Deep-Deep Neural Network Language Models for Predicting Mild Cognitive Impairment

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong and Judyanne Sharmini Gilbert Fernandez
Intelligent Health Research Group
School of Information Technology, Monash University Malaysia
{sylvester.orimaye,jojo.wong,judyanne.gilbert}@monash.edu

Abstract

Early diagnosis of Mild Cognitive Impairment (MCI) is currently a challenge. Currently, MCI is diagnosed using specific clinical diagnostic criteria and neuropsychological examinations. As such we propose an automated diagnostic technique using a variant of deep neural networks language models (DNNLM) on the verbal utterances of MCI patients. Motivated by the success of DNNLM on natural language tasks, we propose a combination of deep neural network and deep language models (D2NNLM) to predict MCI. Results on the DementiaBank language transcript clinical dataset show that D2NNLM sufficiently learned several linguistic biomarkers in the form of higher order n-grams and skip-grams to distinguish the MCI group from the healthy group with reasonable accuracy, which could help clinical diagnosis even in the absence of sufficient training data.

1 Introduction

Early diagnosis of Mild Cognitive Impairment (MCI) is currently a challenge [Abbott, 2011]. More importantly, MCI has been typically diagnosed through extensive neuropsychological examinations using a series of cognitive tests containing a set of questions and images [Mitolo *et al.*, 2015]. For example, the Mini-Mental State Examination (MMSE) and the Montreal Cognitive Assessment (MoCA) screening tools are composed of a series of questions and cognitive tests that assess different cognitive abilities. The challenge with these cognitive tests is that the accuracy depends on the clinician's level of experience and their ability to diagnose different subtypes of the disease [Damian *et al.*, 2011]. Often, researchers and clinicians need to combine other cognitive tests with the MMSE [Mitchell, 2009], and in most cases wait for a reasonably long interval to ascertain diagnosis [Pozueta *et al.*, 2011]. More recently, research has also shown that the reliability of the MMSE as a tool for diagnosing MCI could be limited [Kim and Caine, 2014]. The National Institute on Aging and the Alzheimer's Association has also called for several other clinical criteria that could be used to effectively diagnose MCI and other similar disease in a non-invasive way [Albert *et al.*, 2011].

As opposed to the ad hoc use of neuropsychological examinations, linguistic ability captured from verbal utterances could be a good indication of MCI and other related diseases [Tillas, 2015]. The premise is that, MCI is characterized by the deterioration of nerve cells that control cognitive, speech and language processes, which consequentially translates to how patients compose verbal utterances. According to [Ball *et al.*, 2009], syntactic processing in acquired language disorders such as Aphasia in adults, has shown promising findings, encouraging further study on identifying effective syntactic techniques. Similarly, [Locke, 1997] emphasized the significance of lexical-semantic components of a language, part of which is observable during utterance acquisition at a younger age. That work further highlighted that as the lexical capacity increases, syntactic processing becomes automated, hence leading to lexical and syntactic changes in language.

As such, we are motivated by the effectiveness of deep neural networks language models (DNNLM) in modeling acoustic signals for natural language tasks. In particular, we are inspired by [Schwenk, 2007], which shows that a feed-forward neural network language model can be trained with low error rate and perplexity. Even with just 1 hidden layer, the performance of the NNLM was better than the conventional 4-gram language model. The DNNLM improved on the NNLM with lower error rate and perplexity.

Thus, we explore deep-deep neural networks language models (D2NNLM) to learn the linguistic changes that distinguish the language of patients with MCI from the healthy controls. The ordinary DNNLM uses lower order n-gram N dimensional sparse vectors as discrete feature representations to learn the neural network with multiple hidden layers (DNN). In this paper, we maintain the same DNN architecture and increase the depth of the language models by introducing higher order n-gram and $skip$-gram N dimensional sparse vectors as discrete inputs to the DNN rather than single word N dimensional sparse vectors. In other words, we create n-gram and $skip$-gram vocabulary spaces from which we formed the N dimensional sparse vectors. The premise is that clinical datasets are usually sparse and it is the same for the DementiaBank[1] dataset used in this paper. Thus, using lower order n-gram dimensional sparse vectors alone could limit the vocabulary space and subsume the essential linguis-

[1] http://talkbank.org/DementiaBank/

tic changes and biomarkers, which could potentially distinguish patients with MCI from the healthy controls. On the other hand, n-grams and $skip$-grams have been shown to be good class predictors in several language modeling tasks on sparse data [Sidorov *et al.*, 2014]. In addition, the DNN has been effective in learning discriminating features from sparse feature representations [Li *et al.*, 2015]. To the best of our knowledge, little work has considered deep neural network and deep language models for predicting MCI on sparse clinical language datasets.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 describes the formal architecture of the DNNLM. We explain our D2NNLM in Section 4. Section 5 presents our experimental results. Finally, Section 7 concludes the paper.

