Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Fall risk assessment through a synergistic multi-source DNN learning model

Olga Andreeva [a,d], Wei Ding [a,d], Suzanne G. Leveille [b,d,f,g], Yurun Cai [e], Ping Chen [c,d,*]

[a] *Department of Computer Science, College of Science and Mathematics, United States of America*
[b] *Departments of Nursing, College of Nursing and Health Sciences, United States of America*
[c] *Department of Engineering, College of Science and Mathematics, United States of America*
[d] *University of Massachusetts Boston, Boston, MA, United States of America*
[e] *Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States of America*
[f] *Department of Medicine, Harvard Medical School, Boston, MA, United States of America*
[g] *Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States of America*

ARTICLE INFO

Keywords:
Deep learning
Fall risk factors
Attention
Multi-source learning
Multi-label classification

ABSTRACT

Falls are a complex problem and play a leading role in the development of disabilities in the older population. While fall detection systems are important, it is also essential to work on fall preventive strategies, which will have the most significant impact in reducing disability in the elderly. In this work, we explore a prospective cohort study, specifically designed for examining novel risk factors for falls in community-living older adults. Various types of data were acquired that are common for real-world applications. Learning from multiple data sources often leads to more valuable findings than any of the data sources can provide alone. However, simply merging features from disparate datasets usually will not produce a synergy effect. Hence, it becomes crucial to properly manage the synergy, complementarity, and conflicts that arise in multi-source learning. In this work, we propose a multi-source learning approach called the Synergy LSTM model, which exploits complementarity among textual fall descriptions together with people's physical characteristics. We further use the learned complementarities to evaluate fall risk factors present in the data. Experiment results show that our Synergy LSTM model can significantly improve classification performance and capture meaningful relations between data from multiple sources.

1. Introduction

Falls are the second leading cause of accidental or unintentional injury deaths worldwide and the sixth leading cause of death in the older population of the United States. According to the U.S. Center for Disease Control and Prevention, 59,082 deaths from fall related injuries occurred among people aged 70 and older in 2019. Due to the potentially severe consequences, which include fractures and potential disabilities, a growing number of older adults fear falling and, as a result, limit their activities and social engagements. This can lead to further physical decline, depression, social isolation, and feelings of helplessness. Studies on this topic are of great importance and a lot of effort has been put into automatic fall detection [1–4]. While a fall detection is important, preventive strategies are also essential. A proper analysis of existing and novel fall risk factors can lead to a proper design of fall prevention strategies, which will have the most significant impact in reducing disability in the older population [5–8].

This work focuses on two major aspects: fall environment classification and identification of novel risk factors that contribute to falls, since they are also risk factors for several other adverse consequences in older adults. Fall risk factors are often classified as person specific (or intrinsic) and environmental (or extrinsic). Thus, it is also important to take into consideration the physical characteristics of monitored people. Due to the nature of the task, the quantity and availability of fall event data is low compared to other applications.

In this work we explore a prospective cohort study, specifically designed for examining novel risk factors for falls in community-living older adults, Maintenance Of Balance, Independent Living, Intellect, and Zest in the Elderly of Boston Study (MBS) [9]. Various types of data were acquired that are common for real-world applications — including

* Corresponding author at: Department of Engineering, University of Massachusetts Boston, 100 William T. Morrissey Blvd., Office: S-3-115 Science Building, Boston, MA 02125-3393, United States of America.

E-mail addresses: olga.andreeva001@umb.edu (O. Andreeva), wei.ding@umb.edu (W. Ding), suzanne.leveille@umb.edu (S.G. Leveille), ycai28@jhu.edu (Y. Cai), ping.chen@umb.edu (P. Chen).

https://doi.org/10.1016/j.artmed.2022.102280
Received 17 August 2021; Received in revised form 7 March 2022; Accepted 8 March 2022
Available online 18 March 2022

textual fall descriptions and numeric characteristics. Oftentimes, such datasets consist of multiple modalities, each having its own feature set, distribution, scale, and other characteristics. When these modalities describe the same sample — e.g., any multimedia segment can be simultaneously described by its video and audio signals — they are called views and are actively used in a multi-view learning model. This work focuses on a more general multi-source scenario, where distinct data sources describe the same phenomenon/entity (i.e., a fall environment), but were not necessarily collected at the same time. Many multi-source and multi-view learning models show that the synergy of all data from each source may yield a superior learning model — i.e., "the whole is greater than the sum of its parts". Multiple data sources that describe the same phenomenon/entity provide more comprehensive information and may allow for better generalization capability. Furthermore, they can provide complementary information, which is unlearnable in each of the individual sources alone. How to properly manage the synergies and complementarities across multiple — usually disparate — data sources is currently an important research topic in the AI community.

In many tasks, multi-source learning approaches exploit different data modalities. These approaches are especially crucial in case the collected textual data is sparse and does not contain sufficient knowledge for a learning model. For instance, short texts, unlike any large corpora, do not have enough contextual information, which poses a great challenge for many natural language processing (NLP) tasks. In many fields, data collection process is usually excessive. Researchers often adopt different collection procedures or mechanisms, which can result in data of a different form or modality. For example, during a medical screening, a nurse measures a patient's temperature and asks about the patient's well-being. Without any context, 90 °F and 99 °F do not provide any substantial information for a machine learning model, besides a 2° numeric difference. However, together with the patient's textual description: "I feel great", or "I have a fever" – 97 °F and 99 °F acquire a meaningful difference, which becomes possible for a machine learning model to learn or capture. How to take advantage of synergy among different data views (e.g., texts and other sources of data) is still an open question. In this work, we explore data synergy under a specific task of multi-label fall environment classification based on short text description and personal characteristics. We deal with a dynamic scenario where two divergent data sources — short texts, accompanied by numeric statistics — exist to describe an entity and its associated events. We introduce the Synergy block and incorporate it into an LSTM model. Together, they are capable of learning complementarities among multiple data sources. Experiment results show that our synergy LSTM model can significantly improve performance in a supervised learning setting.

