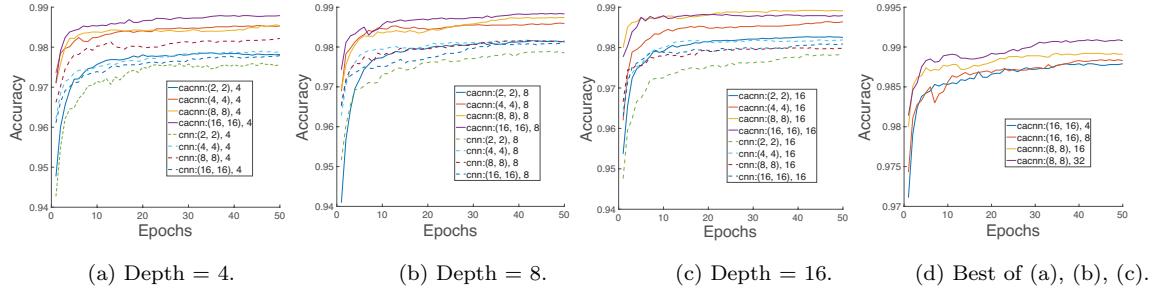


Figure 17: (Best viewed in color) Comparisons between CNN and CA-CNN on MNIST dataset with different hyper-parameters.

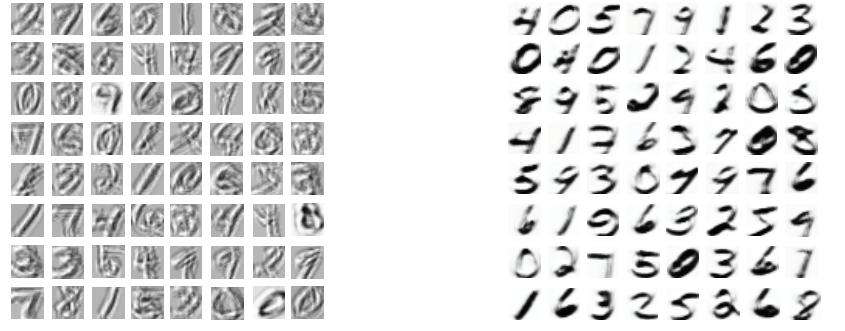


Figure 18: The σ -map (Fig. 6) of the χ -function in Eq. 39, where $\sigma \in [0, 1]$.

estimate the foreground/background segmentation represented by χ -function. In contrast, CNN model performs the best with a relatively small kernel size ((4, 4) or (8, 8)).

Fig. 18 further validates the role of χ -function in helping divide relevant/irrelevant information from the input. Here values of the χ -function for every pixel are visualized as grey-scale images, or also called σ -maps. Samples of these σ -maps from different training steps are shown in Fig. 18 (a) and (b). After 10,000 training steps, it can be seen that the foreground and background are very well separated (Fig. 18 (b)).

4.6.3 CIFAR-10 DATASET

For the CIFAR dataset, we employ a very simple model architecture to test the performance of different configurations: let the input image first go through multiple layers of convolutional networks, then feed the result to a hidden layer of size 512 before the final softmax output (Fig. 19(a)). Here the convolutional network layer can be a CNN, a CA-CNN, or a CA-RES (to be discussed in the next section).

Fig. 19 shows some initial results on running CA-CNN with different configurations. From Fig. 19(b) it can be seen that increasing CA-CNN layer sizes can help improve the accuracy. But when the layer size reaches 16, the model would not learn anything without other strategies like the block-coordinate gradient update scheme mentioned in Sec. 4.2.1. It is also consistent with the MNIST results that for CA-CNN, a moderate kernel size (*e.g.*, (8, 8)) works better than smaller or larger ones (Fig. 19(c)).