2 Related Work

In [Roark *et al.*, 2011], the efficacy of using complex syntactic features to classify MCI was demonstrated. In that work, spoken language characteristics were used to discriminate between 37 patients with MCI and 37 in the healthy elderly group using 7 significant pause and syntactic linguistic annotations as features to train Support Vector Machines (SVM). That technique achieved 86.1% Area Under the ROC Curve (AUC). On the contrary, we use language models, which are more representative of the language space of both the disease and healthy groups without using any handcrafted features.

More recently, [Prud'hommeaux and Roark, 2015] proposed a 'graph-based content word summary score' and 'graph-based content word word-level score' to predict Alzheimer's disease (AD), which is often preceded by MCI. Using SVM on the same DementiaBank, that work achieved 82.3% AUC. However, the graph based techniques require separately built alignment models with sufficiently large datasets.

This paper has two main contributions. (1) We introduce deep language models in the form of higher order n-grams and $skip$-grams N dimensional sparse vectors as discrete inputs to the DNN, hence we derived D2NNLM. (2) We show that D2NNLM predicts MCI with less percentage error, perplexity, and predictive accuracy compared to other baselines especially on sparse clinical language datasets.

3 Deep Neural Network Language Models

The DNNLM architecture has more than one hidden layer with nonlinear activations [Arisoy *et al.*, 2012], and it is built on top of the original feed-forward NNLM architecture [Bengio *et al.*, 2003]. Unlike the DNNLM, NNLM has only two hidden layers. The first hidden layer has a linear activation and often referred to as the projection layer. The second hidden layer uses a nonlinear activation, hence making the NNLM a single hidden layer neural network [Bengio *et al.*, 2003].

In this paper, we follow the notations used in [Schwenk, 2007] and [Arisoy *et al.*, 2012] to describe the components of the DNNLM architecture. Given a vocabulary space, each word in the vocabulary is denoted by a N dimensional sparse vector. In each vector, the index of that particular word is stored with 1 while other indices in the vector are stored with 0s. As inputs to the neural network, the discrete feature representations are concatenated to contain the n-1 previous words in the vocabulary space, which serves as the memory to the previous words history. Given that N is the vocabulary size and P is the size of the projection layer, linear projections of all the concatenated words are used to create the first hidden layer of the network from every ith row of the N x P dimensional projection matrix. This is followed by the hidden layer H with hyperbolic tangent non-linear activation functions as follows:

$$d_j = \tanh\left(\sum_{l=1}^{(n-1)\times P} M_{jl}c_l + b_j\right) \forall j = 1, ..., H \qquad (1)$$

where H is the number of hidden layers, the weights between the projection layer and the subsequent hidden layers are denoted with M_{jl}, and the biases of the hidden layers are represented with b_j.

Note that since the DNNLM follows the NNLM architecture, other hidden layers with the same hyperbolic tangent non-linear activation functions are added to make the network deeper. The output layer uses a softmax function to simultaneously compute the language model probability of each word i giving its history, h_j, thus $P(w_j = i|h_j)$. We present the details of the output layer and the language model probability as follows:

$$o_i = \sum_{j=1}^{H} V_{ij}d_j + k_i \forall i = 1, ..., N \qquad (2)$$

$$p_i = P(w_j = i|h_j) = \frac{\exp(o_i)}{\sum_{l=1}^{N}\exp(o_l)} \forall i = 1, ..., N \qquad (3)$$

where V_{ij} denotes the weights between the hidden layers and the output layer, k_i represents the biases of the output layer, and the p_i computes the language model probability for every ith output neuron.