This paper introduces three novel contributions to the task of fall classification and fall risk identification:

- We introduce the Synergy block, which successfully learns complementarities among multiple data sources. Through the evaluation process we show that an attention-based Synergy block helps to identify dependencies between fall conditions, environments, and physical characteristics of monitored adults.
- We build a multi-source LSTM model integrated with the Synergy block, which takes advantage of synergy among multiple data forms.
- We use the learned synergy to identify possible novel fall risk factors in the older population.

2. Related work

The idea to use data from multiple sources/views has been around for a long time [10]. In multi-view learning scenario, the same object or event is described from different perspectives or views, e.g., English and Russian versions of the same text. Typically, each of these views is fed into a model and learned embedding is jointly optimized to improve generalization performance. Several reviews exist for this matter [11,12]. This work focuses on a more general multi-source scenario, where distinct data sources describe the same phenomenon/entity but were not necessarily collected at the same time. While most works focus on image/video with text or audio with text scenarios [13], in this paper we explore a challenging real-world problem, where numeric statistics and short texts work in tandem to provide insight on fall risk factors in the elderly population. In NLP tasks, multi-source learning approaches often use different data modalities, e.g., audio, video, and text. Several reviews exist regarding this subject [14–16]. Multi-modal strategies have been extensively explored by the audio-video-speech recognition community [17]. Multi-modal fusion has a very broad range of applications, including multi-modal emotion recognition [18], medical image analysis [19], cancer prediction [20], multimedia event detection [21], and multi-modal natural language models [22]. Strategies exploiting attention mechanism [23] were applied to video captioning [24] and video description generation tasks [25]. In [26] authors used attention-based multi-view representations of graph nodes for node classification and link prediction tasks. In [27] authors addressed the task of entity typing for multi-view entity representations by incorporating attention into the fusion process. All these approaches target specific tasks, different in each case. For the fair comparison, we did not include these approaches as our baselines.

3. Methodology

In this section, we describe how to exploit synergy that exists in multiple data views through Synergy block. We hypothesize that our Synergy block is capable of learning complementarities in multiple sources to successfully classify falls described by short texts. Then, we will use the learned complementarities to identify fall risk factors present in the data.

3.1. Dataset description

A diverse group of seniors, aged 70 years and older, were included in the MBS study, where a fall is defined as unintentionally coming to rest on the ground or other lower level not because of a major intrinsic event (e.g., myocardial infarction or stroke) or an overwhelming external hazard (e.g., hit by a vehicle) [28]. We use two types of data collected: health interviews (including information on age, gender, race, education, and pain) and post-fall descriptions that describe circumstances and consequences of each fall during the 18-month follow-up. Analysis of such data is very challenging. First, this dataset is sparse due to having a small number of participants and falls (from the point of view of machine learning research. However, in medical community this study is one of the largest efforts in collecting older adult fall data). Secondly, this dataset is noisy due to variability in fall descriptions, and a fall can be described differently even by the same person. These two types of data provide two sources describing latent features that characterize participants' falls.

3.1.1. Health interview data

During health interviews the following information was collected: participant's age, gender, educational group, race group, and measures

Table 1
Health interview data characteristics.

Variable	Values
gender	0|1
age	70–97
education group	0|1|2
race group	0|1|2
pain count	0|1|2
BPIsev	0–8.7
BPIinterf	0–10

of chronic pain as shown in Table 1. Three different pain scales were used to assess chronic pains: two Brief Pain Inventory (BPI) subscales and a 3-level pain count measure.

BPI subscales — BPI severity (BPIsev) and BPI interference (BPIinterf) — measure global pain severity and pain interference [29,30]. For the severity subscale, older adults rated their pain, which is described as pain “you have today that you have experienced for more than just a week or two,” using a 4-item severity scale. Participants rate their pain in the previous week on a numeric rating scale from 0 to 10, where 0 reflects ‘no pain’ and 10 reflects ‘severe or excruciating pain, as bad as you can imagine.’ The 4 separate items include pain at its worst, least pain, pain on average, and pain now. The BPI severity value is scored as the average of the 4 ratings.

The BPI interference subscale was used to rate level of pain interference with general activity, mood, walking, normal work including housework, relations with other people, sleep, and enjoyment of life. Rating for each item was on a 0–10 numeric rating scale, with 0 indicating no interference and 10 indicating complete interference. The BPIinterf score was calculated as the average of the 7 aforementioned ratings.

The pain count variable is a 3-level measure of chronic musculoskeletal pain representing no pain (0), single-site pain (1), or multi-site pain (2). The measure is derived from the MBS joint pain questionnaire, measuring pain present in the past month and lasting 3 or more months in the previous year, at 6 musculoskeletal sites: shoulder, hand/wrists, back, hips, knees, feet [31].

Health interview data includes statistics for 314 participants who had at least 1 fall during the follow-up period.

3.1.2. Post-fall description data

Participants were given a set of monthly fall calendars and instructed to mark an “F” on the days that a fall occurred and an “N” for each day that no fall occurred. At the end of each month during a follow-up period, participants mailed their calendar postcards to the study center. Whenever a fall was reported, study staff conducted a structured telephone interview to determine the circumstances and location of the fall, injuries sustained, and the presence of external and internal factors that may have contributed to the fall. All of this comprehensive information was saved in phone logs. Participants whose calendar was missing were contacted by telephone to determine whether a fall occurred in the previous month.