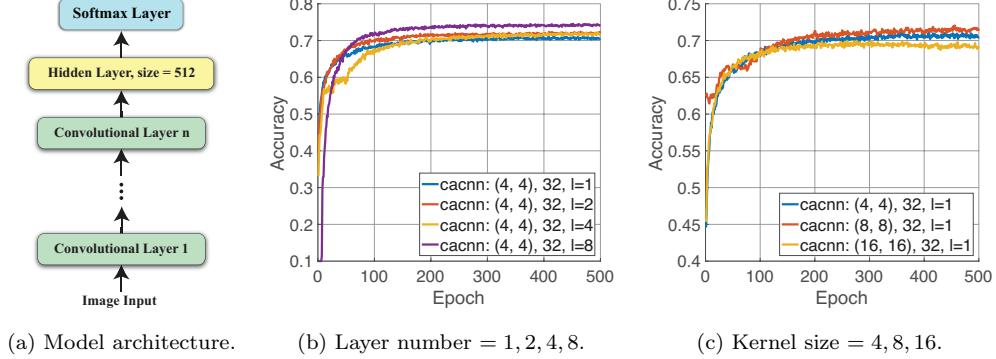


Figure 19: (Best viewed in color) Model architecture (a) and some initial results from various configurations ((b) and (c)).

4.7 Results for CA-RES

Our CA-RES model can be used as a replacement to a regular deep network with multiple hidden layers, or a deep residual network. Our design is flexible enough to subsume them as our special cases, while providing a lot more possible new architectures controlled by a few hyper-parameters. In this section, we only demonstrate a few aspects of our model and leave it to future work to explore its many more properties.

4.7.1 IMPLEMENTATION DETAILS

The CA-RES model is implemented in a way that its complexity can be controlled by a tuple (l, n_v, n_σ, b) where $l, n_v, n_\sigma \in \mathbb{N}^+, b \in \{0, 1\}$. Here l is the total number of layers; n_v specifies the maximal number of previous layers whose activation \vec{w}_i are used for computing $\vec{v}_l(\cdot)$ in Eq. 42, i.e., $\vec{v}_l(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_{l-n_v})$; similarly, n_σ is used for computing $\chi(\cdot)$ in Eq. 42, i.e., $\chi(\vec{c}, \vec{w}_{l-1}, \dots, \vec{w}_{l-n_\sigma})$; b indicates if $\chi(\cdot)$ should always be on ($b = 1$), therefore bypassing \vec{w}_{l0} . This way, it is easy to see that a configuration of $(2, 1, -, 1)^{16}$ with $\vec{v}_2(\vec{c}, \vec{w}_1) = \vec{c} + \vec{w}_1$ and $\vec{w}_1 = \vec{v}_1(\vec{c})$ where \vec{v}_1 is any activation function, would reduce our model to the residual network He et al. (2015). To use CA-RES in a deep neural network, one can either specify a large number of l , or stack multiple blocks of CA-RES layers. For gradient update, the block-coordinate gradient update scheme proposed in Sec. 4.2.1 can also be employed to improve the accuracy.

4.7.2 CIFAR-10 DATASET

Here we use the same model mentioned in Sec. 4.6.3, with the difference that each convolutional layer is a CA-RES network that may contain multiple sublayers (e.g., Fig. 5(b) and (c)). Therefore in addition to testing the impacts of layer number and kernel size to the final accuracy, we also should test the effect of sublayers in our CA-RES model.

Similar to our CA-CNN model, more layers always leads to higher accuracy as shown in Fig. 20(a). However, unlike CA-CNN, a bigger kernel size does not help in improving the accuracy (Fig. 20(b)). Unsurprisingly, more number of sublayers in each CA-RES layer also drives better results (Fig. 20(c)).

16. When $b = 1$, the value of n_σ can be ignored as $\chi(\cdot) \equiv 1$.

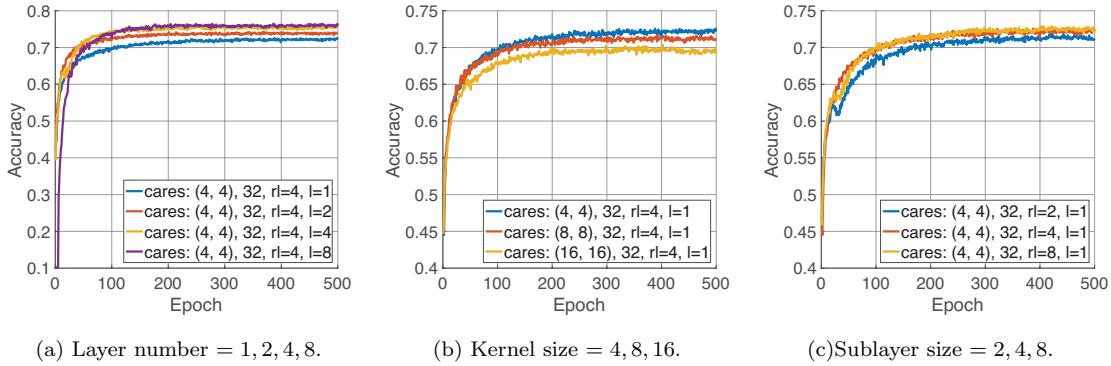


Figure 20: (Best viewed in color) Results for various configurations of using CA-RES in a multilayer network for CIFAR-10 dataset.