4 Deep-Deep Neural Network Language Models

Our D2NNLM uses the same architecture with DNNLM comprising of multiple hyperbolic tangent non-linear activation functions. On top of that, we make the vocabulary space deeper by including additional n-gram and $skip$-gram vocabulary spaces to the ordinary n-gram vocabulary space used in the original DNNLM. Figure 1 shows the architecture of the the D2NNLM.

With regard to predicting language utterances with MCI, it is of paramount importance to our D2NNLM that the language is modeled with a vocabulary space of substantial depth due to the non-trivial nature of the problem [Roark *et al.*, 2011; Fraser *et al.*, 2014]. According to the study conducted by [Roark *et al.*, 2011], many of the handcrafted language and speech measures that have been used in distinguishing patients with MCI from their respective healthy controls – including some statistically significant measures – have shown

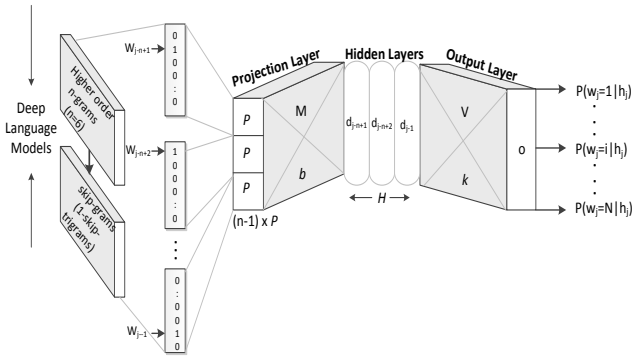

Figure 1: Deep-Deep Neural Network Language Models

the means and the standard deviations to be very close between the MCI and healthy control groups. Thus, it is probable that very little linguistic deficits will characterize either group. Even with DNNLM, which is based on simple n-gram language models with embedded words as a continuous feature space, it is still challenging to generalize over unseen data due to data sparseness problem [Arisoy *et al.*, 2012]. As such, an alternate technique could be found in using a stacked mixture of language models for embedding the vocabulary in a much deeper continuous space [Sarikaya *et al.*, 2009]. As shown in Figure 1, we stacked higher order n-gram and $skip$-gram language models to create deep language models for deep neural network. We refer to such models as Deep-Deep Neural Network Language Models and our preliminary experiments show that deeper language models potentially improve the performance of deep neural network for predicting MCI. We will describe the generation of the n-gram and $skip$-gram vocabulary spaces in the following sections.

Similar to DNNLM, the computational complexity of D2NNLM is characterized by the output layer's $H \times N$ matrix multiplications. As such, we follow [Sarikaya *et al.*, 2009] and performed Singular Value Decomposition (SVD) to produce a reduced-rank approximation of the n-gram and $skip$-gram vocabulary spaces before mapping the vocabularies into the continuous feature space. This is then followed by the stacking together (or concatenation) of the projected vocabulary history vectors. The undecomposed feature space typically has a large but sparse matrix containing few 1s and a lot of 0s. As such, SVD becomes a straightforward option to produce a compact approximation of the original feature space with optimal least-square as it sufficiently models the frequently occurring 0s, which are often not informative [Sarikaya *et al.*, 2009]. Thus, only the low dimensional vocabulary is used to learn the output targets. Note that the DNNLM assigns the probability mass to the output targets. A background language model was used to perform smoothing as done in [Schwenk, 2007]. We trained the neural network using the standard back-propagation algorithm to minimize the error function E_r as follows:

$$E_r = \sum_{i=1}^{N} t_i \log p_i + \epsilon \left(\sum_{jl} M_{jl}^2 + \sum_{ij} V_{ij}^2 \right) \forall j = 1, ..., H \tag{4}$$

where t_i is the target vector, parameter ϵ is determined empirically using the validation set. Note that the first half of the equation computes the cross entropy between the output and the target probability masses, and the second half computes the regularization term, which avoids overfitting the training data.