Post-fall description data used in our experiment includes only a short textual description of a fall given directly by a participant, which may not contain all necessary information to characterize a fall as the comprehensive phone interview logs. The reasons we chose these incomplete descriptions instead of comprehensive phone logs are twofold. Firstly, incomplete or missing data is very common in real-world applications, and we want to assess the performance of our model in a more realistic setting. Secondly, such a setting may reveal the true capability of our synergy model in multi-source data learning scenarios. In total the dataset contains 1721 fall descriptions, where max description length is 58 words, mean length is 21.6, and dataset's vocabulary consists of 2238 words.

3.1.3. Annotation and pre-processing

To evaluate our Synergy LSTM model, post-fall description data needs to be annotated. Two annotators characterized each fall description with the most appropriate labels in six categories shown in Table 2. Altogether, these six categories describe circumstances in which a fall occurs.

A couple of sources were important for this annotation process. First, the scripted fall descriptions themselves. Each sentence contains information that describes a scenario in which a fall occurred. When the annotator had difficulty in making a label assignment, they referred to the second source – the phone interview logs. These logs provide all necessary fall details, not reflected in the fall description. Annotators were constrained to selecting only one label for each category, and in case of ambiguity, they resolved it with the domain expert.

Table 2
An overview of the labels set for the MBS study. First column indicates the category. Second column shows the total amount of labels in each category. Third column lists all possible labels in each category.

Category	#	Labels
1: Did it happen outside/ inside?	2	Inside, outside
2: What was the person doing?	10	Getting in/out of a chair/sofa, getting in/out of a vehicle, going downstairs, going upstairs, lying, sitting, standing, stepping on/off a curb, walking, other activity
3: Where was the person when falling?	17	Basement or cellar, bathroom, bedroom, curb, dining room, escalator, garden/yard, hallway, kitchen, living room, moving walkways, parking lot, sidewalk, stairs, street, train/bus, other location
4: Was the person hurt?	2	Hurt, no hurt
5: Did the person trip or slip?	2	Tripped/slipped, did not trip/slip
6: What were surface conditions?	4	Dry, icy, wet, other surface conditions

Here is the annotation process: First, annotators read a sentence and determine a label for each category according to Table 2. Annotators were constrained to selecting only one label for each category. In case of ambiguity or if a description is not complete, i.e., some labels cannot be determined from a fall description, annotators referred to phone interview logs. For instance, “*I was walking on the street in front of my house, lost my balance and fell on the ground*.” From this description an annotator can determine that the fall happened *outside* on the *street*, and that participant was *walking*. To determine the other labels, annotators need to look at the logs, which record that the participant did *hurt* himself, he had *not tripped or slipped*, and the surface condition was *dry*. As a result, the final label for this fall description is *[outside, walking, street, hurt, no tripped/slipped, dry]*. As was mentioned earlier, we do not use phone logs during the training, thus, to determine the label not presented in texts (e.g., *hurt* in this case), our model would have to exploit learned synergy from other sources.

3.2. Data pre-processing

Data pre-processing was divided into three parts: post-fall phone interviews pre-processing, pre-processing of health interviews, labels' pre-processing. All fall scenarios were collected during post-fall phone interviews, where each response was typed into a computer by a staff member. Thus, collected fall descriptions contain misspellings, abbreviations, and acronyms. To address these problems and improve the quality of our data, we manually checked each description and corrected all misspellings, and substituted abbreviations and acronyms with corresponding full words. This dataset has no technical or ambiguous abbreviations, but rather everyday ones; for instance, ASAP standing for “as soon as possible”. There was no additional pre-processing done for the health interviews. Finally, we combined the post-fall descriptions and health interviews together based on a participant ID.

In data pre-processing each label (refer to Table 2) was assigned a unique value — 37 values in total. We used a 5-fold-cross-validation approach for evaluation, where for every fold the multi-modal dataset was divided into training and test sets. For every fold we converted training sentences into sequences of word ids using the vocabulary of the training dataset. After that, we padded each sequence to reach the maximum sequence length in every fold.

3.3. A synergy LSTM model

The motivation for the Synergy model comes from a simple, but compelling idea. An entity or event is characterized by its features, and these features are often latent or hidden and cannot be measured directly. One measuring or collecting method may only be able to acquire certain features that provide a partial description of an entity. Thus, it is typical to describe such entity with multiple forms, views, or modalities. Existing machine learning methods can capture valuable information from these disparate sources, i.e., create a partial perception model (*ppm*) for each view. Each partial perception model describes hidden latent features for a certain data view or source. Because all of these partial perception models represent a partial set of the same latent features, we can integrate these partial representations into a more comprehensive perception model (*cpm*) to better characterize an entity. In this way we can successfully exploit synergy existing among different data sources.

3.4. *Synergy block*

We propose an approach that utilizes attention mechanism [23] to exploit synergy existing between different data forms. We call this architecture the Synergy block. Fig. 1 illustrates the Synergy block architecture, and we will discuss how to incorporate the Synergy block into a complete LSTM model (Fig. 2) in Section 3.5.

Synergy block is an intermediate attention-based mechanism that receives N sets of inputs from N sources: $\{a_1{}^{<j>}, j = 1\ldots|a_1|\}$ – is a by-product of source 1, …, $\{a_N{}^{<j>}, j = 1\ldots|a_N|\}$ – is a by-product of source N. At a time-step $<t>$ our Synergy block encodes synergy between different data sources, represented by $ppm_i{}^{<t>}$, into a comprehensive perception model $cpm^{<t>}$. To be precise, the Synergy block learns $cpm^{<t>}$ according to Eq. (1).

$$cpm^{<t>} = \Delta\left(\left[ppm_1^{<t>}; \ldots; ppm_N^{<t>}\right]\right) \quad (1)$$

Here $ppm_i{}^{<t>}$ represents a partial perception model, computed for the data from a source i, and $cpm^{<t>}$ represents a comprehensive perception model. Depending on the nature of your data, Δ can be represented as an *RNN*-based network, if your $ppm_i{}^{<t>}$ is sequential, or *CNN*-based, if your $ppm_i{}^{<t>}$ has spatial correlation.