5. Related work

Loosely speaking, existing supervised machine learning models can be categorized based on the type of input they take. Sparse models (*e.g.*, CA-SEM and CA-BSFE) handle input with discrete values such as text or categorical features, etc. Sequence models (*e.g.*, CA-ATT and CA-RNN) handle input with an order and also varying in length. Image models (*e.g.*, CA-RES and CA-RNN) take images as input that are usually represented by $n \times m$ array. Such a taxonomy is only helpful for us to review related works; they are not necessarily exclusive to each other. Because our new framework provides a unified view of many of these models, at the end of this section, we also review related efforts in building theoretical foundations for deep learning.

5.1 Sparse Models

While a single sparse feature can be represented as a vector, there is no consensus on how to represent a bag of features in the embedding space. Although not many work has been done on the general problem of representing a bag of sparse features using a single vector (Mitchell and Lapata, 2010; Hermann and Blunsom, 2014), its special case, phrase/sentence embedding in language modeling, has already drawn increasing attention due to the fast development of distributed representation using neural networks (Bengio et al., 2003; Mikolov et al., 2013; Pennington et al., 2014; Arora et al., 2016). Existing approaches either fall into bag of words (BOW) based, or sequence based. Extensive comparative studies have shown that neither approach has obvious advantage over the other and their performance highly depends on the specific metric (Wieting et al., 2016; Hill et al., 2016).

Among bag of words based models, early attempts including Le and Mikolov (2014) directly extended the idea of Word2Vec (Mikolov et al., 2013) by treating each sentence as a whole with a unique feature id. Later, Hill et al. (2016) showed that a more flexible approach, which simply computes the sentence embedding by summing up the embeddings of individual words, would achieve significant improvement. Wieting et al. (2016) also showed that starting from an existing word embedding model, training them towards a specific task, then simply using the average of new word embeddings in a sentence would achieve satisfactory results. Recently, Arora et al. (2017) proposed a TF-IDF like weighting scheme to composite the embedding of words in a sentence and showed that the performance can be improved simply by removing the first principle component of all the sentence vectors. More recently, the transformer architecture using self-attention proposed

by Vaswani et al. (2017) provides further evidences that bag of words model can be more preferable than sequence models, at least in certain tasks.

As for sequence-based approaches, Kiros et al. (2015) proposed SkipThought algorithm that employs a sequence based model to encode a sentence and let it decode its pre and post sentences during training. By including attention mechanism in a sequence-to-sequence model (Lin et al., 2017), a better weighting scheme is achieved in compositing the final embedding of a sentence. Besides attention, Wang et al. (2016) also considered concept (context) as input, and Conneau et al. (2017) included convolutional and max-pool layer in sentence embedding. Our model includes both attention and context as its ingredients.

In parallel to neural network based models, another popular approach to feature embedding is based on matrix factorization (Koren and Bell, 2015), where a column represents a feature (*e.g.*, a movie) and each row represents a cooccurrence of features (*e.g.*, list of movies watched by a user). In this setting, one can compute the embeddings for both the features and cooccurrences, and the embedding for a new cooccurring case (bag of features) can be computed in a closed form (Hu et al., 2008). Recently, the concept of cooccurrence is generalized to n-grams by reformulating the embeddings of bag-of-features as a compressed sensing problem Arora et al. (2018). Nevertheless, these methods are mostly limited to linear systems with closed-form solutions.

5.2 Sequence Models

Modeling sequential data (*e.g.*, sentences or sound) requires inferring current output from both past observation and current input. Recurrent neural network (RNN) assumes past observations can be summarized in its hidden state, which can then be learned recurrently (recursive in time). Despite its promising modeling power, its basic form is proven to be difficult to train Bengio et al. (1994). Hence various improvements are proposed based on different assumptions (*e.g.*, Pascanu et al. (2013)), among which the LSTM model Sepp Hochreiter (1997) might be among the most successful one that are widely used. Encouraged by its success, various models are proposed to modify its design but very few are game-changing Greff et al. (2016). One of the very few successful challengers of LSTM is the GRU model Cho et al. (2014), which is proven to perform better than LSTM in many cases.