4.1 n-gram vocabulary space

The use of word n-gram is popular in NLP especially for developing language models that are able to characterize the lexical usage of grammar in a dataset. A word n-gram is the sequence of words identified as an independent representation of a part of the grammar in an utterance or a sentence. *'n'* in this case, represents the number of words in the sequence. For instance, when n is 1, it is called a 'unigram', which has only one word. Similarly, a 'bigram' and a 'trigram' have n equal to 2 and 3 respectively, and it is not uncommon to use higher order n-grams (i.e. $n \geq 3$) in learning tasks [Le *et al.*, 2011]. In this paper, our n-gram vocabulary space consist of 6-grams (n-gram=6) features only, which are generated from the transcripts of both the MCI and the healthy control groups. We believe that the 6-grams features could subsume other lower order n-grams such as unigrams, bigrams, and trigrams [Sarikaya *et al.*, 2009]. We put emphasis on higher order n-grams because they are known to have performed with reasonable accuracy in other NLP and ML tasks [Chen and Chu, 2010].

4.2 $skip$-gram vocabulary space

Skip-grams are commonly used in statistical language modelling problems such as speech processing. Unlike the ordinary n-grams, word tokens are skipped intermittently while creating the n-grams. For instance, in the sentence *"take the Cookie Jar"*, there are three conventional *bigrams*: "take the", "the Cookie", and "Cookie Jar". With skip-gram, one might skip one word intermittently for creating additional bigrams, which include "take Cookie", and "the jar". We believe such skip-grams could capture unique linguistic biomarkers in verbal utterances of patients with MCI. Thus, as described [Orimaye *et al.*, 2015], we used a compound of skip-grams to create our $skip$-gram vocabulary space. For each sentence $S = \{w_1 ... w_m\}$ in a verbal dialogue, we define k-skip-n-grams as a set of n-gram tokens $T_{ngram} = \{w_a, ..., w_{a+n-k}, ..., w_{a+n}, ..., w_{m-n}, ..., w_{(m-n)+n-k}, ..., w_m\}$, where n is the specified n-gram (e.g. 2 for $bigram$ and 3 for $trigram$), m is the number of word tokens in S, k is the number of word skip between n-grams given that $k < m$, and $a = \{1, ..., m - n\}$. Thus for the sentence *"take the Cookie Jar from the cabinet"*, 1-skip-2-grams will give {'take Cookie', 'the Jar', 'Cookie from', 'Jar the', 'from cabinet'} and 1-skip-3-grams will produce {'take Cookie Jar', 'take the Jar', 'the Jar from', 'the Cookie from', 'Cookie Jar the', 'Cookie from the', 'Jar the cabinet', 'Jar from cabinet'}. In

Vocabulary	MCI	Control
6-gram	1100	1163
1-skip-trigram	2877	3011
total	3977	4174

Table 1: n-gram and $skip$-gram vocabularies from the MCI and Control groups.

our experiments, we used only 1-skip-3-grams as some of its $skip$-grams often subsume 1-skip-2-grams.

5 Experiment and Results

5.1 Dataset and Baselines

We performed experiments on existing DementiaBank clinical dataset. The dataset was created during a longitudinal study conducted by the University of Pittsburgh School of Medicine on Alzheimer's disease (AD) and related Dementia, which was funded by the National Institute of Aging[2]. The dataset contains transcripts of verbal interviews with healthy controls and patients that were diagnosed with AD and MCI, including those with related dementia. Interviews were conducted in the English language and were based on the description of the Cookie-Theft picture component, which is part of the Boston Diagnostic Aphasia Examination. During the interview, patients were given the picture and were told to discuss everything they could see happening in the picture. The patients' verbal utterances were recorded and then transcribed into a transcription format with the equivalent text. For our experiments, we selected all the available 19 transcripts of MCI and an equivalent 19 transcripts of healthy controls.

Because many existing research works on MCI have mostly used different handcrafted features from self-collected datasets [Roark *et al.*, 2011], it is challenging to compare the D2NNLM with those works because we could not test D2NNLM on their datasets due to ethics constraints. Nevertheless, we compared our model to the DNNLM and NNLM as baselines on the same DementiaBank dataset. Our goal is to show the efficacy of the deep language model and deep neural network technique to the MCI prediction problem.

5.2 D2NN Language Models Settings

We generated the vocabulary for the D2NNLM from all the 38 transcript files containing 210 sentences for MCI and 236 sentences for healthy controls. Table 1 shows the details of the 6-gram and 1-skip-3-grams vocabularies, which were generated from the dataset. The D2NNLM training data consist of 50% of the MCI and the control groups' transcript files, while the test and validation sets consist of 25% of the transcript files, respectively.