Each $ppm_i{}^{<t>}$ follows from Eqs. (2), (3), (4). Similar to attention mechanism, for each set of inputs – $\{a_1, a_2, \ldots, a_N\}$, where $a_i = \{a_i{}^{<j>}, j = 1\ldots|a_i|\}$ – we first compute a set of energies $\{e_i{}^{<t, j>}, j = 1\ldots|a_i|\}$ according to Eq. (2).

$$e_i^{<t,j>} = v_i^{\mathrm{T}} \tau_i\left(W_i\left[\gamma_i\left(s^{<t-1>}\right); a_i^{<j>}\right]\right) \quad (2)$$

Here $\gamma_i(\cdot)$ is a transformation function (linear or non-linear), $\tau_i(\cdot)$ is a non-linear transformation function, v_i is a vector, W_i is a matrix and $e_i{}^{<t,j>}$ is a scalar. $s^{<t-1>}$ is a hidden state from the previous time step $<t-1>$ (see Section 3.5).

$$\varepsilon_i^{<t,j>} = \frac{exp\left(e_i^{<t,j>}\right)}{\sum_{k=1}^{|a_i|} exp\left(e_i^{<t,k>}\right)} \quad (3)$$

$\{\varepsilon_i{}^{<t, j>}, j = 1\ldots|a_i|\}$ is a set of attention weights, which are used to compute partial perception model $ppm_i{}^{<t>}$ according to Eq. (4).

$$ppm_i^{<t>} = \sum_{j=1}^{|a_i|} \varepsilon_i^{<t,j>} a_i^{<j>} \quad (4)$$

The aforementioned process is illustrated in Fig. 1. The Synergy block uses by-products of each view and 1) learns which parts of these inputs are currently more important than the others - this would be encoded in partial perception models; 2) what is the best way to combine those important features from multiple views – encoded in the comprehensive perception model.

3.5. *LSTM-based synergy model*

In this paper we hypothesize that our Synergy block is capable of learning complementarities existing among multiple data views. Our Synergy block architecture described above can be integrated into various deep learning architectures. To make our idea more concrete, in this section we look into a specific task of multi-label short text classification with two views: textual and numerical. For this task, we will build a multi-view model with the Synergy block – the Synergy LSTM model. The process is described as follows. Each view's input sequence – GloVe embeddings for textual input and numeric statistics described in Table 1 – is fed to a corresponding learning model to extract informative features. For our task, we use bidirectional *LSTM* as the learning model. A resulted output can be treated as informational embeddings $\{a_1, a_2\}$, where $a_j = \{a_j{}^{<i>}, i = 1\ldots|a_i|, j = 1, 2\}$. Next, we feed these embeddings into the Synergy block.

The task requires our model to classify an input into $K = 6$ not mutually exclusive categories, i.e. multi-label classification. Table 2 provides a comprehensive break down of each category's class labels. We

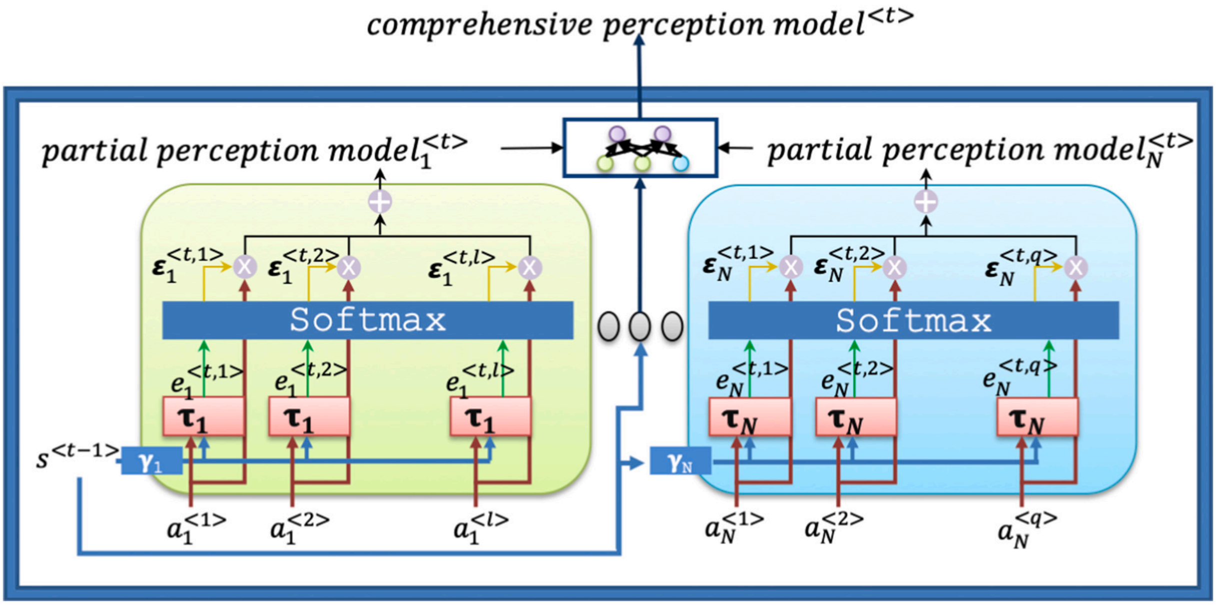

Fig. 1. Synergy block utilizes attention mechanism to learn 1) multiple partial perception models and 2) a comprehensive perception model.