More recently, rather than improving the basic RNN-like structures, it was found that changing how they are stacked and connected can help solve very challenging problems like language translation. The sequence-to-sequence model Sutskever et al. (2014) is one of such successful stories. Then the attention mechanism was proposed able to significantly improve the performance Bahdanau et al. (2014). Recently, it was practically validated that even the basic RNN structure can be rid of by using attention mechanism alone Vaswani et al. (2017).

5.3 Image Models

The visual pathway is one of the most well understood components in human brain, and the convolutional neural network model LeCun et al. (1990) has been proven to model its hierarchical, multi-layer processing functionality very closely. A series of upgrades have been added to the basic CNN structure to improve its performance, such as pooling Lecun et al. (1998), dropout Hinton et al. (2012), or normalization Ioffe and Szegedy (2015); Ba et al. (2016). Recently, the residual network model He et al. (2015) and its variants (*e.g.*, Zagoruyko and Komodakis (2016); Srivastava et al. (2015)) becomes very popular due to their superior performance. Among them, the highway networks Srivastava et al. (2015) shares certain similarity with our CA-NN model in its use of a gate (maps to our χ -function) to switch between an identity activation (we use a default value \vec{w}_0 instead) and a nonlinear activation (maps to $v(\vec{c})$). Each year, combinations of these building blocks are proposed to reach higher accuracies on standard benchmarks for image classification (*e.g.*, Krizhevsky et al. (2012); Szegedy et al. (2015)) or other tasks.

However, human vision is also known to be sensitive to changes in time and attend only to contents that have evolutionary value. Mnih et al. (2014) proposed a reinforcement learning model to explain the dynamic process of fixation in human vision. Xu et al. (2015) directly generalizes the text attention model in the encoder-decoder structure Bahdanau et al. (2014) to image caption generation. Wang and Shen (2018) proposed a multilayer deep architecture to predict the saliency of each pixel. Note that our CA-CNN architecture also allows explicit computation of saliency maps for input images.

5.4 Theoretical Foundations for Deep Learning

Initially, a biological neuron is characterized by its binary active/inactive states McCulloch and Pitts (1943) and hence inspired the perceptron model Rosenblatt (1958). This was then extended to multiple neurons that are divided into visible or hidden, whose states can be modeled by a Boltzmann machine Hinton et al. (1985), or its simplified version, *i.e.*, restricted Boltzmann machine (RBM) Smolensky (1986). Later on, this framework was further extended to the exponential family Welling et al. (2005)¹⁷. Despite their plausibility in explaining certain brain functions (*e.g.*, Hinton et al. (1995)), and deep connections to statistical physics and graphical models, these models are often inefficient to learn and limited in solving large scale problems compared to those heuristically defined, back propagation based neural network models discussed in this paper.

Connecting data-driven machine learning problems to physics-driven geometric approaches can help us understand the problem from a different angle. Intuitively, geometry concerns the shape of the data (topology + metrics), which translates to the embeddings of the data and their distance measures Wainwright (2002); Lebanon (2005). Another direction is to study the process of optimization (gradient descent) as a continuous, dynamical system such that one can easily analyze its optimality properties (*e.g.*, Mei et al. (2018); Chizat and Bach (2018)). These analyses are often based on simplified assumptions of the energy or network structure and therefore may not scale well if the complexity of the system increases.

6. Conclusion

We demonstrated that it is beneficial both in theory and in practice to study neural network models using probability theory with exponential family assumption – it helps us better understand existing models and provides us with the principle to design new architectures. The simple yet fundamental concept of context-awareness enables us to find a mathematical ground for (re)designing various neural network models, based on an equally simple EDF. Its intrinsic connections to many existing models indicates that EDF reveals some *natural* properties of neural networks that is universal. It is foreseeable that many new models would be proposed to challenge the status-quo, based on principles developed in this paper, as we plan to do in the near future.

17. Again, its major difference with our use of exponential family is that we do not deal with the number of visible/hidden units and instead consider a single layer as a variable, whereas the work of Welling et al. (2005) assumes the sufficient statistics (embedding) is applied to each visible/hidden unit.