The 50% training data for the MCI group has 104 sentences, 663 6-grams, and 1659 1-skip-trigrams. On the other hand, the training data for the Control group has 131 sentences, 700 6-grams, and 1793 1-skip-trigrams. Having stacked the discrete continuous vocabulary spaces together, we performed SVD and used the decomposed left-singular matrix from the SVD to map the vocabulary histories into a lower dimensional continuous parameter space for the neural network. The language model smoothing for words outside the output vocabulary was performed as described in [Schwenk, 2007]. The D2NNLM was trained with three hidden layers excluding the projection layer. In order to avoid the risk of reconstructing the identity function [Larochelle *et al.*, 2009], we performed a grid search and set the sizes of the hidden layers to 70% of the input neurons as the number of hidden units. Because the D2NNLM performs a classification task by discriminating between the MCI and the Control classes, we performed finetuning in the form of backpropagation instead of pretraining [Hinton *et al.*, 2012]. For the classification task, the network parameters are used to estimate the likelihood that a vocabulary feature sequence belongs to either the MCI or Control classes. While our training technique is mostly similar to DNNLM [Arisoy *et al.*, 2012], we set the initial learning rate to 0.01, momentum to 0.1, and the weight decay to 0.01, respectively. Finally, we trained the network to convergence.

We estimated the percentage Mean Square Error (MSE) using Pybrain's implementation of the neural network and validates on a held-out test set. The percentage error is used often to evaluate neural network models [Arisoy *et al.*, 2012]. We also estimated the language model perplexity of the D2NNLM in comparison to DNNLM and NNLM. In language modeling, perplexity measures how well a model predicts given examples using an information theoretic approach. A better model minimizes the perplexity. We compute the perplexity as $2^{B(q)}$ as follows:

$$B(q) = -\frac{1}{N}\sum_{i=1}^{N} log_2 q(x_i) \tag{5}$$

$$Perplexity = 2^{B(q)} \tag{6}$$

where $B(q)$ estimates the ability of the model to predict significant portion of the test samples.

5.3 Result Analysis

As shown in Table 2, we performed experiments by comparing the percentage error and perplexity between the D2NNLM, DNNLM, and NNLM. We compared the D2NNLM with DNNLM and NNLM because both models consolidate neural network with language models. Note that previous work have shown that NNLM and DNNLM outperform the conventional 4-gram language models [Schwenk, 2007; Arisoy *et al.*, 2012]. Similarly, in [Schwenk, 2007], NNLM outperformed the 4-gram back-off model even when the modified Kneser-Ney smoothing is used. While it makes sense to compare the D2NNLM with the Hierarchical Pitman-Yor language models (HPYLM) [Huang and Renals, 2007] and the sequence memoizer [Wood *et al.*, 2011], neither the original NNLM nor DNNLM has made such comparison even with very large datasets. Given the relatively small size of the MCI clinical data (i.e. 19 MCI and 19 Control), the parameters for the Pitman-Yor process (prior distribution) could be complex to optimize for the optimal performance of both HPYLM and sequence memoizer, which are Bayesian

[2] http://www.nia.nih.gov/

Models	(%) Error	Perplexity
D2NNLM (ngram=6, skip-gram=1-skip-trigram)	**12.5**	**1.6**
DNNLM (ngram=4)	62.5	3.1
NNLM (ngram=4)	37.5	2.6

Table 2: Percentage error and perplexity on held-out test set

models. Nevertheless, we consider this as part of our plan for future work on a large clinical dataset.

As discussed in Section 3, the three models have different architectures but used the same learning settings. For the 3 models, we used 70% of the input neurons as the number of hidden units for the hidden layers. The D2NNLM (3 hidden-layers) takes a stacked combination of 6-grams and 1-skip-trigram as inputs, while DNNLM (3 hidden layers) and NNLM (1 hidden layer) take only 4-grams as performed in [Arisoy *et al.*, 2012].