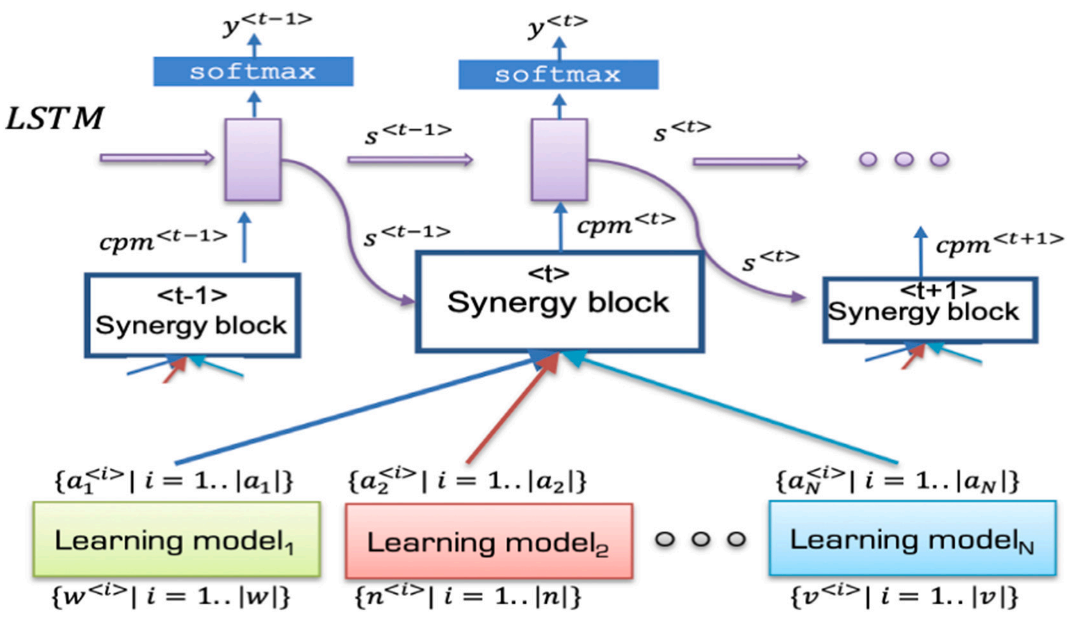

Fig. 2. Synergy model with incorporated synergy block. Here $\{w^{<i>}|i = 1\ldots|w|\}$, $\{n^{<i>}|i = 1\ldots|n|\}$ and $\{v^{<k>}|i = 1\ldots|v|\}$ represent different data from different sources; $cpm^{<t>}$ is a comprehensive learned model at time step t.

follow a common approach used for such tasks – a transformation into a multi-class problem. After this transformation a ground truth class is a combination of every label in the training set, e.g., for labels A, B and C $[1,1,0]$ denotes a situation, when A and B are present, and C is absent.

Our final layer is an *LSTM* layer on top of the Synergy block. The necessity of this layer is twofold: 1) for a time step $<t>$ *LSTM*'s output is used to make a prediction for the category t, where $t = 1\ldots K$; 2) this *LSTM* passes the previous hidden state $s^{<t-1>}$ to the Synergy block on the time step $<t>$. Based on inputs from different views and the hidden state $s^{<t-1>}$, our Synergy block computes the comprehensive perception model $cpm^{<t>}$ for the time step $<t>$, which is used by *LSTM* to make a prediction $y^{<t>}$ for the category t.

In our task we use the *BiLSTM* network due to its important property: each *BiLSTM*'s output encodes information from the past and the future. However, depending on the input's modality, extraction of informative features can be performed by any other network, e.g. *CNN* for images. The general case of the Synergy model for N sources is illustrated in Fig. 2. Here the set $\{w^{<i>}|i = 1\ldots|w|\}$ represents the input from the first data source, $\{n^{<i>}|i = 1\ldots|n|\}$ represents the input from the second data source, and $\{v^{<k>}|i = 1\ldots|v|\}$ - from the source N.

4. Results

To demonstrate the effectiveness of the proposed technique for our problem, it was compared with several baselines as follows. At first, we explored if existing short fall descriptions contained enough information to successfully perform multi-label classification. For this task we used a simple LSTM-based model (Uni-modal LSTM). The second group of baselines includes a typical multi-modal fusion approach, when learned embeddings are simply concatenated with each other and fed into the next stage of a model. We compared the proposed Synergy Block with two different fusion techniques including early fusion (similar to [32]) and late fusion (similar to [33]). In early fusion, the input features are concatenated and then we apply BiLSTM to generate features followed by LSTM to produce the final classification. On the other hand, the late fusion concatenates the learned embeddings (outputs of learning models – refer to Fig. 2) from each modality and apply the LSTM for classification. Finally, we considered a tensor fusion that explicitly models n-modal inter-modal interactions using a Cartesian product from modality embeddings [34].

We performed a 5-fold cross-validation and for each model we reported an average among 5 folds. We evaluated performance of each method according to 2 metrics Hamming Loss (HL) and weighted F_1 score (F_{1w}), which considers class imbalance, existing in non-binary categories. With F_{1w} (Eq. (6)) we can estimate how well each classifier generalizes in each of our 6 categories and HL is used to evaluate a multi-label performance, since it represents the proportion of the misclassified labels to the total number of labels.

$$HL = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i \oplus P_i}{K} \tag{5}$$

Here N represents the total number of samples; K – the total amount of categories (6 in our case); Y_i and P_i – the ground truth and predicted labels respectively.

$$F_{1w} = \frac{1}{\sum_{l\in L}|y_l|}\sum_{l\in L}|y_l|F_1(\widehat{y}_l, y_l) \tag{6}$$

Here y and $\widehat{y}$ represent sets of true and predicted labels respectively; y_l – is the subset of y with label l, and similarly, $\widehat{y}_l$ – the subset of $\widehat{y}$ with label l.