References

- Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI)*, pages 265–283, 2016.
- Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.
- Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations (ICLR)*, 2017.
- Sanjeev Arora, Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A compressed sensing view of unsupervised text embeddings, bag-of-n-grams, and LSTMs. In *International Conference on Learning Representations (ICLR)*, 2018.
- Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. URL <https://arxiv.org/abs/1607.06450>.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL <http://arxiv.org/abs/1409.0473>.
- Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.
- Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055, 2017. URL <http://arxiv.org/abs/1708.00055>.
- Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *CoRR*, abs/1805.09545, 2018. URL <https://arxiv.org/abs/1805.09545>.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.
- Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 160–167, 2008.
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680, 2017.

- Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150, 2004.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.
- K. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, May 1989.
- Klaus Greff, Rupesh K. Srivastava, Jan Koutn \acute{k} , and J $\ddot{u}rgen Schmidhuber Bas R. Steunebrink. LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 2016.$
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.
- Karl Moritz Hermann and Phil Blunsom. Multilingual models for compositional distributed semantics. *CoRR*, abs/1404.4641, 2014. URL <http://arxiv.org/abs/1404.4641>.
- Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. *CoRR*, abs/1602.03483, 2016. URL <http://arxiv.org/abs/1602.03483>.
- D. Ackle Geoffrey E. Hinton, , and T. Sejnowski. A learning algorithm for boltzmann machines. *Cognition Science*, 1985.
- Geoffrey E. Hinton, Peter Dayan, Brendan J. Frey, and Radford M. Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, (26):1158–1161, 1995.
- Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006.
- Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012. URL <http://arxiv.org/abs/1207.0580>.
- Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Eighth IEEE International Conference on Data Mining (ICDM)*, pages 263–272, 2008.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015.
- Eugene M. Izhikevich. *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. MIT press, 2007.
- Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(9):1704–1716, September 2012.
- Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3294–3302, 2015.
- Yehuda Koren and Robert Bell. Advances in collaborative filtering. In *Recommender systems handbook*, pages 77–118. Springer, 2015.

- A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.
- Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1188–1196, 2014.
- Guy Lebanon. *Riemannian Geometry and Statistical Machine Learning*. PhD thesis, 2005.
- Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, R. E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems (NIPS)*, pages 396–404. 1990.
- Yann Lecun, Lon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.
- Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *International Conference on Learning Representations (ICLR)*, 2017.
- Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. Neural machine translation (seq2seq) tutorial. <https://github.com/tensorflow/nmt>, 2017.
- Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, 2015.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, 2011. Association for Computational Linguistics.
- C.D. Manning, P Raghavan, and H Schutze. *An Introduction to Information Retrieval*. 2008.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA), 2014.
- Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, (5):115–133, 1943.
- Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, (33), 2018.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. URL <http://arxiv.org/abs/1301.3781>.
- Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, 2010.

- Andriy Mnih and Geoffrey Hinton. Three new graphical models for statistical language modelling. In *Proceedings of the 24th International Conference on Machine learning (ICML)*, pages 641–648. ACM, 2007.
- Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2204–2212, 2014.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 311–318, 2002.
- Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, pages 386–408, 1958.
- Sebastian Ruder. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098, 2017. URL <http://arxiv.org/abs/1706.05098>.
- Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. Exponential family embeddings. In *Advances in Neural Information Processing Systems (NIPS)*, pages 478–486, 2016.
- Jürgen Schmidhuber Sepp Hochreiter. Long short-term memory. *Neural Computation*, pages 1735–1780, 1997.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.
- P. Smolensky. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Information Processing in Dynamical Systems: Foundations of Harmony Theory, pages 194–281. 1986.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.
- Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2377–2385. 2015.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3104–3112, 2014.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008. 2017.

Martin J. Wainwright. *Stochastic Processes on Graphs with Cycles: Geometric and Variational Approaches*. PhD thesis, 2002.

Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2018.

Yashen Wang, Heyan Huang, Chong Feng, Qiang Zhou, Jiahui Gu, and Xiong Gao. CSE: Conceptual sentence embeddings based on attention model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

Max Welling, Michal Rosen-zvi, and Geoffrey E Hinton. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1481–1488. 2005.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. In *International Conference on Learning Representations (ICLR)*, 2016.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of British Machine Vision Conference (BMVC)*, pages 87.1–87.12, 2016.

Xiao Zhang, Zhiwei Li, Lei Zhang, Wei-Ying Ma, and Heung-Yeung Shum. Efficient indexing for large scale visual search. In *International Conference on Computer Vision (ICCV)*, 2009.