In Table 2, we see that the D2NNLM has a better percentage error of 12.5% and a much lesser perplexity of 1.6, which is comparably better than the DNNLM and NNLM. Interestingly, the single hidden layer NNLM showed a better percentage error and perplexity than the DNNLM with 3 hidden layers. One possible explanation for this behavior is that the higher number of hidden layers in the DNNLM does not necessarily correspond to improved performance especially on small datasets [Arisoy *et al.*, 2012]. On the other hand, increasing the feature dimension could lead to improved performance [Bengio *et al.*, 2003], which is why we introduced the D2NNLM with a much deeper language model vocabulary space by stacking together higher order n-gram and $skip$-gram features. We also observed that D2NNLM needed much longer training iterations (943 epochs) to converge. This is understandable considering the fact that the number of training sample is only 9 per category.

Figure 2 shows a comparison between the perplexities of the three models while varying the number of hidden units at their respective hidden layer(s). Note that all the three models used the same number of hidden units at each step. We see that the D2NNLM consistently shows lower perplexity with respect to the increasing number of hidden units and stabilizes at 70% of the input neurons, which is why we chose 70% of the input neurons as the number of hidden units in the previous experiment.

In Table 3, we show the improvements on the DNNLM and NNLM by using a much deeper language model vocabulary space instead of the 4-gram language model used in [Arisoy *et al.*, 2012](see Table 2). We see that both the percentage error and perplexity reduced considerably by using an increased feature dimension with 6-grams and 1-skip-trigrams. For DNNLM, a deeper vocabulary space reduced the percentage error by 60% and perplexity by 42%. Similarly for NNLM, the percentage error was reduced by 33% and perplexity by 27%. This further confirms the hypothesis that increased feature dimension could lead to improved network performance [Bengio *et al.*, 2003].

We also performed experiments to investigate the contributions of three different hidden layers (H = 2, 3, and 4) to both D2NNLM and DNNLM. Recall that NNLM is a single-layer

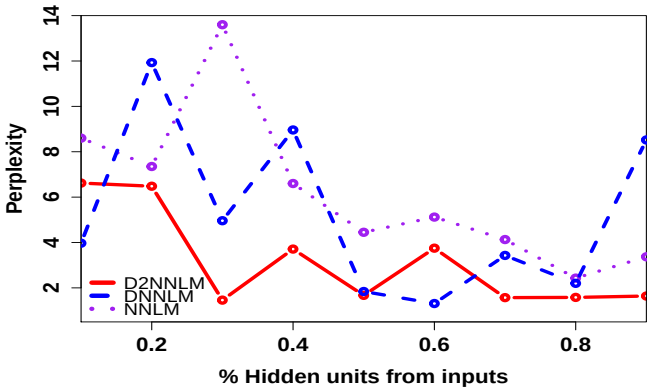

Figure 2: Comparing perplexity on held-out test set with varying hidden units

Models	(%) Error	Perplexity
DNNLM (ngram=6, skip-gram=1-skip-trigram)	25	1.8
NNLM (ngram=6, skip-gram=1-skip-trigram)	25	1.9

Table 3: Improvements on DNNLM and NNLM as a result of deeper vocabulary space

network, hence we cannot increase its hidden layer because doing so will make it a DNNLM. In addition, we included a third D2NNLM-6-grams without $skip$-gram features because we wanted to observe the impact of the $skip$-gram vocabulary space. Figure 3 shows the perplexity plot against different hidden layers of the models. We see that the D2NNLM has a much lower perplexity plot between 2 and 3 hidden layers compared to D2NNLM and D2NNLM-6-grams, albeit with an increased perplexity at the fourth hidden layer. We observed the optimal number of hidden layers to be 3 on the MCI dataset. DNNLM has a marginally better perplexity with 2 hidden layers but performed poorly with increased perplexity at the third and fourth layers. We observed the absence of the $skip$-gram vocabulary space to have substantial effect on the D2NNLM-6-grams with increased perplexity above the full D2NNLM. We believe that a combination of higher order n-grams and $skip$-grams with a maximum of 3 hidden layers led to an improved classification on the MCI dataset.