Table 3 shows model performance evaluated by HL and F_{1w}. Average score is the average performance calculated based on 5 folds. It becomes clear that our Synergy model outperforms all baselines according to both metrics and provides a significant improvement not only in categories 1–6 separately, but also produces less mistakes in a final complex label (has the lowest HL). Low performance of a uni-modal LSTM (low F_1w and high HL) supports the fact that there is not enough information in short fall descriptions to successfully perform multi-label classification (i.e., infer environment in which fall occurred). Additional numerical statistics seem to add meaningful relationships, which improve situation for all categories as shown in Table 3. However, models with a simple fusion (early and late) or a more complex Cartesian fusion are still having a hard time with the most challenging categories 2 and 3. Thus, simply merging two data modalities does not produce a desirable synergy effect. On the other hand, the Synergy model shows the best performance among all evaluated models and effectively uses diverse characteristics within the data to differentiate between multiple classes in categories 2, 3, 5, 6.

As to the time it takes to converge during training, a Cartesian fusion

Table 3
Classification performance: weighted F_1 (the higher the better) and Hamming Loss (HL) (the lower the better).

Model	HL	Weighted F_1					
		Category 1	Category 2	Category 3	Category 4	Category 5	Category 6
Uni-modal LSTM	0.4627 (±0.0066)	0.384 (±0.0055)	0.366 (±0.0343)	0.086 (±0.0121)	0.414 (±0.0351)	0.386 (±0.0182)	0.718 (±0.0342)
Multi-modal LSTM (early fusion) [32]	0.3457 (±0.0174)	0.876 (±0.0195)	0.396 (±0.0532)	0.2314 (±0.1248)	0.566 (±0.0773)	0.714 (±0.0932)	0.82 (±0.0552)
Multi-modal LSTM (late fusion) [33]	0.3625 (±0.0292)	0.878 (±0.0179)	0.382 (±0.0549)	0.264 (±0.0695)	0.574 (±0.041)	0.616 (±0.068)	0.796 (±0.0841)
Multi-modal LSTM (Cartesian fusion) [34]	0.6055 (±0.1727)	0.806 (±0.1496)	0.3044 (±0.2047)	0.2576 (±0.1696)	0.261 (±0.2379)	0.248 (±0.0536)	0.664 (±0.1683)
Synergy model	**0.2729** (±0.0117)	**0.89** (±0.018)	**0.542** (±0.053)	**0.51** (±0.0515)	**0.598** (±0.045)	**0.82** (±0.0291)	**0.852** (±0.037)

multi-view model takes twice as many epochs comparing to our Synergy model. Additionally, construction of a Cartesian fusion takes at least $\mathcal{O}(n^2)$ (in the case of 2 views) and extra space to store this product, which significantly slows down its training.

So far, we have shown how complementary information learned from short texts and numeric values help our Synergy model to achieve the best performance in multi-label classification. In the following example we illustrate an achieved consensus in the data, i.e., the manner how different data modalities embed a compatible (or correlated) latent structure within the data. Fig. 3 illustrates the distribution of attention weights our Synergy model computed during the prediction phase for one test sample: the fall description is: '*I was maybe in my yard, lost my balance and fell.*', and corresponding numeric values are: [2, 75, 2, 5.25, 4.57, 3, 2]. In Fig. 3, the horizontal axis represents textual and numeric values of this sample, and the vertical axis shows the predicted labels (in the middle): [outside, walking, other, no hurt, no tripped/slipped, dry]. For each category, Fig. 3 shows the distribution of attention among words in the sentence and among numeric variables. Darker colors represent higher concentration of attention. For instance, to predict that the fall happened outside, the model concentrates more on the "yard"; at the same time, 2 BPI pain ratings seem to also contribute towards this decision. On the other hand, an activity (category 2) is not clearly stated in the fall description. In this case, it seems like the Synergy model makes a decision based on a consensus between 2 sources: "yard" and 2 BPI pain ratings. Other categories seem to be more complex. For example, the "hurt/no hurt" decision was made based on the texts "in my yard" and "my balance", and an equal contribution from all numeric statistics. This shows an ability of our model to successfully use the latent structure of our data to determine information missing from the texts. We can further evaluate which numeric statistics are used more often to classify the fall.

Fig. 4 shows a combined attention matrix among 5 folds. To build this matrix, we combined attention matrices computed by a trained model for each test case in every fold. It seems to be a general trend that BPIsev plays the key role during the prediction of the first 4 categories. Similar results were found during the previous analyses of the MBS study

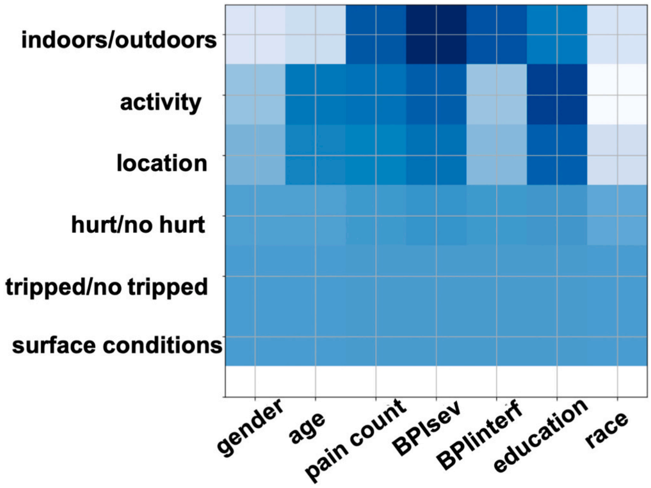

Fig. 4. Combined attention matrix among 5 folds. The darker the color is the more attention model puts towards the metric.

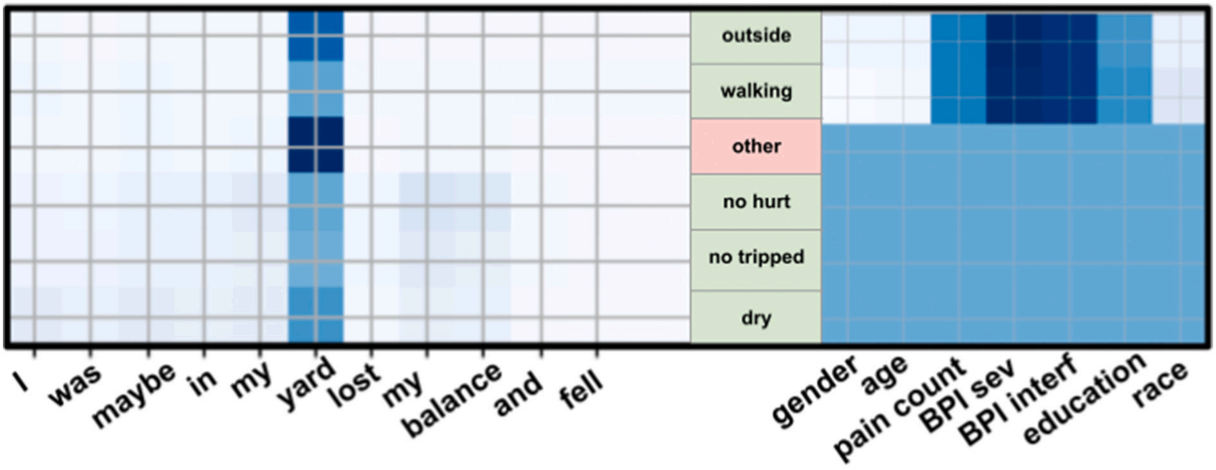

Fig. 3. Attention distribution for each of input views for the test case sample: 'I was maybe in my yard, lost my balance and fell.' – [2, 75, 2, 5.25, 4.57, 3, 2]. The horizontal axis represents an input, and vertical axis – predicted label: [inside, walking, living room, no hurt, no tripped/slipped, dry]. Green color represents the correctly predicted label. Best in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

[31]. In that study, results showed that chronic pain, whether measured by pain count, pain severity or pain interference, was associated with an increased rate of falls in older adults. It is also clear that to predict an activity and location, in addition to pain measurements (except BPIinterf), the Synergy model concentrates heavily on education group and age. Moreover, gender and race do not play a major role in predictions. As for the last two categories, the Synergy model was not able to find any distinct numeric statistic that may play a major role in the prediction. This may be due to the fact that there are no strong correlations between these categories and numeric inputs, and more statistics need to be used. While some of these findings are supported by existing research, it may be worth looking further into correlation between education and fall environment as an additional direction for evaluating fall risk factors.

5. Conclusion

In this paper, we presented the Synergy block - an intermediate attention-based mechanism, which can successfully learn complementarities among multiple data sources. We also show how to incorporate the Synergy block into a multi-source Synergy LSTM architecture and capture more comprehensive information that each data source cannot provide individually. We applied the Synergy model to a challenging task of fall classification and fall risk factors identification. Falls are classified based on short-text descriptions of the fall and physical characteristics of the person. Results show high levels of performance in classifying falls, even when several details are not present in the description (e.g., activity is not explicitly stated). The results demonstrate the capability of using learned complementarities to successfully identify fall risk factors for future investigation in a medical study.

Acknowledgement

This material is based upon work partially supported by (for Dr. Wei Ding, while serving at) the National Science Foundation under NSF grants IIS 1914489 and IIS 2008202. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Galvão YM, Ferreira J, Albuquerque VA, Barros P, Fernandes BJ. A multimodal approach using deep learning for fall detection. Expert SystApplic 2021;168:114226. https://doi.org/10.1016/j.eswa.2020.114226. https://www.sciencedirect.com/science/article/pii/S0957417420309489.

[2] Mastorakis G, Ellis T, Makris D. Fall detection without people: a simulation approach tackling video data scarcity. Expert SystApplics 2018;112:125–37. https://doi.org/10.1016/j.eswa.2018.06.019. https://www.sciencedirect.com/science/article/pii/S0957417418303658.

[3] Guvensan MA, Kansiz AO, Camgoz NC, Turkmen HI, Yavuz AG, Karsligil ME. An energy-efficient multi-tier architecture for fall detection on smartphones. Sensors 2017;17(7). https://doi.org/10.3390/s17071487. https://www.mdpi.com/1424-8220/17/7/1487, https://www.mdpi.com/1424-8220/17/7/1487.

[4] Boutellaa E, Kerdjidj O, Ghanem K. Covariance matrix based fall detection from multiple wearable sensors. J Biomed Inform 2019;94:103189. https://doi.org/10.1016/j.jbi.2019.103189. https://www.sciencedirect.com/science/article/pii/S1532046419301078.

[5] Phelan EA, Ritchey K. Fall prevention in community-dwelling older adults. Ann Intern Med 2018;169(11):ITC81–96.

[6] Hamm J, Money AG, Atwal A, Paraskevopoulos I. Fall prevention intervention technologies: a conceptual framework and survey of the state of the art. J Biomed Inform 2016;59:319–45. https://doi.org/10.1016/j.jbi.2015.12.013. https://www.sciencedirect.com/science/article/pii/S1532046415002932.