To put the performance of D2NNLM in clinical perspective, we achieved 87.5% accuracy on the held-out test set by computing the ratio of the correctly predicted samples to the total number of test samples. The performance is substantial considering the small size of the data set. Moreover, both DNNLM and NNLM gave comparatively lower accuracy of 62.5% on the test set, respectively. In comparison to the results in [Roark *et al.*, 2011], the D2NNLM is in a better position with better predictive accuracy because of its low error rate. Although other evaluations could be performed to substantiate the efficacy of our model in real clinical scenarios, nevertheless, we believe that our experimental results suffi-

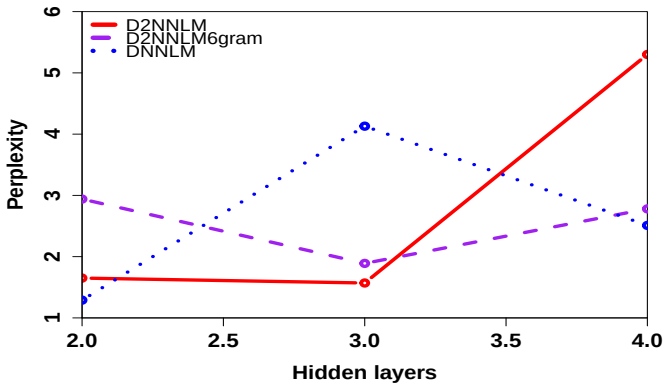

Figure 3: Comparing perplexity on held-out test set with varying hidden units

ciently show the potential of the deep-deep neural network language models for predicting MCI on very sparse clinical dataset.

6 Discussion and Limitations

The use of higher order n-gram and $skip$-gram features in this study is limited to the description of the Cookie-Theft picture in the DementiaBank clinical dataset. This is understandable since the objects within the picture dictate the specific n-gram and $skip$-gram features in the language space of the MCI and control individuals. Unless a picture with similar objects in the Cookie-Theft picture is used for collecting the speech transcript, the use of any other picture with different objects is likely to generate a different set of n-gram and $skip$-gram features.

We believe that D2NNLM could be effective for predicting other language and cognitive impairment related diseases such as Aphasia, Autism spectrum disorder, and Parkinson's disease. For example, linguistic defects could be more pronounced in Aphasia patients because Aphasia mainly affects language. Thus, if D2NNLM can be sensitive to predict MCI, it could as well predict other pronounced language impairments.

Also, our current work has not optimized D2NNLM to perform as a general language model. We focus on MCI, and the possibility of predicting MCI on very sparse clinical dataset. Nevertheless, we believe that future work could validate the D2NNLM on other natural language problems such as sentiment analysis.

Finally, although the dataset used in this study is small, we have used the DementiaBank dataset, which is the largest and publicly available clinical dataset on MCI to date. Most clinical datasets on MCI are self-collected over a short period of time and are mostly not publicly available largely due to ethical constraints. We are currently conducting a longitudinal study to collect large speech samples from several MCI patients across two continents.

7 Conclusion and Future work

In this paper, we proposed the combination of deep neural network and deep language models to predict MCI from clinical language dataset. We learned deep language models using higher order n-gram and $skip$-gram vocabulary spaces. Experimental results show that the model predicts MCI with less percentage error and language model perplexity. In the future, we will consider D2NNLM on other language impaired related diseases such as Aphasia and Parkinson's disease. We will also conduct further evaluations that could put our model to use on large speech samples in actual clinical scenarios.

Acknowledgments

This work was supported by the Malaysian Ministry of Education via the Fundamental Research Grant Scheme (FRGS) - FRGS/2/2014/ICT07/MUSM/03/1.

References

[Abbott, 2011] Alison Abbott. Dementia: a problem for our age. *Nature*, 475(7355):S2–S4, 2011.

[Albert *et al.*, 2011] Marilyn S Albert, Steven T DeKosky, Dennis Dickson, Bruno Dubois, Howard H Feldman, Nick C Fox, Anthony Gamst, David M Holtzman, William J Jagust, Ronald C Petersen, et al. The diagnosis of mild cognitive impairment due to alzheimers disease: Recommendations from the national institute on aging-alzheimers association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia*, 7(3):270–279, 2011.

[Arisoy *et al.*, 2012] Ebru Arisoy, Tara N Sainath, Brian Kingsbury, and Bhuvana Ramabhadran. Deep neural network language models. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 20–28. Association for Computational Linguistics, 2012.

[Ball *et al.*, 2009] Martin J Ball, Michael R Perkins, Nicole Müller, and Sara Howard. *The handbook of clinical linguistics: vol 56*. John Wiley & Sons, United States, 2009.