[7] Rivolta MW, Aktaruzzaman M, Rizzo G, Lafortuna CL, Ferrarin M, Bovi G, Bonardi DR, Caspani A, Sassi R. Evaluation of the tinetti score and fall risk assessment via accelerometry-based movement analysis. Artif Intell Med 2019;95:38–47. https://doi.org/10.1016/j.artmed.2018.08.005. https://www.sciencedirect.com/science/article/pii/S0933365717303901.

[8] de Souto Barreto P, Rolland Y, Vellas B, Maltais M. Association of long-term exercise training with risk of falls, fractures, hospitalizations, and mortality in older adults: a systematic review and meta-analysis. JAMA Intern Med 2019;179(3):394–405. https://doi.org/10.1001/jamainternmed.2018.5406. URL doi: 10.1001/jamainternmed.2018.5406.

[9] Leveille SG, Kiel DP, Jones RN, Roman A, Hannan MT, Sorond FA, Kang HG, Samelson EJ, Gagnon M, Freeman M, et al. The mobilize Boston study: design and methods of a prospective cohort study of novel risk factors for falls in an older population. BMC Geriatr 2008;8(1):16.

[10] Mitchell TM. Machine learning and data mining. CommunACM 1999;42(11):31.

[11] Yan X, Hu S, Mao Y, Ye Y, Yu H. Deep multi-view learning methods: a review. Neurocomputing 2021;448:106–29. https://doi.org/10.1016/j.neucom.2021.03.090. https://www.sciencedirect.com/science/article/pii/S0925231221004768.

[12] Zhao J, Xie X, Xu X, Sun S. Multi-view learning overview: recent progress and new challenges. InformFusion 2017;38:43–54.

[13] Gao J, Li P, Chen Z, Zhang J. A survey on deep learning for multimodal data fusion. Neural Comput 2020;32(5):829–64.

[14] Baltrušaitis T, Ahuja C, Morency L-P. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 2018;41(2):423–43.

[15] Atrey PK, Hossain MA, El Saddik A, Kankanhalli MS. Multimodal fusion for multimedia analysis: a survey. MultimedSyst 2010;16(6):345–79.

[16] Zeng Z, Pantic M, Roisman GI, Huang TS. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 2008;31(1):39–58.

[17] Zhu H, Luo M-D, Wang R, Zheng A-H, He R. Deep audio-visual learning: a survey. IntJAutomComput 2021;18(3):351–76.

[18] Priyasad D, Fernando T, Denman S, Sridharan S, Fookes C. Attention driven fusion for multi-modal emotion recognition. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020. p. 3227–31. https://doi.org/10.1109/ICASSP40776.2020.9054441.

[19] Hosseini M-P, Tran TX, Pompili D, Elisevich K, Soltanian-Zadeh H. Multimodal data analysis of epileptic eeg and rs-fmri via deep learning and edge computing. Artif Intell Med 2020;104:101813. https://doi.org/10.1016/j.artmed.2020.101813. https://www.sciencedirect.com/science/article/pii/S0933365718306882.

[20] Tan K, Huang W, Liu X, Hu J, Dong S. A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction. Artif Intell Med 2022;126:102260. https://doi.org/10.1016/j.artmed.2022.102260. https://www.sciencedirect.com/science/article/pii/S0933365722000252.

[21] Liu H, Zheng Q, Li Z, Qin T, Zhu L. An efficient multi-feature svm solver for complex event detection. Multimed Tools Appl 2018;77(3):3509–32.

[22] Sulubacak U, Caglayan O, Grönroos S-A, Rouhe A, Elliott D, Specia L, Tiedemann J. Multimodal machine translation through visuals and speech. MachTransl 2020;34(2):97–147.

[23] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2014. arXiv preprint arXiv:1409.0473.

[24] Xu J, Yao T, Zhang Y, Mei T. Learning multimodal attention lstm networks for video captioning. In: Proceedings of the 25th ACM international conference on Multimedia; 2017. p. 537–45.

[25] Hori C, Hori T, Lee T-Y, Zhang Z, Harsham B, Hershey JR, Marks TK, Sumi K. Attention-based multimodal fusion for video description. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 4193–202.

[26] Qu M, Tang J, Shang J, Ren X, Zhang M, Han J. An attention-based collaboration framework for multi-view network representation learning. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management; 2017. p. 1767–76.

[27] Yaghoobzadeh Y, Schütze H. Multi-multi-view learning: multilingual and multi-representation entity typing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018. p. 3060–6.

[28] Gibson M, Andres R, Kennedy T, Coppard L, et al. Kellogg international work group on the prevention of falls by the elderly,the prevention of falls in later life. Dan Med Bull 1987;34(4):1–24.

[29] Cleeland C. Measurement of pain by subjective report. AdvPain ResTher 1989;12:391–403.

[30] Cleeland C, Ryan K. Pain assessment: global use of the brief pain inventory. In: Annals. Singapore: Academy of Medicine; 1994.

[31] Leveille SG, Jones RN, Kiely DK, Hausdorff JM, Shmerling RH, Guralnik JM, Kiel DP, Lipsitz LA, Bean JF. Chronic musculoskeletal pain and the occurrence of falls in an older population. JAMA 2009;302(20):2214–21.

[32] Boulahia SY, Amamra A, Madi MR, Daikh S. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. MachVisionApplic 2021;32(6):1–18.

[33] Khaire P, Kumar P, Imran J. Combining cnn streams of rgb-d and skeletal data for human activity recognition. multimodal Fusion for Pattern Recognition Pattern Recog Lett 2018;115:107–16. https://doi.org/10.1016/j.patrec.2018.04.035. URL https://www.sciencedirect.com/science/article/pii/S0167865518301636.

[34] Sahay S, Kumar SH, Xia R, Huang J, Nachman L. Multimodal relational tensor network for sentiment and emotion classification. ACL 2018;2018:20.