[Bengio *et al.*, 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[Chen and Chu, 2010] Stanley F Chen and Stephen M Chu. Enhanced word classing for model m. In *INTERSPEECH*, pages 1037–1040, 2010.

[Damian *et al.*, 2011] Anne M Damian, Sandra A Jacobson, Joseph G Hentz, Christine M Belden, Holly A Shill, Marwan N Sabbagh, John N Caviness, and Charles H Adler. The montreal cognitive assessment and the mini-mental state examination as screening instruments for cognitive impairment: item analyses and threshold scores. *Dementia and Geriatric Cognitive Disorders*, 31(2):126–131, 2011.

[Fraser *et al.*, 2014] Kathleen C Fraser, Jed A Meltzer, Naida L Graham, Carol Leonard, Graeme Hirst, Sandra E

Black, and Elizabeth Rochon. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60, 2014.

[Hinton *et al.*, 2012] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[Huang and Renals, 2007] Songfang Huang and Steve Renals. Hierarchical pitman-yor language models for asr in meetings. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 124–129. IEEE, 2007.

[Kim and Caine, 2014] Scott YH Kim and Eric D Caine. Utility and limits of the mini mental state examination in evaluating consent capacity in alzheimer's disease. *Psychiatric Services*, 2014.

[Larochelle *et al.*, 2009] Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. Exploring strategies for training deep neural networks. *The Journal of Machine Learning Research*, 10:1–40, 2009.

[Le *et al.*, 2011] Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists. *Literary and Linguistic Computing*, page fqr013, 2011.

[Li *et al.*, 2015] Yifeng Li, Chih-Yu Chen, and Wyeth W Wasserman. Deep feature selection: Theory and application to identify enhancers and promoters. In *Research in Computational Molecular Biology*, pages 205–217. Springer, 2015.

[Locke, 1997] John L Locke. A theory of neurolinguistic development. *Brain and Language*, 58(2):265–326, 1997.

[Mitchell, 2009] Alex J Mitchell. A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *Journal of Psychiatric Research*, 43(4):411–431, 2009.

[Mitolo *et al.*, 2015] Micaela Mitolo, Simona Gardini, Paolo Caffarra, Lucia Ronconi, Annalena Venneri, and Francesca Pazzaglia. Relationship between spatial ability, visuospatial working memory and self-assessed spatial orientation ability: a study in older adults. *Cognitive Processing*, 16(2):165–176, 2015.

[Orimaye *et al.*, 2015] Sylvester Olubolu Orimaye, Kah Yee Tai, Jojo Sze-Meng Wong, and Chee Piau Wong. Learning linguistic biomarkers for predicting mild cognitive impairment using compound skip-grams. *arXiv preprint arXiv:1511.02436*, 2015.

[Pozueta *et al.*, 2011] Ana Pozueta, Eloy Rodríguez-Rodríguez, José L Vazquez-Higuera, Ignacio Mateo, Pascual Sánchez-Juan, Soraya González-Perez, José Berciano, and Onofre Combarros. Detection of early alzheimer's disease in mci patients by the combination of mmse and an episodic memory test. *BMC Neurology*, 11(1):78, 2011.

[Prud'hommeaux and Roark, 2015] Emily Prud'hommeaux and Brian Roark. Graph-based word alignment for clinical language evaluation. *Computational Linguistics*, 2015.

[Roark *et al.*, 2011] Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. Spoken language derived measures for detecting mild cognitive impairment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2081–2090, 2011.

[Sarikaya *et al.*, 2009] Ruhi Sarikaya, Mohamed Afify, and Brian Kingsbury. Tied-mixture language modeling in continuous space. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 459–467. Association for Computational Linguistics, 2009.

[Schwenk, 2007] Holger Schwenk. Continuous space language models. *Computer Speech & Language*, 21(3):492–518, 2007.

[Sidorov *et al.*, 2014] Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860, 2014.

[Tillas, 2015] Alexandros Tillas. Language as grist to the mill of cognition. *Cognitive Processing*, pages 1–25, 2015.

[Wood *et al.*, 2011] Frank Wood, Jan Gasthaus, Cédric Archambeau, Lancelot James, and Yee Whye Teh. The sequence memoizer. *Communications of the ACM*, 54(2):91–98, 2011.